

Fang Chen
Kristiina Jokinen
Editors



Speech Technology

Theory and Applications

 Springer

Speech Technology

Fang Chen · Kristiina Jokinen
Editors

Speech Technology

Theory and Applications

 Springer

Editors

Fang Chen
Department of Computing Science &
Engineering
Chalmers University of Technology
412 96 Göteborg
Sweden
fanch@cs.chalmers.se

Kristiina Jokinen
Department of Speech Sciences
University of Helsinki
PO Box 9
FIN-00014 Helsinki
Finland
kristiina.jokinen@helsinki.fi

ISBN 978-0-387-73818-5 e-ISBN 978-0-387-73819-2
DOI 10.1007/978-0-387-73819-2
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010926453

© Springer Science+Business Media, LLC 2010

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In recent years spoken language research has been successful in establishing technology which can be used in various applications, and which has also brought forward novel research topics that advance our understanding of the human speech and communication processes in general. This book got started in order to collect these different trends together, and to provide an overview of the current state of the art, as well as of the challenging research topics that deal with spoken language interaction technologies.

The topics have been broadly divided into two main classes: research on the enabling technology on one hand and applications that exemplify the use of the technology on the other hand. Each chapter is an independent review of a specific topic, covering research problems and possible solutions, and also envisaging development of the field in the near future. The basic technology development covers areas such as automatic speech recognition and speech synthesis, spoken dialogue systems and dialogue modelling, expressive speech synthesis, emotions and affective computing, multimodal communication and animated agents, while the applications concern speech translation, spoken language usage in cars, space, and military applications, as well as applications for special user groups. Discussion of the general evaluation methodologies is also included in the book.

The authors are leading figures of their field. Their experience provides a strong basis for the discussion of various aspects of the specific research topic and, in addition to their own work, for the presentation of a broad view of the entire research area. The core topics are discussed from the research, engineering, and industrial perspectives, and together the chapters provide an integrated view of the research and development in speech and spoken dialogue technology, representing a comprehensive view of the whole area.

The structure of the contributions follows a general format that consists of presenting the current state of the art in the particular research area, including a short history and development of the research as the background, and then discussing and evaluating new trends and view points related to the future of the field. Concerning different applications, the articles address both technical and usability issues, e.g. they provide an analysis of the factors that affect the application's real-time functioning, and usability requirements concerning their specific interaction and design issues.

The book is intended for readers, both in industry and academia, who wish to get a summary of the current research and development in the speech technology and its related fields. The aim of the book is to highlight main research questions and possible solutions so as to clarify the existing technological possibilities for the design and development of spoken language interactive systems, and also to inspire experimentation of new techniques, methods, models, and applications in speech technology and interaction studies. It also serves as a reference book to the historical development of the field and its various subfields, and as a collection of future visions of the research directions.

The spoken interaction research field has made great advances during the past years, and is active, maybe even more than ever, with a wide range of research topics. Many challenges need to be addressed, and new topics will obviously come across during the investigations. As a part of the joint efforts in research and development, the book will, hopefully, help the development of technologies to go forward, so as to satisfy the requirements of the users, and also our needs to understand human communication better.

The book consists of 15 chapters which are divided into three sections: basic speech interaction technology, developments of the basic technology, and applications of spoken language technology. The chapters are briefly introduced below.

The first four chapters discuss the major research areas in speech and interaction technology: speech recognition, speech synthesis, and dialogue systems.

Sadaoki Furui surveys the major themes and advances of the speech recognition research in his chapter *History and Development of Speech Recognition*. The chapter provides a technological perspective of the five decades of speech research that has produced models and solutions for automatic speech recognition. The chapter also points out performance challenges under a broad range of operating conditions, and that a greater understanding of the human speech process is needed before speech recognition applications can approach human performance.

An overview of the speech synthesis research is given by **David Suendermann**, **Harald Höge** and **Alan Black** in their chapter *Challenges in Speech Synthesis*. In the first part, within a historical time scale that goes back about 1,000 years, various types of speaking machines are discussed, from the first mechanical synthesizers to the modern electronic speech synthesizers. The second part concerns speech corpora, standardisation, and the evaluation metrics, as well as the three most important techniques for the speech synthesis. The chapter finishes with the speech synthesis challenges, referring to evaluation competitions in speech synthesis that are believed to boost research and development of the field.

Kristiina Jokinen focuses on dialogue management modelling in her chapter *Spoken Language Dialogue Models*, and outlines four phases within the history of dialogue research, growing from one another as an evolutionary chain according to interests and development lines typical for each period. The chapter reviews the development from the point of view of a Thinking Machine, and also surveys different dialogue models that have influenced dialogue management systems and aimed at enabling natural interaction in speech and interaction technology.

Roberto Pieraccini discusses spoken dialogue technology from the industrial view point in his chapter *The Industry of Spoken Dialogue Systems and the Third Generation of Interactive Applications*. The starting point of the overview is in the change of perspective in spoken language technology, from the basic research of human-like natural understanding of spontaneous speech to technologically feasible solutions and development of dialogue systems. The chapter describes a general architecture of dialogue systems and also compares visual and voice applications. With a reference to spoken dialogue industry, the life cycle of speech applications is presented and the challenges of the third generation of spoken dialogue systems discussed.

The next five chapters move on to describe novel research areas which have grown within the main research and development work in recent years.

Julia Hirschberg presents an overview of the recent computational studies in spoken and written deception. Her chapter *Deceptive Speech: Clues from Spoken Language* reviews common approaches and potential features which have been used to study deceptive speech. Moreover, she reports some results on deception detection using language cues in general and spoken cues in particular.

Roger K. Moore discusses the cognitive perspective to speech communication in his chapter *Cognitive Approaches to Spoken Language Technology*. The chapter argues that, although spoken language technology has successfully migrated from the research laboratories into practical applications, there are shortfalls in the areas in which human beings excel, and that they would seem to result from the absence of cognitive level processes in contemporary systems. There is relatively little spoken language technology research that draws directly on models of human cognition or exploits the cognitive basis of human spoken language. The chapter attempts to redress the balance by offering some insights into where to draw insights and how this might be achieved.

Nick Campbell addresses the issue of human speech communication in his chapter *Expressive Speech Processing and Prosody Engineering*. He does not focus upon the linguistic aspects of speech, but rather on its structure and use in interactive discourse, and shows that prosody functions to signal much more than syntactic or semantic relationships. After considering prosodic information exchange from a theoretical standpoint, he discusses acoustic evidence for the ideas and finally suggests some technological applications that the broader view of spoken language interaction may give rise to.

Elisabeth André and **Catherine Pelachaud** concentrate on the challenges that arise when moving from speech-based human–computer dialogue to face-to-face communication with embodied conversational agents. Their chapter *Interacting with Embodied Conversational Agents* illustrates that researchers working on embodied conversational agents need to address aspects of social communication, such as emotions and personality, besides the objectives of robustness and efficiency that the work on speech-based dialogue is usually driven by. The chapter reviews ongoing research in the creation of embodied conversational agents and shows how these agents are endowed with human-like communicative capabilities:

the agents can talk and use different discourse strategies, display facial expressions, gaze pattern, and hand gestures in occurrence with their speech.

Jianhua Tao presents affective computing history and problem setting in his chapter *Multimodal Information Processing for Affective Computing*. The chapter studies some key technologies which have been developed in recent years, such as emotional speech processing, facial expression, body gesture, and movement, affective multimodal system, and affect understanding and generation. The chapter also introduces some related projects and discusses the key topics which comprise a large challenge in the current research.

The last five chapters of the book focus on applications, and it contains articles on different types of speech applications, including also a chapter on system evaluation.

Farzad Ehsani, Robert Frederking, Manny Rayner, and Pierrette Bouillon give a survey of speech-based translation in their chapter *Spoken Language Translation*. The chapter covers history and methods for speech translation, including the dream of a Universal Translator, and various approaches to build translation engines. The chapter discusses the specific requirements for speech translation engines, concerning especially the component technologies of the speech recognition and speech synthesis. Finally, some representative systems are introduced and their architecture and component technologies discussed in detail.

Fang Chen, Ing-Marie Jonsson, Jessica Villing, and Staffan Larsson in their chapter *Application of speech technology in vehicles* summarise challenges of applying speech technology into vehicles, discuss the complexity of vehicle information systems, requirements for speech-based interaction with the driver, and discuss speech as an input/output modality. The chapter also presents dialogue-based conversational systems and multimodal systems as new next-level speech interfaces, and discusses the effects of emotion, mood, and the driver's personality on the application design.

Manny Rayner, Beth Ann Hockey, Jean-Michel Renders, Nikos Chatzichrisafis, and Kim Farrel describe Clarissa, a voice-enabled procedure browser which is apparently the first spoken dialogue system used in the International Space Station. The chapter *Spoken Dialogue Application in Space* focuses on the main problems and their solutions in the design and development of the Clarissa system. Detailed presentations are given on how to create voice-navigable versions of formal procedure documents, and how to build robust side-effect free dialogue management for handling undo's, corrections, and confirmations. The chapter also outlines grammar-based speech recognition using the Regulus toolkit and methods for accurate identification of cross-talk based on Support Vector Machines.

Jan Noyes and Ellen Haas describe various speech applications in a very demanding context: military domain. Their chapter *Military Applications* introduces the characteristics of the military domain by considering the users, technologies, and the environment in which the applications are used, pointing out harsh requirements for accuracy, robustness, and reliability in safety-critical conditions where applications are used in noisy and physically extreme environments by users who are subject to stress, time pressure, and workload. The chapter also presents some

typical speech-based applications dealing with Command-and-Control, teleoperation, information entry and information retrieval, repair and maintenance, training, and language translation.

Diamantino Freitas summarizes the use of speech technology in an alternative, or we may also say: augmentative way, to enable accessibility in communication technology for people with special needs. Problems as well as solutions are presented concerning specific situations of the visually disabled, the mobility impaired, the speech impaired, the hearing impaired, and the elderly. Problems found generally in public sites or sites providing transportation information require special attention in order to allow easy navigation and access to information. Also instructional games and eBooks are considered in examining what main benefits can be extracted from the use of speech technology in communication technology. It is concluded that solutions to improve communication difficulties for disabled persons may also bring advantages for non-disabled persons by providing redundancy and therefore a higher comfort in the use of communication systems.

Sebastian Möller summarizes the work on the evaluation of speech-based systems in his chapter *Assessment and Evaluation of Speech-based Interactive Systems*. After a brief history of the evaluation tasks, the chapter defines the concepts of performance and quality, which are often used to describe how well the service and the system fulfills the user's or the system designer's requirements. The chapter then moves on to a more detailed analysis of the assessment of individual system components: speech recognition, understanding, dialogue management, and speech synthesis, as well as the system as a whole. It also discusses a number of evaluation criteria that have been used to quantify the evaluation parameters, and which are partially standardized, as well as methods and frameworks for user evaluation. It is pointed out that new principles have to be worked out for capturing user experience and for evaluating multimodal systems, preferably in an automatic way.

Göteborg, Sweden
Helsinki, Finland

Fang Chen
Kristiina Jokinen

Acknowledgments

Several colleagues have kindly read the earlier versions of the book and commented on the individual chapters. We are grateful for their assistance in shaping the book into its final format, and would like to thank: Jens Allwood, Staffan Larsson, Michael McTear, Jan Noyes, Arne Ranta, Manny Rayner, and Graham Wilcock.

Contents

1	History and Development of Speech Recognition	1
	Sadaoki Furui	
2	Challenges in Speech Synthesis	19
	David Suendermann, Harald Höge, and Alan Black	
3	Spoken Language Dialogue Models	33
	Kristiina Jokinen	
4	The Industry of Spoken-Dialog Systems and the Third Generation of Interactive Applications	61
	Roberto Pieraccini	
5	Deceptive Speech: Clues from Spoken Language	79
	Julia Hirschberg	
6	Cognitive Approaches to Spoken Language Technology	89
	Roger K. Moore	
7	Expressive Speech Processing and Prosody Engineering: An Illustrated Essay on the Fragmented Nature of Real Interactive Speech	105
	Nick Campbell	
8	Interacting with Embodied Conversational Agents	123
	Elisabeth André and Catherine Pelachaud	
9	Multimodal Information Processing for Affective Computing . . .	151
	Jianhua Tao	
10	Spoken Language Translation	167
	Farzad Ehsani, Robert Frederking, Manny Rayner, and Pierrette Bouillon	
11	Application of Speech Technology in Vehicles	195
	Fang Chen, Ing-Marie Jonsson, Jessica Villing, and Staffan Larsson	

12 Spoken Dialogue Application in Space: The Clarissa Procedure Browser	221
Manny Rayner, Beth Ann Hockey, Jean-Michel Renders, Nikos Chatzichrisafis, and Kim Farrell	
13 Military Applications: Human Factors Aspects of Speech-Based Systems	251
Jan M. Noyes and Ellen Haas	
14 Accessibility and Design for All Solutions Through Speech Technology	271
Diamantino Freitas	
15 Assessment and Evaluation of Speech-Based Interactive Systems: From Manual Annotation to Automatic Usability Evaluation	301
Sebastian Möller	
Index	323

Contributors

Elisabeth André Multimedia Concepts and Applications, University of Augsburg, Augsburg, Germany, andre@informatik.uni-augsburg.de

Alan Black Carnegie Mellon University, Pittsburgh, PA, USA, awb@cs.cmu.edu

Pierrette Bouillon ISSCO/TIM, University of Geneva CH-1211 Geneva 4, Switzerland, pierrette.bouillon@unige.ch

Nick Campbell Centre for Language and Communication Studies (CLCS), The University of Dublin, College Green, Dublin 2, Ireland, nick@td.ie

Nikos Chatzichrisafis ISSCO/TIM, University of Geneva, CH-1211 Geneva 4, Switzerland, nikos.chatzichrisafis@vozzup.com

Fang Chen Interaction Design, Department of Computer Science and Engineering, Chalmers University of Technology, Göteborg, Sweden, fanch@chalmers.se

Farzad Ehsani Fluential, Inc, 1153 Bordeaux Drive, Suite 211, Sunnyvale, CA 94089, USA, farzad@fluentialinc.com

Kim Farrell RIACS/NASA Ames Research Center, 444 Castro Street, Suite 320, Mountain View, CA 94041, USA; Yahoo, 701 First Avenue, Sunnyvale, CA 94089-0703, USA, k_farrell@mac.com

Robert Frederking Language Technologies Institute/Center for Machine Translation, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA, ref@cs.cmu.edu

Diamantino Freitas Speech Processing Laboratory, University of Porto, Faculty of Engineering, 4250-066 Porto, Portugal, dfreitas@fe.up.pt

Sadaoki Furui Department of Computer Science, Tokyo Institute of Technology, 2-12-1 Okayama, Meguro-ku, Tokyo, 152-8552 Japan, furui@cs.titech.ac.jp

Julia Hirschberg Department of Computer Science, Columbia University, New York, NY, USA, julia@cs.columbia.edu

Beth Ann Hockey UCSC UARC, NASA Ames Research Center, Mail Stop 19-26, Moffett Field, CA 94035-1000, USA, bahockey@bahrc.net

Ellen Haas U.S. Army Research Laboratory, Adelphi, MD 20783-1197, USA, ehaas@arl.army.mil

Harald Höge Siemens Corporate Technology, Munich, Germany, harald.hoege@siemens.com

Kristiina Jokinen University of Helsinki, Helsinki, Finland; University of Tartu, Tartu, Estonia, kristiina.jokinen@helsinki.fi

Ing-Marie Jonsson Toyota Information Technology Center, Palo Alto, CA, USA, ingmarie@gmail.com

Staffan Larsson Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Göteborg, Sweden, sl@ling.gu.se

Roger K. Moore Department of Computer Science, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK, r.k.moore@dcs.shef.ac.uk

Sebastian Möller Deutsche Telekom Laboratories, Technische Universität, Berlin, Germany, sebastian.moeller@telekom.de

Jan M. Noyes University of Bristol, Bristol, UK, j.noyes@bristol.ac.uk

Catherine Pelachaud LINC, IUT de Montreuil, Université de Paris 8, rue de la Nouvelle, France, catherina.pelachaud@telecom-paristech.fr

Roberto Pieraccini SpeechCycle, Inc. 26 Broadway 11th Floor, New York, NY 10004, USA, roberto@speechcycle.com

Manny Rayner ISSCO/TIM, University of Geneva, CH-1211 Geneva 4, Switzerland, emmanuel.rayner@unige.ch

Jean-Michel Renders Xerox Research Center Europe, 6 chemin de Maupertuis, Meylan, 38240, France, jean-michel.renders@xrce.xerox.com

David Suendermann SpeechCycle, Inc. 26 Broadway 11th Floor, New York, NY, USA, david@speechcycle.com

Jianhua Tao Chinese Academy of Sciences, Beijing, China, jhtao@nlpr.ia.ac.cn

Jessica Villing Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Göteborg, Sweden, jessica@ling.gy.se

List of Acronyms

ACAPO	Associação de Cegos e Amblíopes de Portugal
ACSS	Aural Cascading style sheet
ADAS	Advanced Drive Assistance Systems
AFV	Armoured Fighting Vehicle
ALT	Automated Language Translation
API	Application Programming Interface
ARPA	Advanced Research Projects Agency
ASR	Automatic Speech Recognition
ATAG	Authoring Tool Accessibility Guidelines
ATC	Air Traffic Control
ATIS	Air Travel Information System
AUI	Aural User Interface
AVP	Attribute-Value Pair
BDI	Belief-Desire-Intention agent
BNF	Backus-Naur Form
C2OTM	Command and Control On The Move
C3I	Command, Control, Communications, Intelligence
CBCA	Criteria-Based Content Analysis
CCXML	Call Control Extensible Markup Language
CECOM	Communications and Electronics Command
CELL	Cross-Channel Early Lexical Learning
CFG	Context Free Grammar
CLI	Command-line interface
COTS	Commercial Off-The-Shelf (COTS)
CSS2	Cascading Style Sheet 2
C-STAR	Consortium for Speech Translation Advanced Research
DAISY	Digital Accessible Information System
DARPA	Defense Advanced Research Projects Agency
DFA	Design-for-All
DM	Dialogue Manager
DM	Dialog Module
DTD	Document Type Description
DVI	Direct Voice Input

DVO	Direct Voice Output
EAGLES	Expert Advisory Group on Language Engineering Standards
EBL	Explanation Based Learning
ECESS	European Center of Excellence in Speech Synthesis
ECMA	European Computer Manufacturers Association
EBMT	Example-Based Machine Translation
FAP	Face Animation Parameters
FDA	Fisher linear Discriminant Analysis
FIA	Form Interpretation Algorithms
GLM	Grammar-based Language Model
GMM	Gaussian Mixture Model
GSL	Grammar Specification Language
GUI	Graphical User Interface
HMIHY	How May I Help You
HMM	Hidden Markov Model
HTML	Hyper Text Markup Language
HTTP	Hypertext Transfer Protocol
ICA	Independent Component Analysis
ISO	International Standardisation Organisation
ISS	International Space Station
ITU	International Telecommunication Union
ITU-T	Telecommunication Standardization Sector
IVIS	In-Vehicle Information Systems
IVR	Interactive Voice Response
J2EE	Java 2 Platform Enterprise Edition
JAWS	Job access with speech
JSP	Java Server Pages
KBMT	Knowledge-Based Machine Translation
K-NN	K Nearest Neighbors
LASER	Language And Speech Exploitation Resources
LaTeX	Macros for the TeX text processor
LD	Learning disabilities
LIWC	Linguistic Inquiry and Word Count
LPC	Linear Predictive Coding
LVCSR	Large Vocabulary Continuous Speech Recognition
mac-address	Medium Access Control Address
MathML	Mathematical mark-up language
MCI	Minimum Classification Error
MIT	Massachusetts Institute of Technology
MLLR	Maximum Likelihood Linear Regression
MLP	Meridian Lossless Packing
MMI	Maximum Mutual Information
MRCP	Media Resource Control Protocol
MVC	Model View Controller
NASA	National Aeronautics and Space Administration

NATO	North Atlantic Treaty Organisation
NVID	Naval Voice Interactive Device
OAA	Open Agent Architecture
OCR	Optical Character Recognition Software
PARADISE	PARAdigm for DIAlogue System Evaluation
PCA	Principal Component Analysis
PCM	Pulse Code Modulation
PDA	Personal Digital Assistant
PDF	Portable Document Format
PDM	Point Distribute Model
PSTN	Public Switched Telephone Network
QA	Quality Assurance
QLF	Quasi Logical Form
RASTA	Relative Spectral Transform processing
RoI	Region of Interest
SaaS	Software As A Service
SASSI	Subjective Assessment of Speech System Interfaces
SCXML	State Chart Extensible Markup Language
SemER	Semantic Error Rate
SER	Sentence Error Rate
SDS	Spoken Dialogue System
SLM	Statistical Language Model
SLU	Statistical Language Understanding
SPINE	SPeech In Noisy Environments
SPORT	Soldier's On-system Repair Tool
SR	Speech Recognizer
SRGS	Speech Recognition Grammar Specification
SRI	Stanford Research Institute
SS	Speech synthesizer
SSML	Speech Synthesis Markup Language
SUSAS	Speech Under Simulated and Actual Stress
SVM	Support Vector Machine
TDNN	Time Delay Neural Network
TTS	Text-To-Speech Synthesis
UAV	Uninhabited/Unmanned Aerial Vehicle
URL	Uniform Resource Locator
VID	Voice Interactive Display
VQ	Vector quantization
VUI	Voice User Interface
W3C	World Wide Web Consortium
WAI	Web Accessibility Initiative
WCAG	Web contents accessibility guidelines
WER	Word Error Rate
XML	eXtensible Markup Language

About the Authors

Elisabeth André is a full professor of Computer Science at Augsburg University, Germany, and chair of the Laboratory for Multimedia Concepts and their Applications. Prior to that, she worked as a principal researcher at DFKI GmbH. Elisabeth André has been leading various academic and industrial projects in the area of user interfaces and participated in a number of international cooperations funded by the European Commission, such as CALLAS, e-CIRCUS, DynaLearn, IRIS, METABO and HUMAINE, and the German Science Foundation, such as Cube-G. Her current research interests include multimodal man-machine communication, conversational embodied agents, affective computing, and interactive storytelling. Elisabeth André is on the Editorial Board of *Cognitive Processing (International Quarterly of Cognitive Science)*, *Universal Access to the Information Society: An Interdisciplinary Journal*, *AI Communications (AICOM)*, *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)*, *Journal on Multimodal Interfaces*, *IEEE Transactions on Affective Computing (TAC)*, *ACM Transactions on Intelligent Interactive Systems (TIIS)*, and *International Journal of Synthetic Emotions (IJSE)*. In 2007, Elisabeth André was appointed an Alcatel Research Fellow at Internationales Zentrum für Kultur- und Technikforschung of Stuttgart University (IZKT).

Alan Black is associate research professor in language Technologies Institute, Carnegie Mellon University. He has authored or co-authored over 140 refereed papers. His recent research interests cover Speech Processing – Interactive Creation and Evaluation Toolkit for New Languages: automatically building recognition and synthesis support in new languages; Evaluation and Personalization of Synthetic Voices; Flexible voice synthesis through articulatory voice transformation; Speech Synthesis for telling children’s stories; Designing better spoken dialog systems for the elderly and non-natives; Speech-to-speech translation: Transtac (Iraqi), LASER ACTD (Thai), Babylon (Arabic) and Tongues (Croatian); a small fast run-time synthesis engine. Providing fast resource-light scalable speech synthesis for speech technology applications; providing automated methods for building new voices and languages for speech synthesis; Finding automatic training techniques to build

domain specific synthesis voices to capture individual style, domain, and prosodic characteristics.

Pierrette Bouillon started working at TIM/ISSCO in 1989. She has been involved in different projects such as: MULTEXT (LRE project), SLT, MEDTAG, DiET (LE project) and ISLE (LE project). Between 2000 and 2007, she was a Maître d'Enseignement et de Recherche (MER) at the School of Translation and Interpretation, and was made professor in 2007. Her principal research area is in lexical semantics and in particular the generative lexicon.

Nick Campbell is SFI Stokes Professor of Speech & Communication Technology at Trinity College Dublin, Ireland. He received his PhD in experimental psychology from the University of Sussex in the UK. He was chief researcher in the Department of Acoustics and Speech Research at the Advanced Telecommunications Research Institute International (ATR) in Kyoto, Japan, where he also served as research director for the JST/CREST *Expressive Speech Processing* and the SCOPE *Robot's Ears* projects. He was first invited as a Research Fellow at the IBM UK Scientific Centre, where he developed algorithms for speech synthesis, and later at the AT&T Bell Laboratories where he worked on the synthesis of Japanese. He served as Senior Linguist at the Edinburgh University Centre for Speech Technology Research before joining ATR in 1990. His research interests are based on large speech databases, and include nonverbal speech processing, concatenative speech synthesis, and prosodic information modelling. He is visiting professor at the Nara Institute of Science & Technology (NAIST) and at Kobe University in Japan.

Farzad Ehsani is president and CEO of the translation company Fluentia, and has 18 years of experience in research, management, and software development at companies such as Motorola, NEC, and Digital Equipment Corporation. Prior to Fluentia, he was the business and engineering head of Entropic's Application Group, which was sold to Microsoft. He has an MSc degree in electrical engineering and a BSc degree in computer science, both from MIT. He holds several patents and has authored numerous papers on the technological advancements in his field. He is fluent in three languages, with a working knowledge of four other languages.

Kim Farrell is a software development manager in the area of search relevance at Yahoo! Inc., California. She has previously managed R&D projects in the areas of spoken dialogue systems and autonomous planning and scheduling, as well as leading work on several commercial business applications.

Diamantino Freitas is an engineer and associate professor at the Electrotechnical and Computer Engineering Department of the Faculty of Engineering of the University of Porto (FEUP) and is presently in charge of the coordination of LSS – Laboratory of Signals and Systems research unit – Laboratory of Speech Processing. Has participated and coordinated several R&D projects and has been a national delegate to several COST actions, in particular COST 219 ter. The research

areas are: Acoustic Communication, Biomedical Engineering and Rehabilitation and Telecommunications for Persons with Special Needs.

Sadaoki Furui is professor of the Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology. He has also served as Dean of the Graduate School of Information Science and Engineering from 2007 to 2009, and is now serving as Director of Institute Library. His research interests include analysis of speaker characterization information in speech waves and its application to speaker recognition as well as interspeaker normalization and adaptation in speech recognition. He is also interested in vector-quantization-based speech recognition algorithms, spectral dynamic features for speech recognition, speech recognition algorithms that are robust against noise and distortion, algorithms for Japanese large-vocabulary continuous-speech recognition, automatic speech summarization algorithms, multimodal human-computer interaction systems, automatic question-answering systems, and analysis of the speech perception mechanism. He has authored or co-authored over 800 published articles.

Ellen Haas was born in Denver, Colorado. She received her B.A. in psychology from the University of Colorado at Boulder, Colorado; her M.S. in industrial psychology from California State University at Long Beach, California; and her Ph.D. in industrial and systems engineering from Virginia Polytechnic Institute and State University in Blacksburg, Virginia. She has been a Research Engineer at the U.S. Army Research Laboratory (ARL) Human Research and Engineering Directorate (HRED) at Aberdeen Proving Ground for over 25 years, and heads the ARL Multimodal Controls and Displays Laboratory. She is currently Acting Branch Chief of the Translational Neuroscience Branch. She has written and published extensively about auditory and tactile displays, multisensory displays, and speech recognition controls. She co-edited the book *The Ergonomics of Sound* (Santa Monica, California: Human Factors and Ergonomics Society, 2003), a compilation of relevant Human Factors and Ergonomics Society proceedings papers covering a broad range of sound-related applications. Dr. Haas has been a member of the International Ergonomics Association (IEA), and is the founder and chair of the IEA Auditory Ergonomics Technical Committee. She is a member of the Human Factors and Ergonomics Society and of the American Psychology Association Division 21 (Applied Experimental and Engineering Psychology).

Julia Hirschberg is professor of computer science at Columbia University. From 1985–2003 she worked at Bell Labs and AT&T Labs, as member of Technical Staff working on intonation assignment in text-to-speech synthesis and then as Head of the Human Computer Interaction Research Department. Her research focuses on prosody in speech generation and understanding. She currently works on speech summarization, emotional speech, and dialogue prosody. Julia Hirschberg was President of the International Speech Association from 2005-2007 and served as coeditor of *Speech Communication* from 2003-2006. She was editor-in-chief of *Computational Linguistics* and on the board of the Association for Computational

Linguistics from 1993–2003. She has been a fellow of the American Association for Artificial Intelligence since 1994.

Beth Ann Hockey received her PhD in Linguistics, University of Pennsylvania at Philadelphia in 1998. She is Research Scientist in NASA Ames Research Center and Adjunct Associate Professor of Linguistics & Senior Research Scientist at University of California Santa Cruz. She has received several awards, such as NASA Ames Award for Excellence in 2005, and SpeechTek Big Bang Award for the most impactful speech application of 2005, for Clarissa.

Ing-Marie Jonsson is dividing her time between being a visiting assistant professor at the Department of Computer and Information Science at Linköping University, and the CTO of Ansima Inc. a usability and user research startup in the heart of Silicon Valley.

Staffan Larsson was educated at Göteborg University, and gained a PhD in Linguistics there in 2002. From 2003 to 2006 he was assistant professor of Language Technology at Göteborg University, where he is currently a researcher and lecturer. He has participated in several international projects funded by the European Union (TRINDI, SIRIDUS, D'Homme, TALK). In these projects he has lead the development of TrindiKit, a research toolkit of experimenting with and building dialogue systems based on the information state approach, and the GoDiS dialogue system. He is also co-owner and CTO of Talkamatic, a company specialising in multimodal in-vehicle dialogue systems.

Roger K. Moore is the chair of Spoken Language Processing in the Speech and Hearing Research Group (SPandH) at the University of Sheffield. He is Editor-in-Chief of the Journal *Computer Speech & Language* and also serves as Editorial Board member in several other international journals. He has over 30 years' experience in speech technology R&D and much of his work has been based on insights derived from human speech perception and production. In the 1970s, he introduced the *Human Equivalent Noise Ratio (HENR)* as a vocabulary-independent measure of the goodness of an automatic speech recogniser based on a computational model of human word recognition. In the 1980s, he published *HMM Decomposition* for recognising multiple simultaneous signals (such as speech in noise), based on observed properties of the human auditory system. During the 1990s and more recently, he has continued to champion the need to understand the similarities and differences between human and machine spoken language behaviour.

Sebastian Möller is Professor for Quality and Usability at Technische Universität Berlin and leads the Quality and Usability Lab at Deutsche Telekom Laboratories, Berlin, Germany. He works on the quality and usability of telecommunication services, including voice and video transmission, as well as speech-based and multi-modal interactive services. He published a book on the "Quality of Telephone-based Spoken Dialogue Systems" with Springer in 2005, and is currently Rapporteur in

Study Group 12 of the International Telecommunication Union (ITU-T), establishing Recommendations for the evaluation of speech-based services.

Jan M. Noyes PhD, DSc is professor of human factors psychology at the University of Bristol, UK. Her research interests are in human–computer interaction (HCI), and the application of cognitive psychology to design. In 1999, she was awarded the Otto Edholm medal for her significant contribution to the application of ergonomics. She has authored around 250 publications on human factors aspects of interface design including seven books and has worked extensively with industry. She was awarded the IEE Informatics Premium Award in 1998 for her paper on ‘engineering psychology and system safety,’ and has been on numerous journal editorial boards and international conference committees.

Catherine Pelachaud received her PhD in 1991 in computer graphics at the University of Pennsylvania, Philadelphia, USA. Her research interests deal with Embodied Conversational Agents, interactive, multimodal systems, and human-machine interaction. She has published over hundred papers and book chapters in broad research areas, covering nonverbal behaviour (facial expression, gesture, gaze), and models of expressive and emotional behaviour, audio-visual speech (lip movement, co-articulation), as well as feedback behaviour.

Roberto Pieraccini has been in speech technology and research for more than 25 years working at organizations such as CSELT (Italy), Bell Laboratories, AT&T Labs, SpeechWorks, and IBM Research. The development of statistical models for language understanding and the use of reinforcement learning for dialog systems are among his contributions to the field. He is currently the CTO of SpeechCycle (www.speechcycle.com), a company that develops advanced 3rd Generation Spoken Dialog applications. Visit Roberto’s Web site at www.thepieraccinis.com for a full list of his publications.

Manny Rayner is a senior researcher at Geneva University, Switzerland, and has previously held positions at SRI International and NASA Ames Research Center. He is lead developer of the open source Regulus platform, a toolkit used for building grammar-based spoken dialogue systems, and is the first author of *The Spoken Language Translator* and *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler*.

Jean-Michel Renders has a PhD in Applied Sciences from the University of Brussels XEROX. He started his research activities in 1988, in the field of Robotic Dynamics and Control after which he joined the Joint Research Centre of the European Communities to work on biological metaphors (Genetic Algorithms, Neural Networks and Immune Networks) applied to process control. Since then, his research activities have covered Artificial Intelligence and Machine Learning Techniques in Industry. His current research interests mainly focus on Machine Learning techniques applied to Statistical Natural Language Processing and Text

Mining. He is a member of the European Network of Excellence PASCAL (Pattern Analysis, Statistical modelling and ComputAtional Learning) and is involved in the European IST Project REVEAL THIS (Retrieval of Video and Language for The Home user in an Information Society).

David Suendermann is the principal speech scientist of SpeechCycle, Inc., in New York. Before joining SpeechCycle in 2007, he was with Siemens Corporate Technology in Munich and had spent several years as visiting scientist at universities in New York, Los Angeles, Barcelona, and Aachen. He has performed more than 10 years of active research on speech and natural language processing, dialog systems, and machine learning and holds a PhD degree from the Bundeswehr University in Munich.

Jianhua Tao is a professor from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His current research interests include speech synthesis and recognition, human computer interaction, emotional information processing. He has published more than 60 papers in major journals and proceedings, such as IEEE Trans. on ASLP, ICASSP, Interspeech, ICME, ICPR, ICCV, ICIP. Prof. Tao has received several awards from important conferences, including Eurospeech 2001. In 2006, he was elected as vice-chairperson of the ISCA Special Interest Group of Chinese Spoken Language Processing (SIG-CSLP), and he is an executive committee member of the HUMAINE association. He is also the steering committee member of IEEE Trans. on affective computing and the editorial board member of International Journal on Synthetic Emotions and Journal on Multimodal Interface.

Jessica Villing (born 19 February 1968) was educated at the University of Gothenburg and gained a MA in computational linguistics in 2005. She is currently a PhD student at GSLT (Graduate School of Language Technology) and the department of Philosophy, Linguistics and Theory of Science at the University of Gothenburg. She has participated in international as well as national projects with focus on in-vehicle dialogue systems and cognitive workload.

About the Editors

Fang Chen is associate professor at the Computing Science Department, Chalmers University of Sweden. She has been working in Human Factors and Ergonomics research for over 20 years and has published over 20 papers in the cognitive science specially related to speech technology application. Chen has over 20 years teaching and research experience on ergonomics, human factors, and human–computer interaction. She has been teaching on human cognition, human–computer interaction, usability and user-centred design, and research methodology in undergraduate and graduate level. In the past 8 years, her research interests in focused on speech and multimodal interaction design in different applications.

Kristiina Jokinen is adjunct professor of interaction technology at the University of Tampere and a visiting professor of intelligent user interfaces at the University of Tartu, Estonia. She was Nokia Foundation Visiting Fellow at Stanford University in 2006, and she has played a leading role in several academic and industrial research projects. She is the author of the book *Constructive Dialogue Management – Speech Interaction and Rational Agents*. Her research concerns human–computer interaction, spoken dialogue systems, and multimodal communication. She is the secretary of SIGDial, the ACL/ISCA Special Interest Group for Discourse and Dialogue.

Chapter 1

History and Development of Speech Recognition

Sadaoki Furui

1.1 Introduction

Speech is the primary means of communication between humans. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to the desire to automate simple tasks which necessitate human-machine interactions, research in automatic speech recognition by machines has attracted a great deal of attention for five decades.

Based on major advances in statistical modeling of speech, automatic speech recognition systems today find widespread application in tasks that require human-machine interface, such as automatic call processing in telephone networks and query-based information systems that provide updated travel information, stock price quotations, weather reports, etc.

This chapter reviews major highlights during the last five decades in the research and development of automatic speech recognition so as to provide a technological perspective. Although many technological progresses have been made, there still remain many research issues that need to be tackled.

1.2 Five Decades of Progress in Speech Recognition

The progress of automatic speech recognition (ASR) technology in the past five decades [13, 20, 41] can be organized as follows:

- First generation: earliest attempts in the 1950s and 1960s
- Second generation: template-based technology in the late 1960s and 1970s
- Third generation: statistical model-based technology in the 1980s
- Late third generation: advancements in the 1990s and 2000s

S. Furui (✉)

Department of Computer Science, Tokyo Institute of Technology, 2-12-1 Okayama, Meguro-ku, Tokyo, 152-8552 Japan
e-mail: furui@cs.titech.ac.jp

1.2.1 The First-Generation Technology (1950s and 1960s)

Earliest attempts: The earliest attempts to devise ASR systems were made in the 1950s and 1960s, when various researchers tried to exploit fundamental ideas of acoustic phonetics. Since signal processing and computer technologies were yet very primitive, most of the speech recognition systems investigated used spectral resonances during the vowel region of each utterance which were extracted from output signals of an analog filter bank and logic circuits.

Early systems: In 1952, at Bell Laboratories in USA, Davis, Biddulph, and Balashek built a system for isolated digit recognition for a single speaker [7], using the formant frequencies measured/estimated during vowel regions of each digit. In an independent effort at RCA Laboratories, USA, in 1956, Olson and Belar tried to recognize ten distinct syllables of a single speaker, as embodied in ten monosyllabic words [38]. In 1959, at University College in England, Fry and Denes tried to build a phoneme recognizer to recognize four vowels and nine consonants [10]. By incorporating statistical information concerning allowable phoneme sequences in English, they increased the overall phoneme recognition accuracy for words consisting of two or more phonemes. This work marked the first use of statistical syntax (at the phoneme level) in automatic speech recognition. In 1959, Forgie and Forgie at MIT Lincoln Laboratories devised a system which was able to recognize ten vowels embedded in a /b/ – vowel – /t/ format in a speaker-independent manner [9]. In the 1960s, since computers were still not fast enough, several special-purpose hardwares were built. Suzuki and Nakata at the Radio Research Lab in Japan built a hardware vowel recognizer [49]. Sakai and Doshita at Kyoto University built a hardware phoneme recognizer in 1962, using a hardware speech segmenter and a zero-crossing analysis of different regions of the input utterance [44]. Nagata and his colleagues at NEC Laboratories in Japan built a hardware digit recognizer in 1963 [37].

Time normalization: One of the difficult problems of speech recognition exists in the non-uniformity of time scales in speech events. In the 1960s, Martin and his colleagues at RCA Laboratories developed a set of elementary time-normalization methods, based on the ability to reliably detect speech starts and ends that significantly reduced the variability of the recognition scores [34]. Martin ultimately founded one of the first speech recognition companies, Threshold Technology.

1.2.2 The Second-Generation Technology (Late 1960s and 1970s)

DTW: In the late 1960s and 1970s, speech recognition research achieved a number of significant milestones. In 1968, in the Soviet Union, Vintsyuk proposed the use of dynamic programming methods for time aligning a pair of speech utterances (generally known as dynamic time warping (DTW)), including algorithms for connected word recognition [53]. However, his work was largely unknown in other countries

until the 1980s. At the same time, in an independent effort in Japan, Sakoe and Chiba at NEC Laboratories also started to use a dynamic programming technique to solve the non-uniformity problem [46]. Sakoe and Chiba further advanced their technique in the 1970s. Since the late 1970s, dynamic programming in numerous variant forms, including the Viterbi algorithm [54] which came from the communication theory community, has become an indispensable technique in automatic speech recognition.

Template-based isolated word recognition: The area of isolated word or discrete utterance recognition became a viable and usable technology based on fundamental studies in the Soviet Union and Japan. Velichko and Zagoruyko in the Soviet Union advanced the use of pattern recognition ideas in speech recognition [52]. Itakura, while working at Bell laboratories, showed how linear predictive coding (LPC) could be applied to speech recognition systems through the use of an appropriate distance measure based on LPC spectral parameters [16].

Continuous speech recognition: In the late 1960s, Reddy at Carnegie Mellon University (CMU) conducted pioneering research in the field of continuous speech recognition using dynamic tracking of phonemes [43].

IBM Labs: Researchers started studying large-vocabulary speech recognition for three distinct tasks, namely the New Raleigh language for simple database queries [50], the laser patent text language for transcribing laser patents [18], and the office correspondence task, called Tangora, for dictation of simple memos.

AT&T Bell Labs: Researchers began a series of experiments aimed at making speaker-independent speech-recognition systems [42]. To achieve this goal, a wide range of sophisticated clustering algorithms were used to determine the number of distinct patterns required to represent all variations of different words across a wide user population.

DARPA program: An ambitious speech-understanding project was funded by the Defense Advanced Research Projects Agency (DARPA) in USA, which led to many seminal systems and technologies [24]. One of the first demonstrations of speech understanding was achieved by CMU in 1973. Their Hearsay I system was able to use semantic information to significantly reduce the number of alternatives considered by the recognizer. CMU's Harpy system [33] was shown to be able to recognize speech using a vocabulary of 1,011 words with reasonable accuracy. One particular contribution from the Harpy system was the concept of graph search, where the speech recognition language is represented as a connected network derived from lexical representations of words, with syntactical production rules and word boundary rules. The Harpy system was the first to take advantage of a finite-state network (FSN) to reduce computation and efficiently determine the closest matching string.

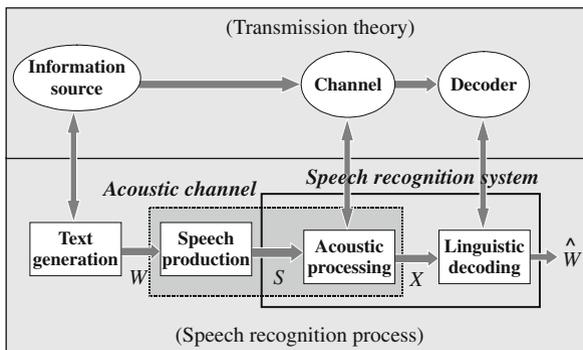
Other systems developed under the DARPA's speech-understanding program included CMU'S Hearsay II and BBN'S HWIM (Hear What I Mean) systems [24]. The approach proposed by Hearsay II of using parallel asynchronous processes that simulate the component knowledge sources in a speech system was a pioneering concept. A global "blackboard" was used to integrate knowledge from parallel sources to produce the next level of hypothesis.

1.2.3 The Third-Generation Technology (1980s)

Connected word recognition: The problem of creating a robust system capable of recognizing a fluently spoken string of connected word (e.g., digits) was a focus of research in the 1980s. A wide variety of the algorithms based on matching a concatenated pattern of individual words were formulated and implemented, including the two-level dynamic programming approach by Sakoe at NEC [45], the one-pass method by Bridle and Brown at Joint Speech Research Unit (JSRU) in UK [4], the level-building approach by Myers and Rabiner at Bell Labs [36], and the frame-synchronous level-building approach by Lee and Rabiner at Bell Labs [27]. Each of these “optimal” matching procedures had its own implementation advantages.

Statistical modeling: Speech recognition research in the 1980s was characterized by a shift in methodology from the more intuitive template-based approach (a straightforward pattern recognition paradigm) toward a more rigorous statistical modelling framework as shown in Fig. 1.1. Today, most practical speech recognition systems are based on the statistical framework developed in the 1980s and their results, with significant additional improvements having been made in the 1990s.

HMM: One of the key technologies developed in the 1980s is the hidden Markov model (HMM) approach [8, 40, 41]. It is a doubly stochastic process in that it has an underlying stochastic process that is not observable (hence the term hidden), but can be observed through another stochastic process that produces a sequence of observations. Although the HMM was well known and understood in a few laboratories (primarily IBM, Institute for Defense Analysis (IDA) and Dragon Systems), it was not until widespread publication of the methods and theory of HMMs in the mid-1980s that the technique became widely applied in virtually every speech recognition research laboratory in the world.



$$\hat{W} = \arg \max_w P(W|X) = \arg \max_w \frac{P(X|W) P(W)}{P(X)}$$

Fig. 1.1 Statistical modeling framework of speech production and recognition system based on information transmission theory

Δ Cepstrum: Furui proposed to use the combination of instantaneous cepstral coefficients and their first- and second-order polynomial coefficients, now called Δ and $\Delta\Delta$ cepstral coefficients, as fundamental spectral features for speech recognition [11]. He proposed this method for speaker recognition in the late 1970s, but no one attempted to apply it to speech recognition for many years. This method is now widely used in almost all speech recognition systems.

N-gram: A primary focus of IBM was the development of a structure of a language model (grammar), which was represented by statistical syntactical rules describing how likely, in a probabilistic sense, was a sequence of language symbols (e.g., phonemes or words) that could appear in the speech signal. The *N-gram* model, which defined the probability of occurrence of an ordered sequence of N words, was introduced, and, since then, the use of *N-gram* language models, and its variants, has become indispensable in large-vocabulary speech recognition systems [17].

Neural net: In the 1980s, the idea of applying neural networks to speech recognition was reintroduced. Neural networks were first introduced in the 1950s, but they did not prove useful because of practical problems. In the 1980s, a deeper understanding of the strengths and limitations of the technology was achieved, as well as an understanding of the relationship of this technology to classical pattern classification methods [22, 30, 55].

DARPA program: The DARPA community conducted research on large-vocabulary, continuous speech recognition systems, aiming at achieving high word accuracy for a 1000-word database management task. Major research contributions resulted from efforts at CMU with the SPHINX system [28], BBN with the BYBLOS system [6], SRI with the DECIPHER system [56], Lincoln Labs [39], MIT [57], and AT&T Bell Labs [26]. The SPHINX system successfully integrated the statistical method of HMM with the network search strength of the earlier Harpy system. Hence, it was able to train and embed context-dependent phone models in a sophisticated lexical decoding network.

1.2.4 The Third Generation, Further Advances (1990s)

Error minimization concept: In the 1990s, a number of innovations took place in the field of pattern recognition. The problem of pattern recognition, which traditionally followed the framework of Bayes and required estimation of distributions for the data, was transformed into an optimization problem involving minimization of the empirical recognition error [19]. This fundamental paradigmatic change was caused by the recognition of the fact that the distribution functions for the speech signal could not be accurately chosen or defined, and that Bayes' decision theory becomes inapplicable under these circumstances. Fundamentally, the objective of a recognizer design should be to achieve the least recognition error rather than provide the best fitting of a distribution function to the given (known) data set as advocated by the Bayes criterion. This error minimization concept produced a number of techniques, such as discriminative training and kernel-based methods.

As an example of discriminative training, the minimum classification error (MCE) criterion was proposed along with a corresponding generalized probabilistic descent (GPD) training algorithm to minimize an objective function which acts to approximate the error rate closely [5]. Another example was the maximum mutual information (MMI) criterion. In MMI training, the mutual information between the acoustic observation and its correct lexical symbol averaged over a training set is maximized. Although this criterion is not based on a direct minimization of the classification error rate and is quite different from the MCE-based approach, it is well founded in information theory and possesses good theoretical properties. Both the MMI and MCE can lead to speech recognition performance superior to the maximum likelihood-based approach [5].

DARPA program: The DARPA program continued into the 1990s, with emphasis shifting to natural language front ends to the recognizer. The central focus also shifted to the task of retrieving air travel information, the air travel information service (ATIS) task [2]. Later the emphasis was expanded to a range of different speech-understanding applications areas, in conjunction with a new focus on transcription of broadcast news (BN) and conversational speech. The Switchboard task is among the most challenging ones proposed by DARPA; in this task speech is conversational and spontaneous, with many instances of so-called disfluencies such as partial words, hesitation, and repairs. The BN transcription technology was integrated with information extraction and retrieval technology, and many application systems, such as automatic voice document indexing and retrieval systems, were developed. A number of human language technology projects funded by DARPA in the 1980s and 1990s further accelerated the progress, as evidenced by many papers published in *The Proceedings of the DARPA Speech and Natural Language/Human Language Workshop*.

Robust speech recognition: Various techniques were investigated to increase the robustness of speech recognition systems against the mismatch between training and testing conditions, caused by background noises, voice individuality, microphones, transmission channels, room reverberation, etc. Major techniques include the maximum likelihood linear regression (MLLR) [29], the model decomposition [51], parallel model composition (PMC) [15], and the structural maximum a posteriori (SMAP) method [47].

Applications: Speech recognition technology was increasingly used within telephone networks to automate as well as enhance operator services.

1.2.5 The Third Generation, Further Advances (2000s)

DARPA program: The effective affordable reusable speech-to-text (EARS) program was conducted to develop speech-to-text (automatic transcription) technology with the aim of achieving substantially richer and much more accurate output than before. The tasks include detection of sentence boundaries, fillers, and disfluencies. The program was focusing on natural, unconstrained human-human speech from broadcasts and foreign conversational speech in multiple languages. The goal was

to make it possible for machines to do a much better job of detecting, extracting, summarizing, and translating important information, thus enabling humans to understand what was said by reading transcriptions instead of listening to audio signals [32, 48].

Spontaneous speech recognition: Although read speech and similar types of speech, e.g., news broadcasts reading a text, can be recognized with accuracy higher than 95% using state-of-the-art speech recognition technology, recognition accuracy drastically decreases for spontaneous speech. Broadening the application of speech recognition depends crucially on raising recognition performance for spontaneous speech. In order to increase recognition performance for spontaneous speech, several projects have been conducted. In Japan, a 5-year national project “Spontaneous Speech: Corpus and Processing Technology” was conducted [14]. The world’s largest spontaneous speech corpus, “Corpus of Spontaneous Japanese (CSJ)” consisting of approximately 7 million words, corresponding to 700 h of speech, was built, and various new techniques were investigated. These new techniques include flexible acoustic modeling, sentence boundary detection, pronunciation modeling, acoustic as well as language model adaptation, and automatic speech summarization [12].

Robust speech recognition: To further increase the robustness of speech recognition systems, especially for spontaneous speech, utterance verification and confidence measures are being intensively investigated [25]. In order to have intelligent or human-like interactions in dialog applications, it is important to attach to each recognized event a number that indicates how confidently the ASR system can accept the recognized events. The confidence measure serves as a reference guide for a dialog system to provide an appropriate response to its users. To detect semantically significant parts and reject irrelevant portions in spontaneous utterances, a detection-based approach has recently been investigated [23]. This combined recognition and verification strategy works well especially for ill-formed utterances.

In order to build acoustic models more sophisticated than conventional HMMs, the dynamic Bayesian network has been investigated [58].

Multimodal speech recognition: Humans use multimodal communication when they speak to each other. Studies in speech intelligibility have shown that having both visual and audio information increases the rate of successful transfer of information, especially when the message is complex or when communication takes place in a noisy environment. The use of the visual face information, particularly lip information, in speech recognition has been investigated, and results show that using both types of information gives better recognition performances than using only the audio or only the visual information, particularly in noisy environment.

1.2.6 Summary of Technological Progress

In the last five decades, especially in the last two decades, research in speech recognition has been intensively carried out worldwide, spurred on by advances in signal processing, algorithms, architectures, and hardware. Technological progress can be

Table 1.1 Summary of the technological process in the last five decades

Past	Present (new)
1 Template matching	Corpus-base statistical modeling, e.g., HMM and <i>N-grams</i>
2 Filter bank/spectral resonance	Cepstral features (cepstrum + Δ cepstrum + $\Delta\Delta$ cepstrum)
3 Heuristic time-normalization	DTW/DP matching
4 "Distance"-based methods	Likelihood-based methods
5 Maximum likelihood approach	Discriminative approach, e.g., MCE/GPD and MMI
6 Isolated word recognition	Continuous speech recognition
7 Small vocabulary	Large-vocabulary recognition
8 Context-independent units	Context-dependent units
9 Clean speech recognition	Noisy/telephone speech recognition
10 Single speaker recognition	Speaker-independent/adaptive recognition
11 Monologue recognition	Dialog/conversation recognition
12 Read speech recognition	Spontaneous speech recognition
13 Recognition	Understanding
14 Single modality (audio signal only)	Multimodal (audio/visual) speech recognition
15 Hardware recognizer	Software recognizer
16 No commercial application	Many practical commercial applications

summarized by the changes in Table 1.1. The majority of technological changes have been directed toward the purpose of increasing robustness of recognition, including many other additional important techniques not noted above.

Recognition systems have been developed for a wide variety of applications, ranging from small-vocabulary keyword recognition over dialed-up telephone lines, to medium-size vocabulary voice interactive command and control systems for business automation, to large-vocabulary speech transcription, spontaneous speech understanding, and limited-domain speech translation. Figure 1.2 illustrates the progress of speech recognition technology in the past three decades according to

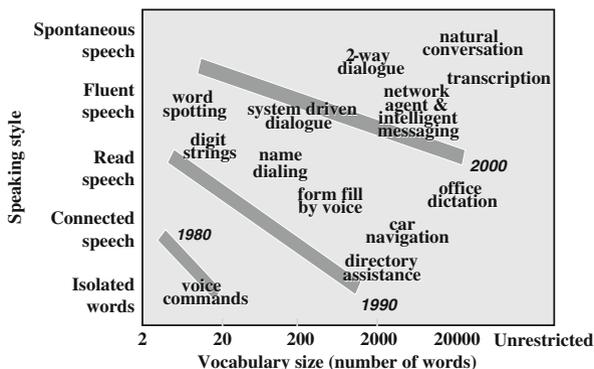


Fig. 1.2 Progress of speech recognition technology in the past three decades

generic application areas. The complexity of these application areas is characterized along two dimensions: the size of the vocabulary and the speaking style. It should be obvious that the larger the vocabulary, the more difficult the application task. Similarly, the degree of constraints in the speaking style has a very direct influence on the complexity of the application; a free conversation full of slurring and extraneous sounds such as “uh,” “um,” and partial words is far more difficult than words spoken in a rigidly discrete manner. Thus, the difficulty of an application grows from the lower left corner to the upper right corner in Fig. 1.2. The three diagonal lines with years demarcate the applications that can and cannot be supported by the technology for viable deployment in the corresponding time frame.

1.2.7 Changes in the Past Three Decades

Although we have witnessed many new technological promises, we have also encountered a number of practical limitations that hinder a widespread deployment of applications and services. Table 1.2 shows the research level of ASR techniques in 1977 [3]. Most of the techniques categorized into C: “a long way to go,” printed in bold-face, still even now have not been able to overcome problems preventing realization of goals. Table 1.3 shows a list of ASR problems in 1977. Roughly speaking, 16 problems out of 28, printed by bold-face, have not yet been solved.

Table 1.2 State of the art of ASR techniques in 1977

Processing techniques	State of the art
(1) Signal conditioning	A, except speech enhancement (C)
(2) Digital signal transformation	A
(3) Analog signal transformation and feature extraction	A, except feature extraction (C)
(4) Digital parameter and feature extraction	B
(5a) Re-synthesis	A
(5b) Orthographic synthesis	C
(6) Speaker normalization	
Speaker adaptation	C
Situation adaptation	
(7) Time normalization	B
(8) Segmentation and labeling	B
(9a) Language statistics	C
(9b) Syntax	B
(9c) Semantics	C
(9d) Speaker and situation pragmatics	C
(10) Lexical matching	C
(11) Speech understanding	B–C
(12) Speaker recognition	A for speaker verification C for all others
(13) System organization and realization	A–C
(14) Performance evaluation	C

A, useful now; B, shows promise; C, a long way to go

Table 1.3 ASR problems in 1977

-
- (1) **Detect speech in noise; speech/non-speech**
 - (2) Extract relevant acoustic parameters (poles, zeros, formant (transitions), slopes, dimensional representation, zero-crossing distributions)
 - (3) Dynamic programming (non-linear time normalization)
 - (4) Detect smaller units in continuous speech (word/phoneme boundaries; acoustic segments)
 - (5) **Establish anchor point; scan utterance from left to right; start from stressed vowel, etc.**
 - (6) **Stressed/unstressed**
 - (7) **Phonological rules**
 - (8) **Missing or extra-added (“uh”) speech sound**
 - (9) **Limited vocabulary and restricted language structure necessary; possibility of adding new words**
 - (10) **Semantics of (limited) tasks**
 - (11) Limits of acoustic information only
 - (12) Recognition algorithm (shortest distance, (pairwise) discriminant, Bayes probabilities)
 - (13) Hypothesize and test, backtrack, feed forward
 - (14) **Effect of nasalization, cold, emotion, loudness, pitch, whispering, distortions due to talker’s acoustical environment, distortions by communication systems (telephone, transmitter–receiver, intercom, public address, face masks), non-standard environments**
 - (15) **Adaptive and interactive quick learning**
 - (16) **Mimicking; uncooperative speaker(s)**
 - (17) **Necessity of visual feedback, error control, level for rejections**
 - (18) Consistency of references
 - (19) Real-time processing
 - (20) **Human engineering problem of incorporating speech understanding system into actual situations**
 - (21) Cost effectiveness
 - (22) **Detect speech in presence of competing speech**
 - (23) Economical ways to adding new speakers to system
 - (24) **Use of prosodic information**
 - (25) **Co-articulation rules**
 - (26) Morphology rules
 - (27) Syntax rules
 - (28) **Vocal-tract modeling**
-

Bold-face indicates problems that have still not been solved

1.3 Research Issues toward the Fourth-Generation ASR Technology

1.3.1 How to Narrow the Gap Between Machine and Human Speech Recognition

It has been shown that human speech recognition performs much better than the state-of-the-art ASR systems. In most recognition tasks, human subjects produce one to two orders of magnitude less errors than machines [31]. There is now increasing interest in finding ways to bridge this performance gap. Recent research in

human speech processing has shown that human beings actually perform speech recognition by integrating multiple knowledge sources [1]. What we know about human speech processing is still very limited, and we have yet to witness a complete and worthwhile unification of the science and technology of speech.

One of the most significant differences between human and automatic speech recognition is robustness against speech variations. The development of statistical methods, which make the system both easy to design, in terms of implementation, and capable of delivering somewhat sufficient performance, in limited tasks, has attracted enthusiasm in technology investment [19]. However, one needs to be rather careful in understanding the permissible operating conditions under which deployment of the system is viable. These conditions include the level of background noise, channel distortion and its variation, speaker dependency, allowable speaking styles and syntactic deviation, spontaneity of the speech, and so on. At present, a system would fail to deliver satisfactory performance if it is not used within the intended, often very narrowly defined, operating condition. Compared to a human listener, most of the spoken-language systems do not perform well when actual operating conditions deviate from the intended ones.

With the statistical method, which is data driven, one can in general improve the system performance by providing training data collected under the exact intended deployment condition. Although a system trained and operated under noisy conditions would still not perform, as well as a system trained and operated in a quiet acoustic ambient, its performance is substantially better than that of a system trained in a quiet, but operated in a noisy (mismatched) condition. The problem is, however, that collecting the “right” data is often very costly. The same notion applies to robustness against language models, speaking styles, as well as other conditions. For most of the applications, it is non-trivial to collect data to ensure coverage of the operating conditions. In other words, one should expect various degrees of condition “mismatch” between design/training in the laboratory and deployment in the field in almost all systems. The issue of robustness, thus, is to address the system’s inherent capability in dealing with the mismatch conditions.

1.3.2 Robust Acoustic Modeling

Let X be a random observation from an information source, consisting of M classes of event. We denote the classes by C_i , $i = 1, 2, \dots, M$. A recognizer’s job is to correctly classify each X into one of the M classes. In the context of statistical pattern recognition, the mismatch means that the maximum a posteriori decision rule is being implemented as, with Y denoting the actually received signal although the model parameter Λ has been obtained based on the training signal X . In general, we assume $Y = h(X, \theta)$ defined on some unknown parameter θ . The approach to the robustness issue can, thus, be addressed in several ways. One is to find and use

an invariant feature to represent the speech. An ideal invariant feature is a representation that will not fluctuate with the signal conditions; the same parameter Λ trained on X is expected to remain applicable for Y . This is obviously difficult to achieve. Another rather prevalent approach is to embed the function h in the a posteriori probability, i.e., to use $P_{\Lambda}(C_i|X) = P_{\Lambda}(C_i|h^{-1}(Y, \theta))$ in the decision rule. The interference parameter θ sometimes can be estimated from Y .

$$C(Y) = C_i, \quad \text{if } P_{\Lambda}(C_i|Y) = \max_j P_{\Lambda}(C_j|Y). \quad (1.1)$$

The most immediate concern in “condition mismatch” is noise and distortion. For additive noise, it is customary to assume, with X and Y being the “clean” and the “noisy” (observed) power spectral sequences, respectively.

$$Y = X + N \quad (1.2)$$

where N is the sequence of noise spectrum of an unknown (and possibly varying) level. If the interference is a linear distortion, then

$$Y = h(X) = \mathbf{H}X, \quad (1.3)$$

where \mathbf{H} is the frequency response of the linear distortion model. Note that in the case of linear distortion, Eq. (1.3) reduces to the form of Eq. (1.2) when a cepstral representation is used. The two types of interference are sometimes lumped together into a simplified function

$$Y = h(X) = \mathbf{H}X + N. \quad (1.4)$$

Much of the work toward robust speech recognition in the past decade focused on estimation of the parameters, \mathbf{H} and N , using Y [21]. Techniques such as spectral mean subtraction, signal bias removal, and maximum likelihood linear regression fall in this category.

The robustness issue can also encompass normalization of the observation to compensate for the variation due to talker differences. One technique that attempts to normalize the spectral difference due to vocal tract length variation among talkers was shown to bring about small but consistent improvement in speech recognition accuracy.

Another thrust to enhance the robustness in system performance is the area of adaptation. Following the above formulation, adaptation is to find $P_{\Lambda}(C_i|Y)$ from $P_{\Lambda}(C_i|X)$ based on a set of newly collected/observed data $\{Y\}$ and some prior knowledge of the distribution of Λ . The technique is effective for converting the speech model (either speaker dependent or speaker independent) to that of another talker using a limited but reasonable amount of new data. Speaker normalization and adaptation techniques in a non-statistical context have been an area of research for decades.

1.3.3 Robust Language Modeling

It is argued that a hidden Markov model with a mixture observation density in each state can adequately represent the acoustic variation manifested in the distribution of spectral parameters. This is due to the density approximation capability of such a model. Beyond the variation at the local acoustic level, however, the probabilistic nature of a language is often less understood.

The lack of a systematic study in probabilistically interpreting a language, as well as a large collection of statistical data, results in two technical areas in need of further research. One pertains to the representation of the linguistic structure ready for the application of probabilistic methods and the other the estimation method for reliable derivation of the relevant statistics for use in speech recognition. The former issue is equivalent to the definition of an event space based upon which a probabilistic model can be developed. Without such a representation, it is difficult to analyze the outcome of the statistical model.

Grammar is a rule system that governs a language. In terms of language processing, the complexity is compounded by the interaction between the structural rules and the lexical elements of expression such as words and phrases. Traditionally, linguists establish a grammar (the structural rules) based on elementary classes such as noun, verb, adjective, noun phrase, and so on, devoid of direct association of specific lexical element. However, the variation in our expression of message or concept comes from three essential components: the choice of lexical elements (words and phrases), the grammatical structure (one may argue that it's less probabilistic), and the interaction between them. It is not straightforward to address these elements of uncertainty and cast them in a formal probabilistic framework.

A number of grammatical representations and parsers exist [35]. The most pervasive is the finite-state grammar, which provides an integrated mechanism for addressing both variations in the sentential structure (traditionally addressed by a parser) and the choice of words. It is, however, a simplified and crude model of language. An N -gram language model is a special case (fixed-order) finite-state grammar; it addresses the probability of observing a word following a particular sequence of $N-1$ words. A finite-state grammar such as an N -gram model has the advantage of implementational ease. The fundamental issue with a finite-state grammar is the difficulty in having a precise coverage. Over-specification (which often happens with a high-order finite-state machine) leads to frequent encounters of out-of-grammar but legitimate expressions (i.e., under-coverage of the overall linguistic expressions). Under-specification, on the contrary, will have over-coverage which, while it alleviates out-of-grammar problems, reduces the accuracy and the effectiveness of the estimated probabilistic language model (e.g., many unlikely or impossible expressions in reality would have non-negligible probability assignments).

The issue of representation for a sequence of linguistic event also exists at the lexical level. People pronounce words differently due to many reasons. The realized phonemic content of a phrase in spoken utterance can vary rather vastly. The range of pronunciation variation is enormous. The need for a pronunciation

dictionary with a proper coverage to accommodate the variation is critical for a high-performance speech recognition system. The same question of structural representation applies here, although the implication of rules (grammatical versus lexical) is different. Research in this area in the past few years only produced slight improvements in recognition accuracy. The fundamental issue of a proper representation is still open.

In spontaneous utterances, ill-formed sentences with disfluencies such as repair, partial, and repetitive words are frequently observed. These ill-formed, as well as many other colloquial sentences, obviously deviate from the grammatical rules and usually lack the regularity to bring about a statistical significance. Language modeling for spontaneous utterances, from the structural representation to model adaptation to a particular talker (people's speaking habits differ), is one of the major challenges in this field of research. Adaptation or acquisition of the language structure to a particular communication context is also a worthwhile and active area of research.

1.3.4 Speech Corpora

As discussed above, data-driven methods have brought about fruitful results in the past decade. Unlike the traditional approach, in which knowledge of the speech behavior is "discovered" and "documented" by human experts, statistical methods provide an automatic procedure to "learn" the regularities in the speech data directly. The need of a large set of good training data is, thus, more critical than ever.

Establishing a good speech database for the machine to uncover the characteristics of the signal is not trivial. There are basically two broad issues to be carefully considered, one being the content and its annotation, and the other the collecting mechanism. The content of a database must reflect the intended use of the database.

For natural dialog applications such as the ATIS System in the DARPA program [2], a wizard setup is often used to collect the data. A wizard in this case is a human mimicking the machine in interacting with the user. Through the interaction, natural queries in sentential forms are collected. A committee is called upon to resolve cases that may be ambiguous in certain aspects. While a wizard setup can produce a useful set of data, it lacks the diversity particularly in situations where the real machine may fail. A human wizard cannot intentionally simulate a machine in error, and thus, the recorded data fail to provide information on real human-machine interaction.

The recorded data need to be verified, labeled, and annotated by people whose knowledge will be introduced into the design of the system through this learning process (i.e., via supervised training of the system after the data have been labeled). Labeling and annotation for isolated word utterances may be straightforward but tedious when the amount of data is large. For continuous speech recognition, nevertheless, this process can easily become unmanageable. For example, how do we annotate speech repairs and partial words, how do the phonetic transcribers reach a consensus in acoustic-phonetic labels when there is ambiguity, and how do we represent a semantic notion? Errors in labeling and annotation will result in system

performance degradation. How to ensure the quality of the annotated results is, thus, of major concern. Research in automating or creating tools to assist the verification procedure is by itself an interesting subject.

Another area of research that has gained interests is a modeling methodology and the associated data collection scheme that can reduce the task dependency. To maximize the performance, one should always strive for data that truly reflect the operating condition. It, thus, calls for a database collection plan that is consistent with the task. This data collection effort would soon become unmanageable if the system designer has to redo data collection for each and every application that is being developed. It is, therefore, desirable to design a task-independent data set and a modeling method that delivers a reasonable performance upon first use and can quickly allow in-field trial for further revision as soon as task-dependent data become available. Research result in this area can offer the benefit of a reduced application development cost.

1.4 Conclusion

Speech is the primary and the most convenient means of communication between people. Whether due to technological curiosity to build machines that mimic humans or desire to automate work with machines, research in speech recognition, as a first step toward natural human-machine communication, has attracted much enthusiasm over the past five decades. Although many important scientific advances have taken place, bringing us closer to the “Holy Grail” of automatic speech recognition and understanding by machine, we have also encountered a number of practical limitations which hinder a widespread deployment of application and services. In most speech recognition tasks, human subjects produce one to two orders of magnitude less errors than machines. There is now increasing interest in finding ways to bridge such a performance gap. What we know about human speech processing is very limited. Significant advances in speech recognition are not likely to come solely from research in statistical pattern recognition and signal processing. Although these areas of investigation are important, the most significant advances in fourth generation systems will come from studies in acoustic phonetics, speech perception, linguistics, and psychoacoustics. Future systems need to have an efficient way of representing, storing, and retrieving the “knowledge” required for natural conversation.

References

1. Allen, J. (2002). From Lord Rayleigh to Shannon: How do we decode speech? In: Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing, Orlando, FL, http://www.auditorymodels.org/jba/PAPERS/ICASSP/Plenary_Allen.asp.html.
2. ATIS Technical Reports (1995). Proc. ARPA Spoken Language Systems Technology Workshop, Austin, TX, 241–280.

3. Beek, B., Neuberg, E., Hodge, D. (1977). An assessment of the technology of automatic speech recognition for military applications. *IEEE Trans. Acoust., Speech, Signal Process.*, 25, 310–322.
4. Bridle, J. S., Brown, M. D. (1979). Connected word recognition using whole word templates. In: *Proc. Inst. Acoustics Autumn Conf.*, 25–28.
5. Chou, W. (2003). Minimum classification error (MCE) approach in pattern recognition. Chou, W., Juang, B.-H. (eds) *Pattern Recognition in Speech and Language Processing*. CRC Press, New York, 1–49.
6. Chow, Y. L., Dunham, M. O., Kimball, O. A. (1987). BYBLOS, the BBN continuous speech recognition system. In: *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, Dallas, TX, 89–92.
7. Davis, K. H., Biddulph, R., Balashek, S. (1952). Automatic recognition of spoken digits. *J. Acoust. Soc. Am.*, 24 (6), 637–642.
8. Ferguson, J. (ed) (1980). *Hidden Markov Models for Speech*. IDA, Princeton, NJ.
9. Forgie, J. W., Forgie, C. D. (1959). Results obtained from a vowel recognition computer program. *J. Acoust. Soc. Am.*, 31 (11), 1480–1489.
10. Fry, D. B., Denes, P. (1959). Theoretical aspects of mechanical speech recognition. The design and operation of the mechanical speech recognizer at University College London. *J. British Inst. Radio Eng.*, 19 (4), 211–229.
11. Furui, S. (1986). Speaker independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoust., Speech, Signal Process.*, 34, 52–59.
12. Furui, S. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Trans. Speech Audio Process.*, 12, 401–408.
13. Furui, S. (2004). Fifty years of progress in speech and speaker recognition. In: *Proc. 148th Acoustical Society of America Meeting*, San Diego, CA, 2497.
14. Furui, S. (2005). Recent progress in corpus-based spontaneous speech recognition. *IEICE Trans. Inf. Syst.*, E88-D (3), 366–375.
15. Gales, M. J. F., Young, S. J. (1993). Parallel model combination for speech recognition in noise. Technical Report, CUED/F-INFENG/TR135.
16. Itakura, F. (1975). Minimum prediction residual applied to speech recognition. *IEEE Trans. Acoust., Speech, Signal Process.*, 23, 67–72.
17. Jelinek, F. (1985). The development of an experimental discrete dictation recognizer. *Proc. IEEE*, 73 (11), 1616–1624.
18. Jelinek, F., Bahl, L., Mercer, R. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Trans. Inf. Theory*, 21, 250–256.
19. Juang, B. H., Furui, S. (2000). Automatic speech recognition and understanding: A first step toward natural human-machine communication. *Proc. IEEE*, 88 (8), 1142–1165.
20. Juang, B. H., Rabiner, L. R. (2005). *Automatic speech recognition: History*. Brown, K. (ed) *Encyclopedia of Language and Linguistics*, Second Edition, Oxford: Elsevier, New York, 11, 806–819.
21. Junqua, J. C., Haton, J. P. (1996). *Robustness in Automatic Speech Recognition*. Kluwer, Boston.
22. Katagiri, S. (2003). Speech pattern recognition using neural networks. Chou, W., Juang, B. H. (eds) *Pattern Recognition in Speech and Language Processing*. CRC Press, New York, 115–147.
23. Kawahara, T., Lee, C. H., Juang, B. H. (1998). Key-phrase detection and verification for flexible speech understanding. *IEEE Trans. Speech Audio Process*, 6, 558–568.
24. Klatt, D. (1977). Review of the ARPA speech understanding project. *J. Acoust. Soc. Am.*, 62 (6), 1324–1366.
25. Koo, M. W., Lee, C. H., Juang, B. H. (2001). Speech recognition and utterance verification based on a generalized confidence score. *IEEE Trans. Speech Audio Process*, 9, 821–832.
26. Lee, C. H., Giachin, E., Rabiner, L. R., Pieraccini, R., Rosenberg, A. E. (1990). Acoustic modeling for large vocabulary speech recognition. *Comput. Speech Lang.*, 4, 127–165.

27. Lee, C. H., Rabiner, L. R. (1989). A frame synchronous network search algorithm for connected word recognition. *IEEE Trans. Acoust., Speech, Signal Process.*, 37, 1649–1658.
28. Lee, K. F., Hon, H., Reddy, R. (1990). An overview of the SPHINX speech recognition system. *IEEE Trans. Acoust., Speech, Signal Process.*, 38, 600–610.
29. Leggetter, C. J., Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Comput. Speech Lang.*, 9, 171–185.
30. Lippmann, R. P. (1987). An introduction to computing with neural nets. *IEEE ASSP Mag.*, 4 (2), 4–22.
31. Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22, 1–15.
32. Liu, Y., Shriberg, E., Stolcke, A., Peskin, B., Ang, J., Hillard, D., Ostendorf, M., Tomalin, M., Woodland, P. C., Harper, M. (2005). Structural metadata research in the EARS program. In: *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, Montreal, Canada, V, 957–960.
33. Lowerre, B. (1980). The HARPY speech understanding system. Lea, W (ed) *Trends in Speech Recognition*. Prentice Hall, NJ, 576–586.
34. Martin, T. B., Nelson, A. L., Zadell, H. J. (1964). Speech recognition by feature abstraction techniques. Technical Report AL-TDR-64-176, Air Force Avionics Lab.
35. Moore, R. C. (1997). Using natural-language knowledge sources in speech recognition. Ponting, K. (ed) *Computational Models of Speech Pattern Processing*. Springer, Berlin, 304–327.
36. Myers, C. S., Rabiner, L. R. (1981). A level building dynamic time warping algorithm for connected word recognition. *IEEE Trans. Acoust., Speech, Signal Process.*, 29, 284–297.
37. Nagata, K., Kato, Y., Chiba, S. (1963). Spoken digit recognizer for Japanese language. *NEC Res. Develop.*, 6.
38. Olson, H. F., Belar, H. (1956). Phonetic typewriter. *J. Acoust. Soc. Am.*, 28 (6), 1072–1081.
39. Paul, D. B. (1989). The Lincoln robust continuous speech recognizer. In: *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, Glasgow, Scotland, 449–452.
40. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77 (2), 257–286.
41. Rabiner, L. R., Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliff, NJ.
42. Rabiner, L. R., Levinson, S. E., Rosenberg, A. E. (1979). Speaker independent recognition of isolated words using clustering techniques. *IEEE Trans. Acoust., Speech, Signal Process.*, 27, 336–349.
43. Reddy, D. R. (1966). An approach to computer speech recognition by direct analysis of the speech wave. Technical Report No. C549, Computer Science Department, Stanford University, Stanford.
44. Sakai, T., Doshita, S. (1962). The phonetic typewriter, information processing. In: *Proc. IFIP Congress*, Munich.
45. Sakoe, H. (1979). Two level DP matching – a dynamic programming based pattern matching algorithm for connected word recognition. *IEEE Trans. Acoust., Speech, Signal Process.*, 27, 588–595.
46. Sakoe, H., Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust., Speech, Signal Process.*, 26, 43–49.
47. Shinoda, K., Lee, C. H. (2001). A structural Bayes approach to speaker adaptation. *IEEE Trans. Speech Audio Process.*, 9, 276–287.
48. Soltau, H., Kingsbury, B., Mangu, L., Povey, D., Saon, G., Zweig, G. (2005). The IBM 2004 conversational telephone system for rich transcription. In: *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, Montreal, Canada, I, 205–208.
49. Suzuki, J., Nakata, K. (1961). Recognition of Japanese vowels – preliminary to the recognition of speech. *J. Radio Res. Lab.*, 37 (8), 193–212.

50. Tappert, C., Dixon, N. R., Rabinowitz, A. S., Chapman, W. D. (1971). Automatic recognition of continuous speech utilizing dynamic segmentation, dual classification, sequential decoding and error recovery. Rome Air Dev. Cen, Rome, NY, Technical Report TR 71-146.
51. Varga, P., Moore, R. K. (1990). Hidden Markov model decomposition of speech and noise. In: Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing, Albuquerque, New Mexico, 845-848.
52. Velichko, V. M., Zagoruyko, N. G. (1970). Automatic recognition of 200 words. *Int. J. Man-Machine Studies*, 2, 223-234.
53. Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Kibernetika*, 4 (2), 81-88.
54. Viterbi, J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Inf. Theory*, 13, 260-269.
55. Waibel, A., Hanazawa, T., Hinton, G., Shiano, K., Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust., Speech, Signal Process.*, 37, 393-404.
56. Weintraub, M., Murveit, H., Cohen, M., Price, P., Bernstein, J., Bell, G. (1989). Linguistic constraints in hidden Markov model based speech recognition. In: Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing, Glasgow, Scotland, 699-702.
57. Zue, V., Glass, J., Phillips, M., Seneff, S. (1989). The MIT summit speech recognition system, a progress report. In: Proc. DARPA Speech and Natural Language Workshop, Philadelphia, PA, 179-189.
58. Zweig, G. (1998). Speech recognition with dynamic Bayesian networks. Ph.D. Thesis, University of California, Berkeley.

Chapter 2

Challenges in Speech Synthesis

David Suendermann, Harald Höge, and Alan Black

2.1 Introduction

Similar to other speech- and language-processing disciplines such as speech recognition or machine translation, speech synthesis, the artificial production of human-like speech, has become very powerful over the last 10 years. This is not only due to the following reasons:

- Extensive scientific work rendered by hundreds of speech synthesis researchers worldwide.
- Ever-growing computational capacity. Approaches like unit selection require a significant processor load to be efficient and real-time able.
- Also, the more speech data is available and the higher its resolution is, the better the achieved quality can be. This, however, requires huge memory capacities able to provide random access to the speech data involved. E.g., for the speech data of the European project TC-Star, 10 h of high-fidelity speech with a sampling rate of 96 kHz/24 bit was used [10]. Even nowadays, only few computers are able to hold the whole speech data in their memory. It comprises almost 10 GB.
- Last but not least, as often in the history of science, money plays a key role for the development of certain research fields. In the speech synthesis domain, extensive financial resources are required in particular for generating the speech resources. The rigorous process of producing high-quality recordings of a synthesis voice requires several steps of speaker selection among professional subjects, special recording environments, careful utterance preparation, recording, transcription, annotation, labeling, and pitch tracking. Corpora like the aforementioned 10-h recording or even larger ones (see Section 2.3 for more examples) could only be provided in recent years since the trust in speech synthesis technology had significantly gained and funds could be raised.

D. Suendermann (✉)
SpeechCycle, Inc., 26 Broadway, 11th Floor, New York, NY, USA
e-mail: david@speechcycle.com

As quality and naturalness of speech synthesis significantly improved in the past 10 years and more and more techniques with supposedly superior quality and intelligibility were presented, a demand for a profound comparison between these techniques emerged. Inspired by its sister fields, the speech synthesis community examined the situation in speech recognition and machine translation research. Already early in the 1980s, a considerable number of speech recognition systems showed decent performance, but nobody could reliably tell which one was the best, since all of them had been trained and tested on different corpora and potentially used different performance metrics making it impossible to have a fair comparison. As a consequence, in 1982, Texas Instruments produced a carefully recorded US English multi-speaker corpus with more than 25,000 digit sequences which was to be used as test corpus for the competing parties [29]. At this time, one also agreed on a standard error measure (edit or Levenshtein distance [30]). Although the introduction of standard corpus and measure was a large step toward a fair comparison, several issues had not yet been resolved:

- Since every development team tested their recognizer against the standard corpus in their own laboratory, the numbers they published were not completely trustworthy.
- The same test corpus was used over and over again, and often the developers used it to actually tune their systems, which could result in a performance significantly higher than if the test corpus would have never seen before.
- The usage of a specific test corpus was entirely voluntary. Say, there were five well-recognized corpora, but a specific recognizer performed well only on one of them, there was no need to publish the worse results on the other corpora.

Fortunately, these problems could be resolved by introducing regular evaluation races which were performed by an independent institution. The very first of such competitions was the DARPA Resource Management project launched in 1987 which involved evaluation turns roughly every 6 months [41]. The National Institute of Standards in Telecommunications (NIST) served as independent evaluation institution. For every evaluation turn, a new test corpus was distributed making it impossible to tune on the test set. Also, once being registered for an evaluation, the respective party was expected to submit its recognition results—a withdrawal was considered a failure—and, furthermore, there was no way for the submitting party to predict the performance of its submission before the publication of all competitors' results.

In the mid-2000s, speech synthesis research tried to learn a lesson from the aforementioned developments in speech recognition and other areas by initiating a number of regular competitions meeting the above formulated criteria for objectivity of results including:

- independent evaluation institutions;
- standardized evaluation corpora changing from evaluation to evaluation;
- standardized evaluation metrics.

Being initiated only a few years ago, these competitions constitute a new trend in speech synthesis research. They build a platform joining the most important players in the field; and the frequently held evaluation workshops are discussion forums for the most recent speech synthesis technology. In this function, the competitions discussed in this chapter are the test bed for new trends in the field.

As speech synthesis competitions are based on historic experiences and, at the same time, markers for future trends in synthesis research, this chapter will first deal with the history of this discipline. In a second part, it will outline the most important state-of-the-art techniques and their representation in the scope of speech synthesis competitions.

2.2 Thousand Years of Speech Synthesis Research

2.2.1 *From Middle Ages Over Enlightenment to Industrial Revolution: Mechanical Synthesizers*

The history of speech synthesis, i.e., the artificial production of human-like speech, is presumably much longer than many of the readers might expect, and it is certainly the oldest speech processing discipline discussed in this book. Legends of talking “machines” go 1000 years back to Pope Sylvester II (950–1003 AD) who was supposed to possess a *brazen head* [12]. This kind of prophetic device was reputed to be able to answer any question [11]. In this spirit, it can be regarded to have been the very first dialog system including components of speech recognition, understanding, generation, and, last but not least, synthesis.

Indeed, referring to the brazen head of Sylvester II as the first automatic speech-processing device is as reasonable as calling Icarus’ wings the first airplane. But as much as human’s dream to fly eventually came true, not only magicians, but some centuries later well-reputed scientists like Isaac Newton addressed the production of artificial speech [28]. In 1665, the Fellow of the Royal Society carried out experiments pouring beer into bottles of different shapes and sizes and blowing into them producing vowel-like sounds. This observation was exploited 300 years later for the development of linear predictive coding [33], one of the most important speech analysis and synthesis techniques.

Before the era of linear predictive coding, there were several attempts to build mechanical devices able to produce human-like speech, as for instance the one of Wolfgang von Kempelen [47]. He was a scientist in the service of Empress Maria Theresa in Vienna at the end of the 18th century. Von Kempelen is considered one of the first experimental phoneticians who built several outstanding devices as the *Turk* [44], a supposedly chess-playing machine which later turned out to be a hoax, and his speaking machine [14]. The latter was an apparatus composed of a bellow representing the lungs, a rubber tube for the mouth, and a wooden extension being the nose (for a construction drawing, see Fig. 2.1). By means of two levers controlling the resonance characteristics, a complete set of a language’s sounds could

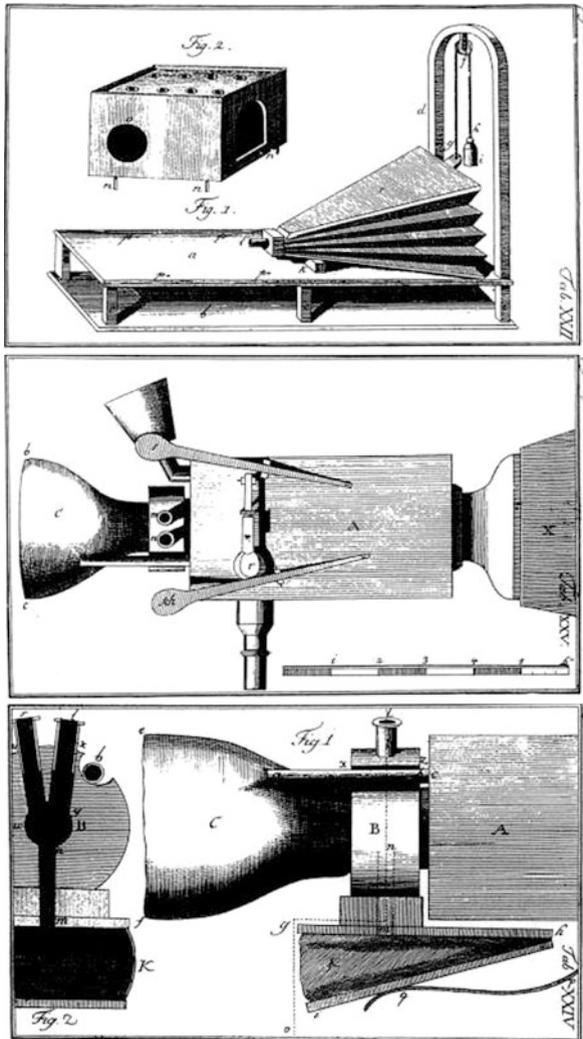


Fig. 2.1 The main components of von Kempelen's speaking machine

be produced. Von Kempelen stated that he successfully produced whole words and sentences in languages such as Latin and French.

In the following years, other inventors produced similar and advanced devices for human-like speech generation, e.g.

- Erasmus Darwin's (the grandfather of the evolutionist) speech synthesizer (1806) [13];
- Joseph Faber's *Euphonia* (1844) [31]; and
- Alexander Graham Bell's (the father of telephony) physical working model of the human vocal tract (end of the 19th century) [3, 17].

2.2.2 The 20th Century: Electronic Synthesizers

The electronic age opened new horizons to speech synthesis technology. After the development of electrical oscillators, filters, amplifiers, loudspeakers, etc., it was possible to generate and control sounds much easier than with mechanical devices whose sound production always was limited to its physical dimensions.

As early as in the year 1936, the UK Telephone Company manufactured an automatic clock, the first instance of practical text-to-speech synthesis [21]. Partial utterances such as numbers and time expressions were stored on a glass disk and were concatenated to form complete sentences.

Only 3 years later, another significant attempt to create a speaking machine (human operated as von Kempelen's) based on electronic sound generation was developed at the Bell Labs (named after the aforementioned Alexander Graham Bell). Homer Dudley's *VODER* (Voice Operated recorDER) was first presented at the 1939 World Fair in New York. It was a keyboard-controlled apparatus composed of a "vocal tract" with 10 parallel band-pass filters spanning the speech frequency range, and an excitation which could be either unvoiced or voiced. Unvoiced sounds were produced by means of a random noise generator; voiced sounds came from a relaxation generator whose fundamental frequency was controlled through a foot pedal. Keyboard and pedal clearly remind of the design of a musical keyboard, and, interestingly, Homer Dudley's visit at Bonn University in 1948 (the year before Bonn became capital of West Germany) inspired Professor Werner Meyer-Eppler to apply synthesis devices such as the *VODER* to music composition pioneering the *Elektronische Musik* movement [27]. In this sense, speech and music synthesizers have the same roots.

All of the attempts to this date had been based on analog signal processing. The development of pulse code modulation (PCM) to transform analog and digital speech representations into each other in the beginning of the 1940s built the foundation of digital speech processing. The very first application of transmitting digitized speech was used during World War II in the vocoder encryption equipment of *SIGSALY*, the secure speech system connecting London and the Pentagon [6].

But PCM was not the only ground-breaking novelty in the mid-20th century. Until the end of the 1950s, experimental research on electronic speech synthesis was focused on specialized devices which were hard-coded and bound to their original purpose. The development of computers, however, gave a lot more flexibility, and as soon as they became strong enough to store and process digitized speech, they were extensively applied to speech-processing tasks.

Again at Bell Labs, probably the first time in history, John L. Kelly used a computer (IBM 704) to create synthetic speech—and not only speech but even a lilting voice singing the song *Bicycle Built for Two*. This incidence later inspired John's friend (and namesake) John Pierce to use this (or a similar) sound sample at the climactic scene of the screenplay for *2001: A Space Odyssey* [48].

At this time, the sound of digital speech synthesis was far from natural, but the clearly recognizable computer voice had its own charm, and synthesis capability was integrated into a number of rather successful electronic devices and computer systems. Very popular was the electronic toy *Speak & Spell*, since 1978

manufactured by Texas Instruments [15]. It contained a single-chip speech synthesizer, the TI TMC0280, based on 10th order linear predictive coding. *Speak & Spell* had a small membrane keyboard and a vacuum fluorescent display and was designed to help young children to become literate, learn the alphabet, and to spell. Again, this synthesizing device played a prominent role in the movies: In Steven Spielberg's motion picture *E.T. the Extra-Terrestrial*, it served as a key component of the alien's home-built interstellar communicator.

Speech synthesis as software or as an integral part of the operating system was introduced in the beginning of the 1980s on computers such as Apple Macintosh or Commodore Amiga.¹ The ever-increasing power of computers witnessed over the last 20 years allowed for the emerging of a wide variety of synthesis techniques whose development and refinement is continued until these days. In the scope of speech synthesis competitions—focus of the remainder of this chapter—many of these techniques are extensively discussed; the most important ones will be revisited in the following section.

2.3 The Many Hats of Speech Synthesis Challenges

2.3.1 Evaluation, Standardization, and Scientific Exchange

In the first years of the new millennium, several groups in Europe, the United States, and Asia met to discuss opportunities for competitions in speech synthesis. Focus elements of all these efforts were:

- the release of *standard speech databases* that would be used by every participant of these challenges to isolate the impact of the speech database from the core synthesis technology,
- the usage of *standard evaluation metrics* to have a fair comparison between the partners and provide an easy measure to express the performance of a system in absolute terms, and
- the involvement of an *independent evaluation party* that would prepare, send out, collect, and evaluate speech synthesis utterances, build an evaluation framework, hire and supervise subjects, and finally report on the evaluation results.

The very first consortium of internationally renowned partners setting up such a framework was the project TC-Star (Technology and Corpora for Speech-to-Speech Translation) funded by the European Commission. Project members were 12 prestigious research institutions from industry and academia. The project's main goal was to significantly reduce the gap between human and machine translation performance. The 36-month project starting on April 1, 2004 was to support basic research in the areas of speech recognition, machine translation, and speech synthesis in the domain of parliamentary and other political speeches.

¹The first author of this chapter used an Amiga 500 computer built in the mid-1980s to synthesize voices for his 1997 drum and bass release *Digital Emperor: Out of O2* [51].

Although the project proposal dates back to as early as 2002 [20], the very first speech synthesis evaluation was conducted only in September 2005. This long delay was not only due to the project being launched in spring 2004, but also due to the very ambitious goal of releasing five multilingual voices for the world's three most frequently spoken languages (Mandarin, English, Spanish). Each of these five voices included 10 h of speech recorded in highly clean studio environments with 96 kHz/24 bit recording precision making the process slow and expensive [10].

The first call for participation to another large-scale initiative on speech synthesis competition also dates back to the year 2004. The Blizzard challenge was specifically founded to compare technology for corpus-based speech synthesis [8]. In doing so, a database that was intentionally designed without the use of copyrighted material was prepared at the Carnegie Mellon University and released free of charge to all participants of the challenge. The first release of the CMU Arctic databases (June 2004) consisted of one female and one male US English voice uttering about 1200 phonetically balanced utterances with a total duration of around 1.5 h each [26]. These corpora were recorded under 16 kHz/16 bit studio conditions and, compared to TC-Star, much faster to record and compile. Consequently, the time span from starting the database preparation and sending to the participants, to the actual evaluation was altogether about a year and a half—the first Blizzard Challenge took place in January 2005, 8 months before the first TC-Star evaluation.

Not only did these competitions differ in the nature of the databases they used, but also the third aforementioned focus of these challenges, the evaluation framework, was treated differently in both instances.

On the one hand, TC-Star took the sister field speech recognition as driving example when using a well-recognized international standardization institution to carry out an independent evaluation as NIST did in the DARPA speech recognition competitions, cf. Section 2.1. The Evaluation and Language Resources Distribution Agency (ELDA, located in Paris, France) took over the responsibility for performing the evaluations including hiring subjects for the subjective assessment and carrying out objective measures if applicable.

On the other hand, the Blizzard challenge relied on in-house resources to perform the evaluation. This was mainly driven by lack of funding for a large independent and human-driven evaluation and by the conviction that, the more subjects participate in the rating, the more trustful the outcomes will be. Also, the evaluation approach itself and the selection of most appropriate subjects were subject to the curiosity of the involved researchers, since it had not yet been fundamentally investigated, to which extend the nature of the subjects is reflected in the outcomes of evaluations. At any rate, in the case of the Blizzard challenge 2005, the set of subjects was composed of:

- *Speech experts.* Every participant in the challenge was to provide 10 experts of its group to contribute.
- *Volunteers.* Random people using the web interface.
- *Undergraduate students.* These people were paid for task completion.

Also early in 2004, undoubtedly the initiation year of speech synthesis challenges, a third consortium was founded to bring together international experts of the speech synthesis community (mostly from Europe and Asia). Main goal of the European Center of Excellence in Speech Synthesis (ECESS) is the standardization, modularization, and evaluation of speech synthesis technology.

This consortium defined a standard text-to-speech system to consist of three main modules: the text processing, the prosody generation, and the acoustic synthesis. Each participant agrees on delivering an executable version of at least one of these three modules following a well-defined interface terminology [42]. Text-to-speech synthesis is considered a plug-and-play system where every partner can focus his research on one module making use of highly performing modules of other parties. To this end, module-wise evaluations are carried out in regular intervals to rate the performance of the evaluating modules.

All three, TC-Star, Blizzard Challenge, and ECESS, have a lot in common. They must not be considered as competing platforms where progress is gained only by trying to best the other participants. Instead, they are scientific forums that are driven by shared resources (such as the corpora in the Blizzard Challenges, or technological modules in ECESS), the grand effort of standardization (corpora, evaluation metrics, module interfaces, etc.), and an intensive dialog in the scope of regular workshops with scientific presentations and proceedings compiling peer-reviewed publications of the respective consortia. Such workshops were held every year during the life cycle of TC-Star. Blizzard Challenges are focus of annual workshops mostly organized as satellite events of major speech conferences such as Interspeech. ECESS workshops are held at least semi-annually since February 2004 in diverse places mostly in Europe. These events quickly became important venues for the major players of the field—the sheer numbers of participants suggests which important role these challenges and their by-products are playing in the speech synthesis community—see Table 2.1.

Table 2.1 Evolution of number of participants in speech synthesis challenges

	TC-Star [10, 36, 37]	Blizzard Challenge [8, 5, 18]	ECESS [42, 23]
2004	3	-	6
2005	3	6	13
2006	10	14	16
2007	6	16	17

2.3.2 Techniques of Speech Synthesis

This section is to give a brief overview about the most popular and successful techniques covered in evaluation campaigns and scientific workshops of speech synthesis challenges.

2.3.2.1 Concatenative Synthesis

As aforementioned, the emergence of powerful computers was main driver for the development of electronic speech synthesis. Mass storage devices, large memory, and high CPU speed were conditions of a synthesis paradigm based on concatenation of real speech segments. The speech of a voice talent is recorded and stored after digitizing. At synthesis time, the computer produces the target speech by cutting out segments from the recording and concatenating them according to a certain scheme.

This principle allows for very simple applications such as a talking clock where an arbitrary time expression is composed of a number of prerecorded utterances. You can imagine that one records the utterances “the time is,” “a.m.,” “p.m.,” and the numbers from 1 to 59 and concatenates them according to the current time. This is indeed a straightforward and practical solution that was already used in the semi-mechanical talking clock by the UK Telephone Company mentioned earlier in this chapter.

Going from a very specialized application to general-purpose text-to-speech synthesis, the above described solution is not feasible anymore. We do not know in advance which words and phrases will be synthesized. Even when one would record every single word in the vocabulary of a language (which is theoretically possible), word-by-word concatenation would sound very fragmented and artificial. In natural speech, words are often pronounced in clusters (without pauses between them), and, most importantly, natural speech features a sentence prosody, i.e., depending on position, role in the sentence, sentence type, etc., word stress, pitch, and duration are altered.

In order to produce synthetic speech following these principles, Hunt and Black [22] proposed the *unit selection* technique. Basically, unit selection allows for using a speech database of arbitrary textual content, phonetically labeled. When a sentence is to be synthesized, the phonetic transcription is produced, and the sentence prosody is predicted. Now, units (of arbitrary length, i.e., number of consecutive phonemes) are selected from the database such that the phonetic transcription matches the target. In doing so, also the prosody of the selected units is considered to be closest as possible to the predicted target prosody. This optimization is carried out as a large search through the space of exponentially many possible selections of units matching the target transcription and is only tractable by applying dynamic programming [4] and beam search [19]. To account for prosody artifacts (pitch jumps, etc.), often, signal processing is applied to smooth concatenation points.

Over the years after the introduction of unit selection, there have been major improvements to optimize the quality of the output speech such as:

- paying extreme attention to a superior quality of the speech recordings to make them consistent, having accurate phonetic labels and pitch marks, highest possible SNR, etc. (see the careful recording specifications for TC-Star [10]),
- reducing involved signal processing as much as possible by producing a speech corpus that as best as possible reflects the textual and prosodic characteristics of the target speech [7],

- enlarging the corpus, since this increases the probability that nice units will be found in the database (see, e.g., the 16-h corpus ATRECCS provided for the Blizzard Challenge 2007 [39]),
- improving prosody models to produce more natural target speech [43].

Unit selection is definitely the most popular text-to-speech synthesis technique to date as for instance the submissions to the Blizzard Challenge 2007 show: 14 out of 15 publications at the evaluation workshop dealt with unit selection synthesis.

2.3.2.2 HMM-Based Synthesis

Hidden Markov models (HMMs) [1] were used since the mid-1970s as a successful approach to speech recognition [45]. The idea is that constant time frames of the PCM-encoded acoustic signal are converted into feature vectors representing the most important spectral characteristics of the signal. Sequences of feature vectors are correlated to phoneme sequences mapped to the underlying text through a pronunciation lexicon. The HMM models the statistical behavior of the signal and is trained based on a significant amount of transcribed speech data. In recognition phase, it allows for estimating the probability of a feature vector sequence given a word sequence. This probability is referred to as *acoustic model*.

Besides, HMMs allow for estimating the probability of a word sequence, called the *language model*. In this case, the training is carried out using a huge amount of training text.

The combination of acoustic and language model according to Bayes' theorem [2] produces the probability of a word sequence given a feature vector sequence. By trying all possible word sequences and maximizing this probability, the recognition hypothesis is produced.

In speech synthesis, this process has to be reversed [35]. Here, we are given the word sequence and search for the optimal sequence of feature vectors. Since the word sequence is known, we do not need the language model in this case. For applying the acoustic model, we first have to convert the word sequence into a phonetic sequence using pronunciation lexicon or (in case of an unknown word) grapheme-to-phoneme conversion (as also done in other synthesis techniques such as unit selection). The phoneme sequence is then applied to the acoustic model emitting the most probable feature sequence as described, e.g., in [9].

In contrast to the application to speech recognition, in synthesis, the consideration of spectral features alone is not sufficient, since one has to produce a wave form to get audible speech. This is done by generating a voiced or unvoiced excitation and filtering it by means of the spectral features on a frame-by-frame basis. This procedure is very similar to the VODER discussed in Section 2.2. To overcome the limitation to just two different excitation types resulting in rather synthetic and muffled speech, there are several approaches to more sophisticated excitation modeling like, for instance, the harmonic + noise model [50], residual prediction [52], or mixed excitation with state-dependent filters [32].

Convincing advantage of HMM-based speech synthesis is its small footprint. As this technique only requires the model parameters to be stored (and not the speech data itself), its size is usually limited to a few megabytes. In contrast, the above-mentioned 16-h synthesis corpus ATRECCS, sampled at 48 kHz and a final 16 bit precision requires more than 5 GB of storage.

In addition, HMM-based synthesis requires less computation than unit selection. This is because the latter searches a huge space becoming larger and larger as more speech data are available. Real-time ability can be an issue with such systems, whereas HMM-based synthesis operates at much lower expense.

Last but not least, HMM-based synthesis has the potential to produce high-quality speech. This is mainly because HMMs produce continuous spectral contours as opposed to unit selection synthesis where we may face inconsistencies at the concatenation points resulting in audible speech artifacts. Blizzard Challenge 2005 delivered the proof that HMM-based synthesis is even able to outperform the well-established unit selection technique, as the contribution [55] achieved the best results in both speech quality and intelligibility. This, however, was partially because the training corpus (the Arctic database as introduced above) contained the relative small amount of 1.5 h of speech. Unit selection synthesis may face unit sparseness problems with too small databases lacking appropriate data for the target utterances: In the Blizzard Challenge 2006 which provided a 5-h corpus, the best system was based on unit selection synthesis [25].

2.3.2.3 Voice Conversion

In the above sections, we emphasized the importance of large and carefully designed corpora for the production of high-quality speech synthesizers. The compilation of corpora such as those built for TC-Star or the ATRECCS database incorporates a significant time effort as well as high monetary expenses. This explains that, usually, synthesizers come along with at most two or three voices for a given language constituting a lack of variety and flexibility for the customer of speech synthesis technology.

Voice conversion is a solution to this problem. It is a technology that allows for rapidly transforming a source voice into a target voice [38]. Statistical approaches to voice conversion discussed since the mid-1990s [49, 24] are the most popular ones. These approaches link parallel source and target speech by means of spectral feature vectors and generate a joint statistical model, usually a Gaussian mixture model (GMM), representing the relation between both voices. In this context, *parallel speech* means that both source and target speaker uttered the very same text preferably with a similar timing pattern. A one-to-one speech frame mapping is then produced by alignment techniques such as dynamic time warping [46] or HMM-based forced alignment [54]. This approach is called *text dependent*, since it requires the target speaker to utter the same text as the source speaker. In contrast, *text-independent* techniques provide flexibility regarding the target speech as required if voice conversion is to be applied on earlier recorded databases, including the case

that both voices use different *languages*. In this case, one speaks of *cross-language* voice conversion [34].

The application of voice conversion to speech synthesis was one of the focuses of the TC-Star project dedicating entire workshop sessions to this topic, e.g., the 2006 workshop in Barcelona dealt with GMM-based [40], text-independent [16], as well as cross-language voice conversion [53].

2.4 Conclusion

Competitions (or challenges) in speech synthesis emerging only few years ago are playfields for research and industry playing several roles at once.

- Originally, they were mainly intended to serve as a platform where different synthesis techniques could be compared.
- To design this comparison as objective as possible, standard corpora, standard evaluation measures, and a standard evaluation framework were proposed. In this way, and particularly as the number of participants grew significantly, challenges became forums for standardization of corpora, evaluation criteria, and module interfaces.
- Challenge-related workshops, conferences, and meetings evolved from pure result-reporting events to scientific venues with talks, poster presentations, and published proceedings, hence a platform for the exchange of ideas.
- They serve challenges not only for the exchange of ideas but also for the exchange of technology. Special agreements between participants led to a publication of significant parts of used resources and software. Corpora are made available for download in the Internet; source codes and binaries of entire speech synthesizers, modules, tools, and evaluation kits are distributed.

The authors believe that a major part of future speech synthesis research will be carried out in the scope of such challenges leading to a boost of quality and applicability of speech synthesis solutions and a strengthening of the whole research field.

References

1. Baum, L., Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statistics*, 37, 1554–1563.
2. Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. Roy. Soc. Lond.*, 53, 370–418.
3. Bell, A. (1922). Prehistoric telephone days. *Natl. Geographic Mag.*, 41, 223–242.
4. Bellmann, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, USA.
5. Bennett, C., Black, A. (2006). The Blizzard Challenge 2006. In: *Blizzard Challenge Workshop*, Pittsburgh, USA.

6. Bennett, W. (1983). Secret telephony as a historical example of spread-spectrum communications. *IEEE Trans. Commun.*, 31(1), 98–104.
7. Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., Syrdal, A. (2006). The AT&T Next-Gen TTS system. In: Proc. TC-Star Workshop, Barcelona, Spain.
8. Black, A., Tokuda, K. (2005). Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In: Proc. Interspeech, Lisbon, Portugal.
9. Black, A., Zen, H., Tokuda, K. (2007). Statistical parametric synthesis. In: Proc. ICASSP, Honolulu, USA.
10. Bonafonte, A., Höge, H., Tropsch, H., Moreno, A., v. d. Heuvel, H., Sündermann, D., Ziegenhain, U., Pérez, J., Kiss, I. (2005). TC-Star: Specifications of language resources for speech synthesis. Technical Report.
11. Butler, E. (1948). *The Myth of the Magus*. Cambridge University Press, Cambridge, UK.
12. Darlington, O. (1947). Gerbert, the teacher. *Am. Historical Rev.*, 52, 456–476.
13. Darwin, E. (1806). *The Temple of Nature*. J. Johnson, London, UK.
14. Dudley, H., Tarnoczy, T. (1950). The speaking machine of Wolfgang von Kempelen. *J. Acoust. Soc. Am.*, 22(2), 151–166.
15. Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht, Netherlands.
16. Duxans, H., Erro, D., Pérez, J., Diego, F., Bonafonte, A., Moreno, A. (2006). Voice conversion of non-aligned data using unit selection. In: Proc. TC-Star Workshop, Barcelona, Spain.
17. Flanagan, J. (1972). Voices of men and machines. *J. Acoust. Soc. Am.*, 51, 1375–1387.
18. Fraser, M. King, S. (2007). The Blizzard challenge 2007. In: Proc. ISCA Workshop on Speech Synthesis, Bonn, Germany.
19. Hand, D., Smyth, P., Mannila, H. (2001). *Principles of Data Mining*. MIT Press, Cambridge, USA.
20. Höge, H. (2002). Project proposal TC-STAR – Make speech to speech translation real. In: Proc. LREC, Las Palmas, Spain.
21. Holmes, J., Holmes, W. (2001). *Speech Synthesis and Recognition*. Taylor and Francis, London, UK.
22. Hunt, A., Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. ICASSP, Atlanta, USA.
23. Kacic, Z. (2004–2007). Proc. 11th–14th Int. Workshops on Advances in Speech Technology. University of Maribor, Maribor, Slovenia.
24. Kain, A., Macon, M. (1998). Spectral voice conversion for text-to-speech synthesis. In: Proc. ICASSP, Seattle, USA.
25. Kaszczuk, M., Osowski, L. (2006). Evaluating Ivona speech synthesis system for Blizzard Challenge 2006. In: Blizzard Challenge Workshop, Pittsburgh, USA.
26. Kominek, J., Black, A. (2004). The CMU arctic speech databases. In: Proc. ISCA Workshop on Speech Synthesis, Pittsburgh, USA.
27. Kostelanetz, R. (1996). *Classic Essays on Twentieth-Century Music*. Schirmer Books, New York, USA.
28. Ladefoged, P. (1998). *A Course in Phonetics*. Harcourt Brace Jovanovich, New York, USA.
29. Leonard, R., Doddington, G. (1982). *A Speaker-Independent Connected-Digit Database*. Texas Instruments, Dallas, USA.
30. Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Phys. Dokl.*, 10, 707–710.
31. Lindsay, D. (1997). Talking head. *Am. Heritage Invention Technol.*, 13(1), 57–63.
32. Maia, R., Toda, T., Zen, H., Nankaku, Y., Tokuda, K. (2007). An excitation model for HMM-based speech synthesis based on residual modeling. In: Proc. ISCA Workshop on Speech Synthesis, Bonn, Germany.
33. Markel, J., Gray, A. (1976). *Linear Prediction of Speech*. Springer, New York, USA.
34. Mashimo, M., Toda, T., Shikano, K., Campbell, N. (2001). Evaluation of cross-language voice conversion based on GMM and STRAIGHT. In: Proc. Eurospeech, Aalborg, Denmark.

35. Masuko, T. (2002). HMM-based speech synthesis and its applications. PhD thesis, Tokyo Institute of Technology, Tokyo, Japan.
36. Mostefa, D., Garcia, M.-N., Hamon, O., Moreau, N. (2006). TC-Star: D16 Evaluation Report. Technical Report.
37. Mostefa, D., Hamon, O., Moreau, N., Choukri, K. (2007). TC-Star: D30 Evaluation Report. Technical Report.
38. Moulines, E. and Sagisaka, Y. (1995). Voice conversion: State of the art and perspectives. *Speech Commun.*, 16(2), 125–126.
39. Ni, J., Hirai, T., Kawai, H., Toda, T., Tokuda, K., Tsuzaki, M., Sakai, S., Maia, R., Nakamura, S. (2007). ATRECSS – ATR English speech corpus for speech synthesis. In: Proc. ISCA Workshop on Speech Synthesis, Bonn, Germany.
40. Nurminen, J., Popa, V., Tian, J., Tang, Y., Kiss, I. (2006). A parametric approach for voice conversion. In: Proc. TC-Star Workshop, Barcelona, Spain.
41. Pallet, D. (1987). Test procedures for the March 1987 DARPA Benchmark Tests. In: Proc. DARPA Speech Recognition Workshop, San Diego, USA.
42. Pérez, J., Bonafonte, A., Hain, H.-U., Keller, E., Breuer, S., Tian, J. (2006). ECESS inter-module interface specification for speech synthesis. In: Proc. LREC, Genoa, Italy.
43. Pfitzinger, H. (2006). Five dimensions of prosody: Intensity, intonation, timing, voice quality, and degree of reduction. In: Proc. Speech Prosody, Dresden, Germany.
44. Poe, E. (1836). Maelzel’s Chess Player. *Southern Literary Messenger*, 2(5), 318–326.
45. Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2), 257–286.
46. Rabiner, L., Rosenberg, A., Levinson, S. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Trans. Acoustics, Speech Signal Process.*, 26(6), 575–582.
47. Ritter von Kempelen, W. (1791). Mechanismus der menschlichen Sprache nebst der Beschreibung einer sprechenden Maschine. J. V. Degen, Vienna, Austria.
48. Stork, D. (1996). HAL’s Legacy: 2001’s Computer as Dream and Reality. MIT Press, Cambridge, USA.
49. Stylianou, Y., Cappé, O., Moulines, E. (1995). Statistical methods for voice quality transformation. In: Proc. Eurospeech, Madrid, Spain.
50. Stylianou, Y., Laroche, J., Moulines, E. (1995). High-quality speech modification based on a harmonic + noise model. In: Proc. Eurospeech, Madrid, Spain.
51. Suendermann, D., Raeder, H. (1997). Digital Emperor: Out of O2. d.l.h.-productions, Cologne, Germany.
52. Sündermann, D., Bonafonte, A., Ney, H., Höge, H. (2005). A study on residual prediction techniques for voice conversion. In: Proc. ICASSP, Philadelphia, USA.
53. Sündermann, D., Höge, H., Bonafonte, A., Ney, H., Hirschberg, J. (2006). TC-Star: Cross-language voice conversion revisited. In: Proc. TC-Star Workshop, Barcelona, Spain.
54. Young, S., Woodland, P., Byrne, W. (1993). *The HTK Book, Version 1.5*. Cambridge University Press, Cambridge, UK.
55. Zen, H., Toda, T. (2005). An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In: Proc. Interspeech, Lisbon, Portugal.

Chapter 3

Spoken Language Dialogue Models

Kristiina Jokinen

3.1 Introduction

Spoken language interactive systems range from speech-enabled command interfaces to dialogue systems which conduct spoken conversations with the user. In the first case, spoken language is used as an alternative input and output modality, so that the commands, which the user could type or select from the menu, may also be uttered. The system responses can also be given as spoken utterances, instead of written language or drawings on the screen, so the whole interaction can be conducted in speech. Spoken dialogue systems, however, are built on models concerning spoken conversations between participants so as to allow flexible interaction capabilities. Although interactions are limited concerning topics, turn-taking principles and conversational strategies, the systems aim at human–computer interaction that would support natural interaction which enables the user to interact with the system in an intuitive manner. Moreover, trying to combine insights of the processes that underlie typical human interactions, spoken dialogue modelling also seeks to advance our knowledge and understanding of the principles that govern communicative situations in general.

The terms *interactive system*, *dialogue system* and *speech-based system* are often used interchangeably. In order to restrict attention to a particular type of interactive systems, we distinguish the terms in this chapter as follows. *Interactive system* is used as a generic term which refers to any automatic machine that the user can interact with. Interaction is usually conducted using natural language, although various non-verbal modalities such as gestures, facial expressions and gazing can also be used as means of interaction: it is possible to have multimodal interactive systems. *Dialogue system* is a subtype of interactive systems such that it includes a particular dialogue management component, the dialogue manager, which handles the interaction between the user and the system. The Dialogue Manager does not only manage interaction, but can also model the participant's communicative

K. Jokinen (✉)

University of Helsinki, Helsinki, Finland; University of Tartu, Tartu, Estonia
e-mail: kristiina.jokinen@helsinki.fi

goals, beliefs and preferences, and encode rather complicated reasoning rules for cooperation and communicative principles. It creates abstract representations from the user input, and the representations can be further processed by the other system components. Thus the input in dialogue systems is manipulated on different analysis levels which determine the appropriate output depending on the particular dialogue context, unlike in human-machine *interfaces* which typically relate the input to the output directly. Finally, the term *speech-based system* refers to interactive systems with a specific interaction mode, i.e. speech; likewise, it is possible to specify gesture-based systems, tactile systems, etc., if the main mode of interaction is conducted using gestures, tactile interfaces, etc. In industrial-related research it is also common to talk about voice user interfaces (VUI) as opposed to graphical user interfaces (GUI), if the emphasis is on speech rather than on mouse-and-menu interface. Analogously, a *spoken dialogue system* is a dialogue system which deals with speech input and output. If there is a need to emphasise conversational capabilities of the system as opposed to simple commands, the term (*spoken*) *conversational system* is commonly used.

This chapter provides a short introduction to dialogue systems and dialogue research in general. An overview of the development of dialogue systems from the industrial point of view is given in Chapter 4, and the other chapters specify particular research topics in more detail.

3.2 Historical Overview

The idea of an intelligent interactive system goes back to the 1940s when the modern computer was developed. Initially, interaction was associated with a science fiction type view of the computer possessing near human-like communicative capabilities, but the task has been much harder. The richness of human communication does not easily lend itself to simple algorithmic models, and the goal also requires advanced engineering and modelling technology. However, the knowledge of the functioning of human-human as well as human-computer interaction has increased to support the technology and to drive research with new observations and models, and in practice, the work has progressed gradually to the current state of dialogue technology where spoken dialogue systems are available for certain well-defined task domains, and even commercially feasible. Often the goal of the research and system development is described as to build applications with improved interaction capabilities and advanced technology, and one of the main vehicles for this is to furnish the interface with richer and more natural communication possibilities, including both language and multimodal interaction. Although the original ideas of human-human like communication may not be entertained as such, we can say that much of the research activity within the dialogue system design and development is implicitly related to the original AI goal of realising a system which could assist humans in their various tasks via intelligent interaction.

The 60-year history span of dialogue system research and development can be roughly divided into four eras which cover the development from the early ideas (during 1950–1980), via experimental prototypes in the 1980s and large-scale dialogue projects in 1990s, to dialogue technology and ubiquitous computer challenges in the new millennium (2000–). We briefly discuss these below, and for a more comprehensive overview refer to [66].

3.2.1 *Early Ideas of a Thinking Machine*

The idea of human–computer interaction can be traced back to Alan Turing who in his influential article sketched a *Thinking Machine* [138]. This was an abstract machine which would operate according to complex algorithms and produce behaviour which could be described as intelligent. As a sign of its intelligence, it could, e.g. be engaged in natural language conversations. Turing proposed a test, later to be called the Turing test, to judge whether such a Thinking Machine would indeed be an intelligent entity, despite the fact that its observed behaviour would be produced by following algorithmic, albeit complex, rules: if the machine’s behaviour could not be distinguished from that of humans, then it could be judged as intelligent. Of course, these early discussions in the beginning of the modern artificial intelligence research cannot be considered dialogue modelling as such, since the main interest lay in the logic-philosophical pondering and in engineering efforts concerning the computer’s possible intelligence and how, if at all, such intelligent behaviour would differ from the one exhibited by humans. However, it is worth noticing that intelligent behaviour was tied to verbal communication: the capability to express oneself in words was understood as an inherent sign of intelligence.

During the early years, studies in formal grammars and language structures [28], semantics [49] and pragmatics [13, 124] were crucial in developing interactive systems based on natural language. Moreover, significant development took place in the AI research which was relevant also for the dialogue system research and development later on. For instance, symbolic representation and manipulation [105], frames [96], scripts [121] and dialogue focus [50] were important topics also for dialogue system development.

The first interactive system, and probably also one of the most famous conversational computer programs, was the Eliza system [148]. It plays the role of a psychotherapist and can fool the user for some time acting as a sympathetic conversational partner who asks relevant questions concerning the user’s problems. Eliza provides the basis for the present-day chatbots, although of course it was quite different from the creatures that now populate the internet enabled by web-technology. From the dialogue modelling point of view, however, Eliza (and similar programs) has no real understanding of utterances or conversational principles. It uses pattern matching techniques and randomly selects a suitable response from its database depending on the keywords found in the user’s input. The user soon realises that the dialogue leads nowhere: the questions are of the same type, and the system seems

to contradict with what the user just said or what the system has said previously in the interaction. What is missing is the mechanism to take care of the speakers' intentions and the coherence of the dialogue.

The SHRDLU-program [151] was a milestone in the sense that it showed that automatic natural language processing was indeed possible. The program could interpret natural language commands with respect to a small block world, disambiguate references, and ask the user questions. Also systems such as LUNAR [152] and HEARSAY-II [42] demonstrated natural language-based interaction, and Power [111] showed that simulated interactions were possible using planning techniques. As a result of the early exploratory work, the development of dialogue systems could be undertaken more realistically.

3.2.2 Experimental Prototypes and Dialogue Models

Research in automatic language understanding, discourse processing and dialogue systems gained momentum in the 1980s, and extensive investigations within specific research frameworks started both in Europe and in the United States. The idea of the Thinking Machine faded in the background as the activities were driven by exploring the new and exciting possibilities that formal modelling and computer technology provided for consolidating different theoretical aspects of language communication, and enabling automatic interaction. Research prototypes such as HAM-ANS [140], EES [97] and MINDS [154] were built with research activities especially focussing on natural language processing, text coherence and discourse structure. One of the first multimodal systems was also built at that time: the Put-That-There-system [17] allowed users to interact with the world through the projection on the wall by using speech and pointing gestures. In Japan, the Ministry of International Trade and Industry announced the Fifth Generation Computer Systems programme in 1982 to develop massive parallel computers, and this also supported research on natural language engineering [98].

The research emphasised deep understanding of the language and the world, and theories and models for language and dialogue were implemented and tested using computational means. Necessary rules and data were designed and produced manually, and the feasibility of the algorithms and processes was considered strong support for the correctness of the particular idea or theory. Cooperation, rational communication, speech acts and action planning were among the core topics in the AI-community, and interests focused on the BDI (Belief-Desire-Intention) agents and how their intentions and beliefs drive communication and action: how the agents build plans and how these plans give rise to their actions, of which communication is one type, and how cooperation is based on the agents' rational action (see overview in [15, 36]). Various aspects of discourse processing were also investigated, such as coherence and discourse relations [61, 62, 113], cooperative answers [69] and misconceptions [91]. Grosz and Sidner [53] developed a discourse theory which combines three independent, yet interconnected descriptive levels: linguistic

level, intentional level and attentional state. Each level uses its own organisation that captures different aspects of discourse (and dialogue) management. Coherence is conveyed by language expressions, while the speakers' intentions and the discourse purposes are modelled in the intentional level, and the attentional state encodes focus spaces associated with discourse segments and the most salient objects in them. Together the levels determine the well-formedness of the discourse (and the dialogue).

The research also focused on user modelling so as to help the system provide information that would be appropriate to the user's level of expertise [27, 109]. Also the generation of dialogue contributions was studied [11], although much of the generation research focused on text-generation, see e.g. [155]. Clark and colleagues [31, 32] developed a model of how speakers establish common ground through their contributions and collaborate in utterance production.

3.2.3 Large-Scale Projects: From Written to Spoken Dialogues

When approaching 1990s, research results on dialogues and dialogue models accumulated, together with the advanced technical level of the computers. It became feasible and practical to form large-scale dialogue projects which aimed at integrating the natural language front-end with the dialogue management component into a working dialogue system. Funding frameworks facilitated the launching of large-scale projects where the goal was not only to highlight some particular aspects of communication, but to build a practical dialogue system that would demonstrate how the techniques would work in limited contexts.

The European projects SUNDIAL [92] and PLUS [16] implemented communicative principles, pragmatic reasoning and planning techniques into their dialogue managers, with applications dealing with Yellow Pages and flight information. Another European project ARISE [12] focused on a train timetable system for Dutch, French and Italian, and studied different dialogue strategies and technologies. On the national level, large dialogue projects were also funded. For instance, the French Telecom project produced ARTIMIS, a dialogue system that is based on strict logical reasoning about the participants' beliefs and intentions [119], while the German national project *Verbmobil* [139] integrated various spoken language and dialogue-processing techniques into a system that allowed users to agree on meeting dates and times using spoken language. The system also monitored the users interaction so as to be available for assistance if needed. The KTH Waxholm project [26] was an experimental spoken dialogue system that gave information about boat traffic in the Stockholm archipelago.

In the United States, the national research initiative DARPA Communicator supported large-scale dialogue system research and was crucial in establishing speech-based, multimodal interaction technology. Focussing on the system architecture and dialogue management components, the project produced the Communicator architecture for dialogue system development (Galaxy Communicator [125]; CMU

Communicator [117]). The Open Agent Architecture (OAA, [90]) was developed at SRI (Stanford Research Institute). This is a flexible distributed computer architecture which was used, e.g., in the CommandTalk spoken dialogue interface to a battlefield simulator [130] and in the Witas project [79] to develop a multimodal dialogue system to instruct and monitor mobile robots.

The TRAINS project [4] and its follow-up TRIPS [2] started from the BDI-agent view point and focused on plan-based dialogue modelling. The task of the systems was to negotiate suitable routes for certain train transports, and this required an elaborated dialogue model with the possibility to recognise the partner's goals and to plan one's own responses accordingly. The TRIPS architecture has a special module, Behavioral Agent, that takes care of the planning of the system's actions, and it also features other aspects, typical especially for conversational dialogue systems, such as various types of communicative acts, speaker intentions, discourse referents and the process of grounding of utterances [136]. Also Circuit-Fix-It-Shop-system [129] dealt with dialogue management in collaborative problem-solving situation where the agents possess complementary information and have to initiate communication to complete the task. Guinn [55] demonstrated how this kind of mixed-initiative negotiation can be implemented using logical problem-solving techniques. The goal of the agent is to prove that the task is done, but if the agent's knowledge is not sufficient for a logical proof, the Missing Axiom principle directs the agent to look for the missing information by asking the partner, and the process continues until it can be proved that the task is successfully resolved.

In Japan, much of the spoken dialogue research concerned engineering issues and building systems that also integrated speech technology. For instance, at NTT Basic Research Laboratories, the WIT toolkit was developed for real-time dialogue responding, and demonstrated the talking head, Noddy, which gave feedback to the speaker by nodding on conversationally adequate points [58, 102]. Research activities at ATR (Advanced Telecommunications Research Laboratories), the highly international research institute in Kyoto, supported several different aspects of spoken dialogue technology, from speech recognition and synthesis to dialogue management and machine translation. The work culminated in the MATRIX, a speech-to-speech translation system that exemplified how multilingual hotel reservations could be made via a spoken language interface [134]. The system incorporated dialogue research especially in the utterance disambiguation. The project was part of the CSTAR consortium, twinning for instance with the Verbmobil project and the CMU JANUS [132].

The research projects also initiated dialogue system evaluation. It became important to assess the usability of the integrated systems as a whole, and also to enable comparison of different technologies and approaches concerning their applicability to a particular task. For instance, the EAGLES evaluation group surveyed different areas in language technology [38], and emphasised that system evaluation is needed for forming and refining system design, and assessing the impact, usability and effectiveness of the overall system performance. In Japan, the DiaLeague initiative [56] aimed at objective and automated evaluation using simulated dialogues between two systems that were to find a path between two stations using their slightly different maps of the possible routes.

Various evaluations of dialogue systems were reported [40, 112, 128], and the PARADISE framework was described in [142, 143] as a general paradigm for evaluating dialogue systems. The overall objective is to maximise user satisfaction by maximising task success and minimising dialogue cost. The impact of the various evaluation parameters on the user satisfaction is calculated by linear regression which produces a function that characterises the system performance. The function also allows comparison of different systems, given they use the same evaluation parameters, in an objective way. However, the framework requires a lot of manual preparation, and it has also been criticised for mixing the system designer's and the system user's view points which do not necessarily coincide. Moreover, the attribute-value type task representation is applicable only to tasks that have a small number of parameters with a fixed set of values, i.e. not good for modelling tasks such as negotiation dialogues.

In the beginning of the 1990s dialogue systems mainly used written language, but the developments in speech technology (see Chapter 1) made spoken language the predominant input/output modality for the dialogue systems in the latter part of the decade. Besides technological challenges in integrating spoken language and dialogue systems, speech also brought in new phenomena for the research, and largely introduced statistical and probabilistic modelling as techniques for spoken dialogue research too. For instance, spoken language studies have dealt with discourse-related phenomena such as using prosodic cues for discourse and topic segmentation [51, 60, 104] and disambiguating cue-phrases [59], and research has also focused on various conversational speech phenomena, such as error handling and corrections [57, 78, 103, 133], the role of syntax and prosody in turn-taking [77], incremental understanding of spoken utterances [101] and non-verbal feedback [58, 145]. Prosody has also been used to recognise the speaker's intentions in recognising dialogue acts [84, 126].

3.2.4 Dialogue Technology: Industrial Perspectives

In the beginning of new millennium, spoken dialogue research was mature enough to be deployed by commercial enterprises (see Chapter 4). Compared with the academic research, industrial settings brought forward other preferences than just modelling the internal workings of interaction: the main goal was to apply dialogue technology to commercially viable products and thus to build robust applications which would enable automatic interaction between users and particular service like banking, travel information and call routing. This set both limits and challenges for dialogue research: research geared more towards systems, techniques, and improved performance, but also towards the user's role in the interaction cycle and the possibility to use other modalities besides speech.

When the experimental research was applied to real usage situations, it was also observed that the users behaved quite differently in real situations than in laboratory settings. For instance, real users appeared to used more barge-ins and help requests than the recruited subjects who, on the other hand, talked more and faster than the

real users [1]. Hence, besides the requirement of the system having to work in a robust manner with different users, the design and development also had to take the genuine needs and motivations of the users into consideration.

The users of commercial dialogue applications and services expect fast and reliable performance that helps them in their particular task. Thus usability evaluations are important: the dialogue system should not only function in the required manner, but it should provide reliable functioning, be easy to learn and enjoyable to use, i.e. it should score high in user satisfaction. Moreover, the system should have value to the users so that they are willing to spend time and effort in learning how to use the system, and also be committed to (buying and) using the system.

In the industrial context, the modelling of complex dialogue phenomena was thus replaced by careful dialogue design. This aims at crafting system prompts and utterance sequences so as to provide robust technology for the questions and answers that are typical of a given application domain. One of the techniques is “How may I help you?” [48, cf. also 30] which allows users to state their question in natural language, and uses then pattern matching and machine-learning techniques to further specify the query and match it with an appropriate problem class that the user may want to get help about. Dialogue design also aims at overcoming technological shortcomings and reducing user frustration: system prompts should guide the user to respond in a particular way which helps the system to recognise user utterances and to keep track of the underlying task. For instance, Yankelovich [153] emphasized that it is important that the system responses that guide the user forward with the task completion and also provide help in case of problems. She suggested different design strategies to cope with different users so that the interaction appears fluent and the system can guide the user to a successful completion of the task.

In order to enable rapid prototyping and compatibility with other modules, also standards and tools have been under active development. Besides the ISO (International Standardisation Organisation) framework, the work is also carried out, e.g., in X+V and SALT communities, concerning component interfaces and scripting languages, and also in the W3C consortium which focuses on standard architectures and languages especially for web applications. VoiceXML is an example of a widely used industrial standard for interaction management, while representation languages such as XML, RDF and EMMA are used to provide standard representations for the data from multiple knowledge sources, and thus make portability of system components easier.

As already mentioned, machine-learning techniques were established to dialogue research largely through contacts with the speech technology community where speech recognisers had been developed using statistical and probabilistic techniques. Since hand-crafted rules can be difficult as well as time and resource-consuming to design, machine-learning techniques are seen as an alternative that provides a quick engineering approach to building natural language interfaces. For instance, reinforcement learning has been used to learn optimal dialogue strategies [83, 87, 141], cf. also experiments with chatbots [71]. The requirements for annotated data and a large enough training set limit the use of the techniques, and in some cases it may, in fact, be faster to hand-craft symbolic rules than to try to learn

correlations through data. In a more comprehensive statistical dialogue management framework, reinforcement learning has been used for user simulation and for optimising and comparing different dialogue management strategies [116, 122, 123, 150]. However, there are still a number of issues that need to be addressed when applying reinforcement learning in practical dialogue system development [108].

3.2.5 Current Trends: Towards Multimodal Intelligent Systems

Current dialogue technology allows us to build dialogue systems which can interact with the user in a fairly robust albeit fixed manner. However, many research questions are still open, and technology is improved towards systems with richer and more flexible input and output possibilities. For instance, dialogue systems accept a rather limited set of spoken utterances so techniques for error handling as well as improved conversational speech recognition are actively investigated. Also the question-answering patterns based on straightforward task structures can be extended towards more conversational management models that allow users to be engaged in negotiations or even friendly chatting if necessary. Moreover, other modalities than speech are also getting integrated in the systems. Natural communication is conveyed by non-verbal means such as facial expressions, hand gestures and body posture, which not only add to the naturalness of communication but also invite engineering efforts to develop new techniques and interface technologies. Much of the current dialogue research thus focuses on multimodal dialogue systems [88] and non-verbal communication [43]. Also less usual interaction modalities such as eye-gaze [89] are topics of concern: The novel interaction techniques also allow interaction for users who cannot use keyboard, mouse and/or speech, and they thus support universal access to digital data (cf. Chapter 14). Moreover, multimodal interaction is relevant in human-robot communication where the robot acts on the basis of the information that it has received from the world through its sensory channels [18, 76]. Robot's action and communication depend on the varying settings of the environment and on the robot's observations about the situation, including the human instructions given in speech, gestures, etc.

Another important research direction focuses on social aspects of communication, concerning multiparty interactions and the use of language as a means of constructing our social reality. Communication among groups of people is different from two-person dialogues, since social dynamics of the group influences the individual behaviour. Group opinions and decisions tend to emerge from social interaction with certain individuals being the opinion setters, others being followers, others mediators, etc. [137]. Moreover, non-verbal communication (gesturing, body posture, [74]) plays an important role in the controlling and coordination of interaction besides explicit verbal utterances.

Finally, current research also focuses on modelling techniques and methods, as well as semi-automatic evaluation of the systems and technologies. Work also continues on architectures and tools that provide open source platforms for system development and evaluation.

We will return to research directions in the last section, but will first discuss topics related to dialogue models aiming to enable more natural interactions.

3.3 Dialogue Modelling

It is useful to distinguish dialogue modelling from the techniques for dialogue control, although the concepts come close to each other in practical dialogue systems. By dialogue control we refer to computational techniques used to implement the system's interaction management, whereas the goal of the dialogue modelling is to produce generalisations and models about communicative situations in which natural language is used. Computational dialogue modelling aims at models which can also be implemented in dialogue systems, and thus be used as the basis for the system's dialogue management. Below we refer to the latter type of models with the term "discourse model", in order to emphasise the fact that they have not been exclusively developed for spoken communication, but also applied to written language, i.e. texts.

3.3.1 Dialogue Management Models

Different dialogue management techniques can be distinguished for the implementation of dialogue control [94]. The simplest technique is the scripted dialogue management model which defines appropriate actions at each dialogue state as a kind of a predefined script. The states and state transitions can be described with the help of particular scripting languages, such as VoiceXML, which can also include subroutines that enable fairly sophisticated dialogue structures. However, the scripting technique does not distinguish procedural dialogue knowledge from the static task structure, and each possible dialogue action must be explicitly defined for each task step in the script. The approach becomes untenable, if complex interactions are to be modelled with several parameters and states.

A more flexible technique is to separate the procedural and static knowledge from each other. In the frame-based dialogue management a special frame or a form is defined to encode the information that is needed to complete the underlying task, while dialogue actions can be selected on the basis of which information is known at each dialogue state. The frame-based approach is typically used in information providing systems where dialogue actions can be executed in various orders and the interaction be driven by the information needed. However, it also falls short, if the underlying task becomes complicated and requires, e.g., negotiations or dealing with multiple alternatives simultaneously.

The agent-based approach builds on the software agent technology and supports modular architecture. Different dialogue and domain tasks are decomposed into smaller tasks and implemented as independent modules, while the main interaction cycle (interpretation, decision and generation) can be implemented as an

asynchronous process between the modules and the general data flow. Agent-based architectures provide a suitable framework for conversational dialogue systems, as they allow flexible control between the modules, and in principle, advanced dialogue features can be experimented by just adding and removing relevant dialogue modules, see, e.g., [73] of an overview of distributed dialogue management. The approach also calls for advanced computational techniques and enables hybrid systems to be built where modules using rule-based and machine-learning techniques are combined in the same system.

Recently also statistical dialogue modelling has become popular. The goal is to learn relevant relations and generalisations through the frequent patterns that occur in the data, and ultimately, to condition the operation of the whole dialogue system on the statistical properties of input and output data. The approach has gained interest in both industrial and academic contexts, and as discussed above, reinforcement learning is actively used in experimenting with dialogue control strategies. Of course, the requirements for annotated data and large enough training set limits to the use of the techniques, and in some cases it may, in fact, be faster to hand-craft symbolic rules rather than to try to learn correlations through data. Although a fully stochastic dialogue system is still a research issue, early experiments with artificial neural networks and hybrid systems have shown that such systems may be possible [95, 149].

3.3.2 Discourse Modelling

The starting point for discourse and dialogue modelling has been the purposeful behaviour by the dialogue participants. Basic concepts concern dialogue actions related to turn-taking and feedback giving processes which deal with the exchange of meaningful pieces of information between the partners. Two main approaches can be distinguished concerning the nature of the rules in dialogue descriptions: the rules can function as predictive top-down rules which specify the types of utterances that can occur in particular dialogue situations, or they appear as local well-formedness constraints which bind dialogue actions into larger structures in a bottom-up manner.

3.3.2.1 Top-Down Approach

The top-down approach is based on certain theoretical assumptions that deal with the units and basic mechanisms considered important in the dialogue processing. Two main approaches, the grammar-based and the intention-based modelling, have been widely used in dialogue research.

The grammar-based approach is based on the observation that dialogues have a number of sequencing regularities (questions are followed by answers, proposals by acceptances, etc.), and on the assumption that these regularities can be captured by a dialogue grammar [127]. A dialogue grammar defines dialogue units such as moves, exchanges and segments, as well as their possible combinations, and the rules can range from straightforward re-writing rules to those with elaborated

embedded constructions. Popularity of dialogue grammars is mainly due to their formal properties which allow efficient implementation via finite state automata or context free grammars. Nowadays the grammar approach appears in the form of script-based dialogue management where the script describes the overall dialogue structure and the state transitions implement dialogue rules.

However, as already discussed with respect to the scripts, the dialogue grammar approach has shortcomings which make it less suitable for elaborated dialogue modelling. If the situation becomes more complicated, the number of grammar rules (state transitions) can grow rapidly to the extent where it will be difficult to handle the rules manually. Moreover, it is not possible to reason about why moves in the exchanges are what they are or what purposes they fulfil in the structure. In order to manage miscommunication or deviations from a predefined dialogue structure, other knowledge sources and constraints must thus be evoked. Often utterances are also multifunctional and can fill in two different structural positions simultaneously, but this is difficult to describe in strict structure models.

An alternative top-down approach, intention-based modelling, is rooted in the logic-philosophical work on speech acts by Austin [13] and Searle [124], and views language as action: utterances do not only serve to express propositions, but also to perform communicative actions. Communication is described in terms of mental states, i.e. in terms of beliefs and intentions of the participants, and the ultimate goal of a communicative act is to influence the listener's mental state so that their future actions and attitudes will accord with the intentions of the speaker, cf. [49]. People do not act randomly but plan their actions to achieve certain goals, and consequently each communicative act is tied to some underlying plan that the speaker intends to follow and intends to communicate to the partner. Successful communication means that the partner recognises the speaker's purpose in using language, adopts the goal (at least temporarily for the time of the interaction) and plans her own actions so as to assist the speaker to achieve the underlying goal, or minimally so as not to prevent the speaker from achieving her goal.

Purely intention-based dialogue management requires mechanisms to take care of the inferences concerning the speakers' beliefs and intentions. Since reasoning about pre- and post-conditions may become complicated and computationally infeasible even for simple reactions, approach has been less desirable in practical applications. However, the TRAINS system [4] shows the feasibility of the approach, and ARTIMIS [119] demonstrates that a commercial application can be built using a purely intention-based rational cooperation approach.

3.3.2.2 Dialogue Act and Plan-Based Approaches

Dialogue acts (also called communicative or conversational acts, dialogue move) have been an important means to model interaction, related especially to the intention-based dialogue models. We discuss them in a separate section, however, since they have also been used in dialogue grammar approaches and bottom-up analysis of dialogue contributions, capturing a theoretically motivated view of the language as action.

Dialogue acts extend the original concept of speech act [13, 124], with dialogue information: they describe the whole dialogue state and thus also include contextual information besides the action itself. In his Dynamic Interpretation Theory (DIT), Bunt [20–22] defines dialogue acts as a combination of the communicative function and semantic content. The communicative function is an operation on the context, and thus, when a sentence is uttered, its semantic content is expressed and also a set of background assumptions is changed. Bunt also provides a multidimensional taxonomy of dialogue acts, with the main distinction being between task-oriented and dialogue control acts. The former are related to the content (asking, informing) and the latter to managing the dialogue flow (greeting, thanking, confirming).

Dialogue acts can also be seen as a special class of communicative acts. Cohen and Levesque [33] consider specific utterance events in the context of the speaker's and hearer's mental states, and derive the different effects of the acts from general principles of rational agenthood and cooperative interaction. As they point out, this view actually renders illocutionary act recognition redundant since intention recognition is reduced to rationality considerations.

Also Allwood [5, 6] advocates this view. He discusses difficulties in determining illocutionary forces and suggests that communicative actions be identified taking into account features such as the intended and the actual achieved effects (which are not necessarily the same), overt behaviour and the context in which the communicative action is performed. This results in the definition of speech-acts having expressive, evocative and evoked functions, and consequently, dialogue acts express and evoke the agents' beliefs and intentions, and also give rise to expectations concerning the course of next actions in the communication. The agent's dialogue decisions are also constrained by certain normative obligations based on the agent's social interaction and coordinated action. Examples of the application of the principles in dialogue situations are discussed in [9].

In the development of dialogue systems, an important milestone was the observation that speech acts could be operationalised with the help of planning operators, commonly used in the AI modelling of actions [3, 37]. Planning operators define a set of preconditions that need to be fulfilled in order for the operator to apply, and a set of post-conditions that describe what the world will be like after the application of the operator. Given a goal, planning techniques can then be used to chain suitable actions into a sequence which describes a feasible plan to achieve the goal. In task-dialogues the speaker's communicative intentions are mainly linked to the underlying task, and dialogue plans tend to follow the task structure. However, although dialogue acts emerge from the actions required by the task, the speakers' dialogue strategies form a separate level of planning. Accordingly, Litman and Allen [86] specified particular metaplans, with domain and discourse plans as parameters, as a way of liberating discourse planning from a strict task structure, while Carberry [23] distinguished discourse, domain and problem-solving plans.

Much research has also been conducted on the segmentation, recognition and prediction of dialogue acts. It is based on probability estimations of the transcribed words and word sequences, conditioned on the previous dialogue act sequences and often on the prosodic information of the speech signal (pitch, duration, energy,

etc.). A wide variation of statistical and machine-learning techniques can be applied to data, the work has included, e.g. n-grams [100, 114], decision trees [70], transformation-based learning [120], neural networks [115], LVQ [68], Bayesian models [72] and rule-induction and memory-based algorithm [80].

To provide a general basis for dialogue coding and dialogue analysis, research also focused on dialogue act taxonomies and annotation manuals. For instance, DAMSL (Dialogue Act Markup in Several Layers, [39]) and the DRI annotation (Discourse Research Initiative, [25]) worked towards a general definition of dialogue acts for annotation purposes. DRI developed the taxonomy especially in accordance with the ideas from Clark [31] and Allwood [5], and emphasises the two-way functionality of the utterances: utterances are related to the previous utterances as a reaction to what has been presented earlier, and they produce expectations concerning what can be uttered next. Dialogue acts are thus said to have backward- and forward-looking functions, as they link the utterance to what has been said before, and simultaneously also commit the speaker to future actions. Consequently, interpretation of communicative acts is constructed by the participants together in the course of the dialogue, rather than being a property of a single utterance. The main discussion points in dialogue act classification and taxonomies are summarised in [135], while dialogue act standardisation activities are started in [22].

The concept of dialogue act has been a widely used tool in dialogue management although, as already mentioned, the notion of speech act has also been critically reviewed and reduced to rationality considerations. It has also been pointed out that the basic assumptions about segmenting and sequencing are problematic since this is not the way that humans participate in conversations [85]. For instance, dialogue acts can span over several turns and speakers as the participants jointly produce the dialogue act by overlapping contributions and by finishing off the partner's contributions. Utterances can also be multifunctional and it is impossible to determine one single category for the act. In some cases, it is not possible to interpret the act without taking the whole context into account: interpretation of what is being said is jointly constructed in the course of the interaction by the speaker's utterance and the partner's response to it.

3.3.2.3 Bottom-Up Approach

The descriptive soundness of the top-down approaches has been challenged because of the obvious diversity of human conversations: the true nature of the conversation, as it seems, is not a structural product, but an interactive and constructive process. Another controversy has concerned methodological issues such as whether or not formalisation is possible in dialogue studies at all. If the dialogue is understood as a dynamic process which is jointly constructed by the speakers, dialogue structure may be a side effect of the speaker's activity, and it can be observed only after the dialogue is produced. Thus regularities concerning well-formed dialogues may not be possible to be formalised as rules that the speakers follow in order to produce such dialogues, but they appear as features that characterise the result when it is studied afterwards from outside. However, if the aim is to build interactive

systems, discussion on the transient nature of dialogues also requires some formal accounts of the processing since this is necessary to enable automatic interaction in the first place.

The main representative of the bottom-up approach is Conversation Analysis (CA). It started as a branch of social anthropology called ethnomethodology, and the seminal work by Sacks et al. [118] on telephone conversations opened a new way to investigate human interactions: analysis of language in its context of use. CA emphasises the study of everyday human–human dialogues which may have no obvious purpose at all, and important aspects of language are extracted through observations of how the language is used in various human interactions and activities. CA aims at producing insights of the phenomena that occur in conversations rather than providing explicit formal rules that describe dialogue behaviour, and conversational “regularities” are thus best described as preferences and expectations of what will follow rather than prescriptive causative rules of dialogue behaviour. CA has also emphasised the role of language in establishing social reality. Analysis of interactions and institutional language reveals how attitudes, preferences and understanding of the world are maintained in language, how power-relations are encoded in language expressions, and how language is used to constitute the reality that the speakers live in. An overview of the CA approach can be found, e.g., in [47].

Although it may be difficult to apply typical CA research directly in the design of dialogue systems, it has introduced terminology which is generally used when talking about dialogue phenomena (e.g. terms like adjacency pairs, turn-taking, back-channelling, repairs, opening and closing sequences). With the focus on spontaneous spoken interaction, CA has also brought forward insights which are important in spoken dialogue design: utterances are not continuous stretches of speech but contain hesitations and pauses, they also overlap and contain non-verbal feedback, back-channelling. For instance, research on repairs has inspired computational models [57, 78, 93], and the notion of shared construction of utterances and conversational meaning has influenced models of grounding and dialogue cooperation [31, 32].

Recently data-driven approaches to dialogue management have become popular, and the bottom-up approach can also be related to this. As the methodology shifts from rule-based modelling to statistical techniques, large enough training and test data are necessary in order to experiment with the techniques. This provides an empirical basis for dialogue studies which can benefit from the mathematical algorithms that capture meaningful relations among the data. Many earlier projects have focused on collecting, analysing and annotating dialogue data, and among the widely used corpora are, e.g., MapTask [10], Verbmobil [64] and Switchboard corpora [131]. Recent efforts have centred on multimodal corpus collection such as the AMI project which concerns meeting interactions with augmented multimodal setting [24] and the CALO project (www.calo.org) which aims at conversational tools to assist meeting minutes and meeting decision-making procedures. Much effort has also been put on defining annotations and coding manuals so as to provide more standardised views of the dialogue phenomena. While the earlier research focused especially on dialogue act classifications, recently multimodal aspects have become

important, and hand gestures, facial expressions and body posture are being annotated (see, e.g., MUMIN annotation scheme, [8] and the references therein). Work has also been conducted on enabling efficient manual annotation, and various tools are available such as the MATE workbench [63] for dialogue annotation or Anvil [75] for multimodal annotation.

3.3.3 *Conversational Principles*

Many kinds of conversational principles play a role in communication. These include, e.g., cooperation, politeness and rational dialogue planning. Conversational principles are often difficult to operationalise in dialogue systems as they seem to be emergent properties of the dialogue partners' adherence to some local coherence relations rather than explicitly produced dialogue behaviour. Although people give judgements about each other's behaviour in terms of cooperation and politeness, it is less obvious how these descriptive evaluations can be translated into an algorithmic format.

Usually, conversational principles are guidelines for the dialogue designer who takes them into account when designing conditions and properties for appropriate response strategies. Grice's Conversational Maxims [49] are commonly used as such conversational principles in dialogue system design. E.g. Dybkjaer et al. [41] provide best-practice guidelines for developing practical dialogue systems and include Gricean maxims into the software design. They discuss how system designer can incrementally refine the system's robustness according to the principles and the user's behaviour.

Politeness, however, has usually been part of the dialogue system's conventional "thanks" and "good-bye" acts, and it is often noted that experienced users, who want to get the task done quickly (e.g. ask for certain information), prefer a system that trades politeness in favour of fast performance. The opposite is usually true of novice users, who prefer "chattier" and more human-human like conversations. In general, politeness is an important part of human social behaviour, and includes complex issues ranging from the linguistic expressions to cultural conventions [19]. Its role and function in interactions has been studied more recently, especially in regard to Embodied Conversational Agents and their interaction with human users (see Chapter 8 in this volume).

Intelligent agents have several different desires, some of which they have chosen as their goals. Moreover, agents are assumed to have rather selfish view points: each agent wants to achieve their own goals first and in the most appropriate way. However, the basic setting of social interaction requires that the agents' selfish view points fit together: if the agents are to interact a lot, it is better to cooperate since this pays back best in the long run [14]. The speakers are thus bound by various social commitments and obligations that regulate their actions [5, 7]. For instance, [136] consider different types of obligations which are then implemented as separate action rules, while [65] provides another view of implementing cooperation in a dialogue system: based on Allwood's Ideal Cooperation and communicative

obligations, cooperation is implemented as a series of constraints that restrict the message content and intentions to be conveyed to the partner. However, in both cases, to achieve their goals, the agents must construct a mutual context in which the goals can be recognised and collaboration can take place. In the context of pursuing goals and building the shared context, the agents have intentions that direct them to take initiative: the agent introduce new goals and have rights to control the dialogue. In dialogue system research, intentions are related to asking questions that initiate the dialogue and drive the dialogue forward, and the right to ask questions can be on the system (system-initiative), user (user-initiative) or on both (mixed-initiative), see, e.g., [29].

We can distinguish cooperation from collaboration. The former deals with the participants' shared awareness and understanding of what is communicated, while the latter is a type of coordinated activity. Cooperation thus supports collaboration and the construction of the mutual context in which information can be exchanged and specific task goals achieved. For instance, Allwood's Ideal Cooperation [5] presupposes this type of awareness of the social norms: the partners have a joint goal and they cognitively and ethically consider each other to reach the appropriate reaction in the context, and they also trust each other to behave according to similar principles.

On the other hand, Galliers [44] notes that situations where the agents have conflicting goals are common in the real world, and actually play "a crucial and positive role in the maintenance and evolution of cooperation in social systems". Cooperation is thus not only benevolent conforming to social norms and obligations, but active seeking for the achievement of goals and the resolution of conflict if the agents have conflicting goals. She extends the Cohen and Levesque teamwork formalism with the agents reasoning about obligations.

Collaboration is a type of coordinated activity and requires both planning and action in collaborative setting: beliefs and intentions of an individual agent are extended to cover beliefs and intentions that the agent possesses as a member of a group. Some goals require other agents' assistance to be achieved and some may require collaboration as part of the task itself (as, e.g. lifting a heavy item or playing tennis). Levesque et al. [82] and Cohen and Levesque [35] define the joint goal in teamwork analogously to the individual goal, except that the agents are only weakly committed to achieve the mutual goal: if they privately discover that the goal is achieved or it is impossible to achieve or irrelevant, they have to make this mutually known. The SharedPlan approach [52, 54] assumes that the members are jointly committed to bringing about a joint goal within a (partial) shared plan, which plan controls and constrains the agent's future behaviour and individual intentions. The SharedPlan approach is also used as the basis for Collagen, a collaborative interface agent framework [81] that first determines the user's plan and then in a collaborative fashion works on its way to help the user to achieve the plan.

Rational cooperation has been studied in AI-based BDI-agent interaction [34], and also in terms of utilitarian cooperation which models the balance between rational but selfish partners who seek for the best strategy to maximise their benefits with least risk taking in the long run [45, 46].

3.3.4 *HCI and Dialogue Models*

Research activities within the Human–Computer Interaction (HCI) community also relate to dialogue modelling. Although the main goal of the HCI research has been to design user-friendly interfaces rather than to model dialogues as such, it has had impact on the models and methods that enable interaction between the user and the system. Especially, issues related to the “human factor”, i.e. robust and natural interaction as well as user evaluation and usability, are also concern of the dialogue modelling community.

In HCI, much research has focused on ergonomics: trying to match task and system requirements with the user’s thinking and cognitive limitations. Starting from the design of the computer hardware (keyboard, screen, mouse, etc.), HCI deals with the interface design (graphical user interfaces with layout, buttons and menus, as well as speech user interfaces which include spoken commands) so as to provide systems with transparent functionality and easy to operate interfaces which impose as small a cognitive load as possible on the user. These aspects directly concern the enablements of interaction and the usefulness of the different modalities (graphics, text, speech, touch) in order to make interfaces more natural and intuitive. From the spoken dialogue point of view, we can argue that the interface should also include natural language capability, so as to provide an interface that is truly affordable for human users. Interactive computers also bring new social dimensions to the interface design, and the design metaphor has changed from the computer-as-a-tool to computer-as-an-agent [66].

The emphasis in interface design has been on clear, unambiguous prompts which provide the user with explicit information about the task and the system’s capabilities. The golden design principles concern such aspects as linguistic clarity, simplicity, predictability, accuracy, suitable tempo, consistency, precision, forgiveness and responsiveness [146]. An important aspect has also been to help the users to feel they master the usage of the tool and are in control of the application: the users should be able to decide when and how to accustom the system to their particular needs. Recently, it has also been emphasised that the usage situation should encourage positive experience, and that the users should feel satisfied with the service that the interactive system provides.

To assess how the systems address task requirements and the users’ various needs, user evaluation is carried out in the HCI studies. The factors related to how the users perceive the application are gathered under the notion of usability: the system is easy to use, easy to learn and easy to talk to [144]. Much of the evaluation research concerns user interfaces: e.g. heuristic evaluation [106] involves inspection of the interface by special evaluators with respect to recognised usability principles (heuristics) such as the visibility of system status, match between the system and the real world, user control, consistency, error prevention, recognition, flexibility and efficiency of use, aesthetic design and help. In the user-centred design [107], users are taken into account early in the design cycle in order to provide information on the desired features and to influence the system development in the early

stages. The design proceeds in an iterative way, i.e. in repeated cycles during which the system is refined and tested.

Often in system evaluation the objective and subjective criteria do not match, i.e. task success and user satisfaction may show opposite values. Paradoxically, users can tolerate problems and difficulties such as long waiting times and mere errors, as long as the system is interesting and the users motivated at using it. Also, the users seem to rank a system and its speech capabilities higher if they assume speech is an extra facility that the interface provides, compared with a situation where the users expect to interact with an ordinary speech-based system [67]. This means that the main issue in the evaluation process is to select design features and quality measures that appropriately operationalise the user's perception of the system. There is no clear framework for measuring the impact of the system on the user experience as a whole, but various aspects that need to be taken into account have been proposed and discussed. For instance, [99, and Chapter 15 in this volume] presents the quality of experience evaluation and lists different factors that should be taken into account when evaluating the quality of practical systems. The quality factors include e.g. recognition of communicative principles that support easy and natural interaction as well as trust in the system's ability to function reliably and to provide truthful information. Recently, it has also been emphasised that the evaluation should not only deal with the system's performance as perceived by the users, but also with what the users desire or expect from the system, i.e. there is a need to quantify the system's value for the users.

3.4 Conclusion

Dialogue research has developed much during the 60 years of computer technology and now allows us to build dialogue systems which interact with the user fairly robustly albeit in a fixed way. Research directions deal with the sophistication and development of modelling techniques and widening the interaction repertoire towards multimodal aspects. Research thus continues in investigating such issues as non-verbal communication, emotion and cognition in language interaction, as well as extending interaction strategies towards multi-party conversations and social communication, and developing new multimodal interaction techniques. As for applications, speech-based interaction technology is deployed, e.g. in tutoring systems, in-car navigation and question-answering service systems. Also command interfaces to human-robot interaction and in the games industry are application domains for spoken dialogue interaction.

In the ubiquitous computing context [147] it is envisaged that future information and communication systems will concern computers that are embedded in our daily environment, and which can communicate easily with other intelligent objects and also have intuitive human-computer interfaces that interpret human expressions. In other words, human agents are required to interact with a complex environment which will not only consist of people, but of intelligent dialogue agents too. Even though it is not necessarily to agree with all the fantasy views of the future society, it

is obvious that there will be an increasing number of intelligent interactive systems that can cater for many everyday needs for the users and can mediate between the users and automatic services which have become increasingly complicated.

Going back to Turing's idea of a Thinking Machine, it is thus tempting to say that such a machine is, in principle, achievable. In fact, this view is entertained in the annual Loebner Prize Competition (<http://www.loebner.net/Prizef/loebner-prize.html>) which are held as Turing tests: human users converse with a system over a keyboard and try to establish whether their partner is a human or a program. So far the results have very clearly shown human superiority, and criticism has also been raised because the general aim of the interaction is to fool the human user, which may not lead to strategies that are expected from cooperative dialogue agents and even less from reliable service agents. Although interaction should be natural and enjoyable, many dialogue situations do not deal with chatting or affecting the partner as such, but have some underlying goals to fulfil, too. The intelligence of the partner is thus measured with respect to the rationality of the plan that they construct to achieve the goals, and to the appropriateness of the means that the agent takes to realise the plan in practice. Consequently, also exhibit some planning capability besides engaging the partner in a pleasant chat.

The situation seems to resemble the one in the beginning of 1950s: the idea of a Thinking Machine inspires research and development but in practice seems to evade realisation. Back then the main problems were related to the conceptual modelling of the world knowledge and enabling computer agents to use the models automatically in their interactions with human users. These issues now to have been resolved for limited domains and simple interactions as evidenced by the current state of dialogue technology: industrial prospects of dialogue systems can be regarded as a proof of the dialogue research having reached the level where its results can be successfully applied to practical tasks. However, at the same time, the complexity of the interaction task seems to have grown, too: e.g., talking robots can move autonomously and navigate in the complex information space. Thus it is not only the mere amount of information and its efficient modelling that is needed, but learning how to acquire and use the information.

This is related to the better understanding of the ways in which the information is clustered and classified into meaningful concepts and used to master the surrounding world. It is thus important to study how intelligent agents learn the necessary skills needed to cope with their environment (including language communication to coordinate their actions with the other agents), and how they accumulate knowledge through their experience, adopting best practices either by being taught or by trying out solutions themselves. The complex and dynamic nature of learning has not really been discussed in dialogue research, although it has been important in the areas such as cognitive science, robotics and neurolinguistics.

The agent-metaphor emphasis that the computer system functions as an agent among human agents, and so it is important that humans will be able to communicate with the systems. As the computer has different capabilities from the human users, e.g. tireless repetition of the same actions and effortless calculation of long mathematical formulas, it is likely that these differences form the core

topic areas that the system should learn to communicate about with humans in a natural way. Essential features of intelligence are large amounts of information and the number of action choices which cannot simply be enumerated. Intelligent conversational agents can replace direct manipulation interfaces and prove their usefulness as well as “intelligence” exactly through their skills to manage interaction on issues which are perceived as complex and uncontrollable from the user’s point of view. One of the features of intelligent agents (intelligent dialogue systems) is simply their ability to communicate by using natural language and natural interaction strategies.

References

1. Ai, H., Raux, A., Bohus, D., Eskenazi, M., Litman, D. (2007). Comparing spoken dialog corpora collected with recruited subjects versus real users. In: Proc. 8th SIGDial Workshop on Discourse and Dialogue, Antwerp, Belgium.
2. Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., Stent, A. (2000). An architecture for a generic dialog shell. *Nat. Lang. Eng.*, 6 (3), 1–16.
3. Allen, J., Perrault, C.R. (1980). Analyzing intention in utterances. *Artif. Intell.*, 15, 143–178.
4. Allen, J. F., Schubert, L. K., Ferguson, G., Heeman, P., Hwang, C. H., Kato, T., Light, M., Martin, N. G., Miller, B. W. Poesio, M., Traum, D. R. (1995). The TRAINS Project: A case study in building a conversational planning agent. *J. Exp. Theor. AI*, 7, 7–48. Also available as TRAINS Technical Note 94–3 and Technical Report 532, Computer Science Department, University of Rochester, September 1994.
5. Allwood, J. (1976). *Linguistic Communication as Action and Cooperation*. Department of Linguistics, University of Göteborg. Gothenburg Monographs in Linguistics, 2.
6. Allwood, J. (1977). A critical look at speech act theory. In: Dahl, Ö. (ed.) *Logic, Pragmatics, and Grammar*, Studentlitteratur, Lund.
7. Allwood, J. (1994). Obligations and options in dialogue. *Think Q.*, 3, 9–18.
8. Allwood, J. Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In: Martin, J. C., Paggio, P., Kuenlein, P., Stiefelhagen, R., Piansi F. (eds), *Multimodal Corpora For Modelling Human Multimodal Behaviour*. *Int. J. Lang. Res. Eval. (Special Issue)*, 41 (3–4), 273–287.
9. Allwood, J., Traum, D., Jokinen, K. (2000). Cooperation, dialogue, and ethics. *Int. J. Hum. Comput. Studies*, 53, 871–914.
10. Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., Weinert, R. (1991). The HCRC map task corpus. *Lang. Speech*, 34 (4), 351–366.
11. Appelt, D. E. (1985). *Planning English Sentences*. Cambridge University Press, Cambridge.
12. Aust, H., Oerder, M., Seide, F., Steinbiss, V. (1995). The Philips automatic train timetable information system. *Speech Commun.*, 17, 249–262.
13. Austin, J. L. (1962). *How to do Things with Words*. Clarendon Press, Oxford.
14. Axelrod, R. (1984). *Evolution of Cooperation*. Basic Books, New York.
15. Ballim, A., Wilks, Y. (1991). *Artificial Believers*. Lawrence Erlbaum Associates, Hillsdale, NJ.
16. Black, W., Allwood, J., Bunt, H., Dols, F., Donzella, C., Ferrari, G., Gallagher, J., Haidan, R., Imlah, B., Jokinen, K., Lancel, J.-M., Nivre, J., Sabah, G., Wachtel, T. (1991). A pragmatics based language understanding system. In: Proc. ESPRIT Conf. Brussels, Belgium.
17. Bolt, R.A. (1980). Put-that-there: Voice and gesture at the graphic interface. *Comput. Graphics*, 14 (3), 262–270.

18. Bos, J., Klein, E., Oka T. (2003). Meaningful conversation with a mobile robot. In: Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), Budapest, 71–74.
19. Brown, P., Levinson, S. C. (1999) [1987]. Politeness: Some universals in language usage. In: Jaworski, A., Coupland, N. (eds) *The Discourse Reader*. Routledge, London, 321–335.
20. Bunt, H. C. (1990). DIT – Dynamic interpretation in text and dialogue. In: Kálmán, L., Pólos, L. (eds) *Papers from the Second Symposium on Language and Logic*. Akadémiai Kiadó, Budapest.
21. Bunt, H. C. (2000). Dynamic interpretation and dialogue theory. In: Taylor, M. M. Néel, F., Bouwhuis, D. G. (eds) *The Structure of Multimodal Dialogue II*, John Benjamins, Amsterdam, 139–166.
22. Bunt, H. C. (2005). A framework for dialogue act specification. In: Fourth Workshop on Multimodal Semantic Representation (ACL-SIGSEM and ISO TC37/SC4), Tilburg.
23. Carberry, S. (1990). *Plan Recognition in Natural Language Dialogue*. MIT Press, Cambridge, MA.
24. Carletta, J. (2006). Announcing the AMI Meeting Corpus. *ELRA Newslett.*, 11 (1), 3–5.
25. Carletta, J., Dahlbäck, N., Reithinger, N., Walker, M. (eds) (1997). *Standards for Dialogue Coding in Natural Language Processing*. Dagstuhl-Seminar Report 167.
26. Carlson R. (1996). The dialog component in the Waxholm system. In: LuperFoy, S., Nijholt, A., Veldhuijzen van Zanten, G. (eds) *Proc. Twente Workshop on Language Technology. Dialogue Management in Natural Language Systems (TWLT 11)*, Enschede, The Netherlands, 209–218.
27. Chin, D. (1989). KNOBE: Modeling what the user knows in UC. In: Kobsa, A., Wahlster, W. (eds) *User Modeling in Dialogue Systems*. Springer-Verlag Berlin, Heidelberg, 74–107.
28. Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague/Paris.
29. Chu-Carroll, J., Brown, M. K. (1998). An evidential model for tracking initiative in collaborative dialogue interactions. *User Model. User-Adapted Interact.*, 8 (3–4), 215–253.
30. Chu-Carroll, J., Carpenter, B. (1999). Vector-based natural language call routing. *Comput. Linguist.*, 25 (3), 256–262.
31. Clark, H. H., Schaefer, E. F. (1989). Contributing to discourse. *Cogn. Sci.*, 13, 259–294.
32. Clark, H. H., Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1–39.
33. Cohen, P. R., Levesque, H. J. (1990a). Persistence, intention, and commitment. In: Cohen, P. R., Morgan, J., Pollack, M. E. (eds) *Intentions in Communication*. The MIT Press, Cambridge, MA, 33–69.
34. Cohen, P. R., Levesque, H. J. (1990b). Rational interaction as the basis for communication. In: Cohen, P. R., Morgan, J., Pollack, M. E. (eds) *Intentions in Communication*. The MIT Press, Cambridge, MA, 221–255.
35. Cohen, P. R., Levesque, H. J. (1991). Teamwork. *Nous*, 25 (4), 487–512.
36. Cohen, P. R., Morgan, J., Pollack, M. (eds) (1990). *Intentions in Communication*. MIT Press, Cambridge.
37. Cohen, P. R., Perrault, C. R. (1979). Elements of plan-based theory of speech acts. *Cogn. Sci.*, 3, 177–212.
38. Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V. (eds) (1996). *Survey of the State of the Art in Human Language Technology*. Also available at <http://www.cse.ogi.edu/CSLU/HLTSurvey/>
39. Core, M. G., Allen, J. F. (1997). Coding dialogs with the DAMSL annotation scheme. In: *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA.
40. Danieli M., Gerbino E. (1995). Metrics for evaluating dialogue strategies in a spoken language system. In: *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, Stanford University, 34–39.

41. Dybkjaer, L., Bernsen, N. O., Dybkjaer, H. (1996). Evaluation of spoken dialogue systems. In: Proc. 11th Twente Workshop on Language Technology, Twente.
42. Erman, L. D., Hayes-Roth, F., Lesser, V. R., Reddy, D. R. (1980). The HEARSAY-II speech understanding system: Integrating knowledge to resolve uncertainty. *Comput. Surv.*, 12 (2), 213–253.
43. Esposito, A., Campbell, N., Vogel, C., Hussain, A., and Nijholt, A. (Eds.). *Development of Multimodal Interfaces: Active Listening and Synchrony*. Springer Publishers.
44. Galliers, J. R. (1989). A theoretical framework for computer models of cooperative dialogue, acknowledging multi-agent conflict. Technical Report 17.2, Computer Laboratory, University of Cambridge.
45. Gmytrasiewicz, P. J., Durfee, E. H. (1993). Elements of utilitarian theory of knowledge and action. In: Proc. 12th Int. Joint Conf. on Artificial Intelligence, Chambry, France, 396–402.
46. Gmytrasiewicz, P. J., Durfee, E. H., Rosenschein, J. S. (1995). Towards rational communicative behavior. In: AAAI Fall Symp. on Embodied Language, AAAI Press, Cambridge, MA.
47. Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. Academic Press, New York.
48. Gorin, A. L., Riccardi, G., Wright, J. H. (1997). How may i help you? *Speech Commun.*, 23 (1/2), 113–127.
49. Grice, H. P. (1975). Logic and conversation. In: Cole, P., Morgan, J. L. (eds) *Syntax and Semantics. Vol 3: Speech Acts*. Academic Press, New York, 41–58.
50. Grosz, B. J. (1977). *The Representation and Use of Focus in Dialogue Understanding*. SRI Stanford Research Institute, Stanford, CA.
51. Grosz, B. J., Hirschberg, J. (1992). Some international characteristics of discourse. *Proceedings of the Second International Conference on Spoken Language Processing (ICSLP'92)*, Banff, Alberta, Canada, 1992, 429–432.
52. Grosz, B. J., Kraus, S. (1995). Collaborative plans for complex group action. Technical Report TR-20-95, Harvard University, Center for Research in Computing Technology.
53. Grosz, B. J., Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Comput. Linguist.*, 12 (3), 175–203.
54. Grosz, B. J., Sidner, C. L. (1990). Plans for discourse. In: Cohen, P. R., Morgan, J., Pollack, M. E. (eds) *Intentions in Communication*. The MIT Press. Cambridge, MA, 417–444.
55. Guinn, C. I. (1996). Mechanisms for mixed-initiative human-computer collaborative discourse. In: Proc. 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, California, USA, 278–285.
56. Hasida, K., Den, Y., Nagao, K., Kashioka, H., Sakai, K., Shimazu, A. (1995). Dialeague: A proposal of a context for evaluating natural language dialogue systems. In: Proc. 1st Annual Meeting of the Japanese Natural Language Processing Society, Tokyo, Japan, 309–312 (in Japanese).
57. Heeman, P. A., Allen, J. F. (1997). International boundaries, speech repairs, and discourse markers: Modelling spoken dialog. In: Proc. 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain.
58. Hirasawa, J., Nakano, M., Kawabata, T., Aikawa, K. (1999). Effects of system barge-in responses on user impressions. In: Sixth Eur. Conf. on Speech Communication and Technology, Budapest, Hungary, 3, 1391–1394.
59. Hirschberg, J., Litman, D. (1993). Empirical studies on the disambiguation of cue phrases *Comput. Linguist.*, 19 (3), 501–530.
60. Hirschberg, J., Nakatani, C. (1998). Acoustic indicators of topic segmentation. In: Proc. Int. Conf. on Spoken Language Processing, Sydney, Australia, 976–979.
61. Hobbs, J. (1979). Coherence and coreference. *Cogn. Sci.*, 3 (1), 67–90.
62. Hovy, E. H. (1988). *Generating Natural Language under Pragmatic Constraints*. Lawrence Erlbaum Associates, Hillsdale, NJ.

63. Isard, A., McKelvie, D., Cappelli, B., Dybkjær, L., Evert, S., Fitschen, A., Heid, U., Kipp, M., Klein, M., Mengel, A., Møller, M. B., Reithinger, N. (1998). Specification of workbench architecture. MATE Deliverable D3.1.
64. Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M., Quantz, J. (1995). Dialogue acts in VERBMOBIL. Technical Report 65, BMBF Verbmobil Report.
65. Jokinen, K. (1996). Goal formulation based on communicative principles. In: Proc. 16th Int. Conf. on Computational Linguistics (COLING - 96) Copenhagen, Denmark, 598–603.
66. Jokinen, K. (2009). Constructive Dialogue Modelling – Speech Interaction and Rational Agents. John Wiley, Chichester.
67. Jokinen, K., Hurtig, T. (2006). User expectations and real experience on a multimodal interactive system. In: Proc. 9th Int. Conf. on Spoken Language Processing (Interspeech 2006 – ICSLP) Pittsburgh, US.
68. Jokinen, K., Hurtig, T., Hynnä, K., Kanto, K., Kerminen, A., Kaipainen, M. (2001). Self-organizing dialogue management. In: Isahara, H., Ma, Q. (eds) NLPRS2001 Proc. 2nd Workshop on Natural Language Processing and Neural Networks, Tokyo, Japan, 77–84.
69. Joshi, A., Webber, B. L., Weischedel, R. M. (1984). Preventing false inferences. In: Proc. 10th Int. Conf. on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics, 1984, Stanford, California, USA, 34–138.
70. Jurafsky, D., Shriberg, E., Fox, B., Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In: ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Montreal, Quebec, Canada.
71. Kearns, M., Isbell, C., Singh, S., Litman, D., Howe, J. (2002). CobotDS: A spoken dialogue system for chat. In: Proceedings of the Eighteenth National Conference on Artificial Intelligence, Edmonton, Alberta.
72. Keizer, S., Akker, R. op den, Nijholt, A. (2002). Dialogue act recognition with Bayesian Network for Dutch dialogues. In: Jokien, K., McRoy, S. (eds.) Proc. 3rd SIGDial Workshop on Discourse and Dialogue, Philadelphia, US.
73. Kerminen, A., Jokinen, K. (2003). Distributed dialogue management. In: Jokinen, K., Gambäck, B., Black, W. J., Catizone, R., Wilks, Y. (eds.) Proc. EACL Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management. Budapest, Hungary.
74. Kendon, A. (2004). Gesture: Visible action as utterance. Cambridge. In: Proc. 13th Eur. Conf. on Artificial Intelligence (ECAI).
75. Kipp, M. (2001). Anvil – A generic annotation tool for multimodal dialogue. In: Proc. 7th Eur. Conf. on Speech Communication and Technology, (Eurospeech), Aalborg, Denmark, 1367–1370.
76. Koeller, A., Kruijff, G.-J. (2004). Talking robots with LEGO mindstorms. In: Proc. 20th COLING, Geneva.
77. Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y. (1998). An analysis of turn taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. Lang. Speech, 41 (3–4), 295–321.
78. Krahmer, E., Swerts, M., Theune, M., Weegels, M. (1999). Problem spotting in human-machine interaction. In: Proc. Eurospeech '99, Budapest, Hungary, 3, 1423–1426.
79. Lemon, O., Bracy, A., Gruenstein, A., Peters, S. (2001). The WITAS multi-modal dialogue system I. In: Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), Aalborg, Denmark.
80. Lendvai, P., Bosch, A. van den, Krahmer, E. (2003). Machine learning for shallow interpretation of user utterances in spoken dialogue systems. In: Jokinen, K., Gambäck B., Black, W. J., Catizone, R., Wilks, Y. (eds) Proc. ACL Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management, Budapest, Hungary, 69–78.
81. Lesh, N., Rich, C., Sidner, C. L. (1998). Using plan recognition in human-computer collaboration. MERL Technical Report.

82. Levesque, H. J., Cohen, P. R., Nunes, J. H. T. (1990). On acting together. In: Proc. AAAI-90, 94–99. Boston, MA.
83. Levin, E., Pieraccini, R. (1997). A stochastic model of computer-human interaction for learning dialogue strategies. In: Proc. Eurospeech, 1883–1886, Rhodes, Greece.
84. Levin, E., Pieraccini, R., Eckert, W. (2000). A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Trans. Speech Audio Process.*, 8, 1.
85. Levinson, S. (1983). *Pragmatics*. Cambridge University Press, Cambridge.
86. Litman, D. J., Allen, J. (1987). A plan recognition model for subdialogues in conversation. *Cogn. Sci.*, 11(2), 163–200.
87. Litman, D., Kearns, M., Singh, S., Walker, M. (2000). Automatic optimization of dialogue management. In: Proc. 18th Int. Conf. on Computational Linguistics (COLING 2000) Saarbrücken, Germany, 502–508.
88. Lopez Cozar, R., Araki, M. (2005). *Spoken, multilingual and multimodal dialogue systems*. Wiley, New York, NY.
89. Majaranta, P., Räihä, K. (2002). Twenty years of eye typing: Systems and design issues. In: Proc. 2002 Symp. on Eye Tracking Research & Applications (ETRA '02), ACM, New York, 15–22.
90. Martin, D., Cheyer, A., Moran, D. (1998). Building distributed software systems with the Open Agent Architecture. In: Proc. 3rd Int. Conf. on the Practical Application of Intelligent Agents and Multi-Agent Technology, Blackpool, UK. The Practical Application Company, Ltd.
91. McCoy, K. F. (1988). Reasoning on a highlighted user model to respond to misconceptions. *Comput. Linguist.*, 14 (3), 52–63.
92. McGlashan, S., Fraser, N. M., Gilbert, N., Bilange, E., Heisterkamp, P., Youd, N. J. (1992). Dialogue management for telephone information services. In: Proc. Int. Conf. on Applied Language Processing, Trento, Italy.
93. McRoy, S. W., Hirst, G. (1995). The repair of speech act misunderstandings by abductive inference. *Comput. Linguist.*, 21 (4), 435–478.
94. McTear, M. (2004). *Spoken Dialogue Technology: Toward the Conversational User Interface*. Springer Verlag, London.
95. Miikkulainen, R. (1993). *Sub-symbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. MIT Press, Cambridge.
96. Minsky, M. (1974). *A Framework for Representing Knowledge*. AI Memo 306. M.I.T. Artificial Intelligence Laboratory, Cambridge, MA.
97. Moore, J. D., Swartout, W. R. (1989). A reactive approach to explanation. In: Proc. 11th Int. Joint Conf. on Artificial Intelligence (IJCAI), Detroit, MI, 20–25.
98. Motooka, T., Kitsuregawa, M., Moto-Oka, T., Apps, F. D. R. (1985). *The Fifth Generation Computer: The Japanese Challenge*. Wiley, New York, NY.
99. Möller, S. (2002). A new taxonomy for the quality of telephone services based on spoken dialogue systems. In: Jokinen, K., McRoy, S. (eds) Proc. 3rd SIGdial Workshop on Discourse and Dialogue, Philadelphia, PA, 142–153.
100. Nagata, M., Morimoto, T. (1994). First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Commun.*, 15 (3–4), 193–203.
101. Nakano, M., Miyazaki, N., Hirasawa, J., Dohsaka, K., Kawabata, T. (1999). Understanding unsegmented user utterances in real-time spoken dialogue systems. In: Proc. 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Maryland, USA, 200–207.
102. Nakano, M., Miyazaki, N., Yasuda, N., Sugiyama, A., Hirasawa, J., Dohsaka, K., Aikawa, K. (2000). WIT: Toolkit for building robust and real-time spoken dialogue systems. In: Dybkjær, L., Hasida, K., Traum, D. (eds) Proc. 1st SIGDial workshop on Discourse and Dialogue – Volume 10, Hong Kong, 150–159.

103. Nakatani, C., Hirschberg, J. (1993). A speech-first model for repair detection and correction. In: Proc. 31st Annual Meeting on Association for Computational Linguistics, Columbus, OH, 46–53.
104. Nakatani, C., Hirschberg, J., Grosz, B. (1995). Discourse structure in spoken language: Studies on speech corpora. In: Working Notes of the AAAI-95 Spring Symposium on Empirical Methods in Discourse Interpretation, Palo Alto, CA.
105. Newell, A., Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Commun. ACM*, 19, 113–126.
106. Nielsen, J. (1994). Heuristic evaluation. In: Nielsen, J., Mack, R. L. (eds) *Usability Inspection Methods*, Chapter 2, John Wiley, New York.
107. Norman, D. A., Draper, S. W. (eds) (1986). *User Centered System Design: New Perspectives on Human-Computer Interaction*. Lawrence Erlbaum Associates, Hillsdale, NJ.
108. Paek, T., Pieraccini, R. (2008). Automating spoken dialogue management design using machine learning: an industry perspective. In: McTear, M. F., Jokinen, K., Larson, J. (eds) *Evaluating New Methods and Models for Advanced Speech-Based Interactive Systems. Special Issue of Speech Commun.*, 50 (8–9).
109. Paris, C. L. (1988). Tailoring object descriptions to a user's level of expertise. *Comput. Linguist.*, 14 (3), 64–78.
110. Power, R. (1979). Organization of purposeful dialogue. *Linguistics*, 17, 107–152.
111. Price, P., Hirschman, L., Shriberg, E., Wade, E. (1992). Subject-based evaluation measures for interactive spoken language systems. In: Proc. Workshop on Speech and Natural Language, Harriman, New York, 34–39.
112. Reichman, R. (1985). *Getting Computers to Talk Like You and Me. Discourse Context, Focus, and Semantics (An ATN Model)*. The MIT Press, Cambridge, MA.
113. Reithinger, N., Maier, E. (1995). Utilizing statistical dialogue act processing in *Verbmobil*. In: Proc. 33rd Annual Meeting of ACL, MIT, Cambridge, US, 116–121.
114. Ries, K. (1999). HMM and neural network based speech act detection. ICASSP. Also available: citeseer.nj.nec.com/ries99hmm.html
115. Roy, N., Pineau, J., Thrun, S. (2000). Spoken dialog management for robots. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong.
116. Rudnicky, A., Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu, W., Oh, A. (1999). Creating natural dialogs in the Carnegie Mellon Communicator System. In: Proc. 6th Eur. Conf. on Speech Communication and Technology (Eurospeech-99), Budapest, 1531–1534.
117. Sacks, H., Schegloff, E., Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50 (4), 696–735.
118. Sadek, D., Bretier, P., Panaget, F. (1997). ARTIMIS: Natural dialogue meets rational agency. In: Proc. IJCAI-97, Nagoya, Japan, 1030–1035.
119. Samuel, K., Carberry, S., Vijay-Shanker, K. (1998). Dialogue act tagging with transformation-based learning. In: Proc. 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (ACL-COLING), Montreal, Quebec, Canada, 1150–1156.
120. Schank, R. C., Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum Associates, Hillsdale, NJ.
121. Schatzmann, J., Weilhammer, K., Stuttle, M. N., Young, S. (2006). A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowledge Eng. Rev.*, 21 (2), 97–126.
122. Scheffler, K., Young, S. (2000). Probabilistic simulation of human-machine dialogues. In: Proc. IEEE ICASSP, Istanbul, Turkey, 1217–1220.
123. Searle, J. R. (1979). *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press, Cambridge.
124. Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., Zue, V. (1998). GALAXY-II: A reference architecture for conversational system development. In: Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP 98), Sydney, Australia.

125. Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Lang. Speech*, 41, 3–4, 439–487.
126. Sinclair, J. M., Coulthard, R. M. (1975). *Towards an Analysis of Discourse: The English Used by Teacher and Pupils*. Oxford University Press, Oxford.
127. Smith, R. W. (1998). An evaluation of strategies for selectively verifying utterance meanings in spoken natural language dialog. *Int. J. Hum. Comput. Studies*, 48, 627–647.
128. Smith, R. W., Hipp, D. R. (1994). *Spoken Natural Language Dialog Systems – A Practical Approach*. Oxford University Press, New York, NY.
129. Stent, A., Dowling, J., Gawron, J. M., Owen-Bratt, E., Moore, R. (1999). The CommandTalk spoken dialogue system. In: *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, College Park, Maryland, US, 20–26.
130. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., Meteer, M. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Comput. Linguist.*, 26 (3), 339–373.
131. Suhm, B., Geutner, P., Kemp, T., Lavie, A., Mayfield, L., McNair, A. E., Rogina, I., Schultz, T., Sloboda, T., Ward, W., Woszczyna, M., Waibel, A. (1995). JANUS: Towards multilingual spoken language translation. In: *Proc. ARPA Spoken Language Workshop*, Austin, TX.
132. Swerts, M., Hirschberg, J., Litman, D. (2000). Correction in spoken dialogue systems. In: *Proc. Int. Conf. on Spoken Language Processing (ICSLP-2000)*, Beijing, China, 615–618.
133. Takezawa, T., Morimoto, T., Sagisaka, Y., Campbell, N., Iida, H., Sugaya, F., Yokoo, A., Yamamoto, S. (1998). A Japanese-to-English speech translation system: ATR-MATRIX. In: *Proc. (ICSLP-98)*, Sydney, Australia, 957–960.
134. Traum, D. R. (2000). 20 questions on dialogue act taxonomies. *J. Semantics*, 17, 7–30.
135. Traum, D. R., Allen, J. F. (1994). Discourse obligations in dialogue processing. In: *Proc. 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, USA, 1–8.
136. Traum, D., Roque, A., Leuski, A., Georgiou, P., Gerten, J., Martinovski, B., Narayanan, S., Robinson, S., Vaswani Hassan, A. (2007). A virtual human for tactical questioning. In: *Proc. 8th SIGDial Workshop on Discourse and Dialogue*, Antwerp, Belgium, 71–74.
137. Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 49, 433–460.
138. Wahlster, W. (ed) (2000). *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin.
139. Wahlster, W., Marburger, H., Jameson, A., Busemann, S. (1983). Overanswering yes-no Questions: Extended responses in a NL interface to a vision system. In: *Proc. 8th Int. Joint Conf. on Artificial Intelligence (IJCAI'83)*, Karlsruhe, 643–646.
140. Walker, M. A., Fromer, J. C., Narayanan, S. (1998). Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In: *Proc. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* Montreal, Quebec, Canada.
141. Walker, M. A., Hindle, D., Fromer, J., Di Fabbrizio, G., Mestel, G. (1997a). Evaluating competing agent strategies for a voice email agent. In: *Proc. 5th Eur. Conf. on Speech Communication and Technology (Eurospeech 97)*, Rhodes, Greece.
142. Walker, M. A., Litman, D. J., Kamm, C. A., Abella, A. (1997b). Evaluating spoken dialogue agents with PARADISE: Two case studies. *Comput. Speech Lang.*, 12 (3), 317–347.
143. Wallace, M. D., Anderson, T. J. (1993). Approaches to interface design. *Interacting Comput.*, 5 (3), 259–278.
144. Ward, N., Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *J. Pragmatics*, 23, 1177–1207.
145. Weinschenk, S., Barker, D. (2000). *Designing Effective Speech Interfaces*. Wiley, London.
146. Weiser, M. (1991). The computer for the twenty-first century. *Sci. Am.*, September 1991 (Special Issue: Communications, Computers and Networks), 265(3), 94–104.
147. Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9, 36–45.

148. Wermter, S., Weber, V. (1997). SCREEN: Learning a flat syntactic and semantic spoken language analysis using artificial neural networks. *J. Artif. Intell. Res.*, 6 (1), 35–85.
149. Williams, J. D., Young, S. J. (2007). Partially observable Markov decision processes for spoken dialog systems. *Comput. Speech Lang.*, 21 (2), 231–422.
150. Winograd, T. (1972). *Understanding Natural Language*. Academic Press, New York.
151. Woods, W. A., Kaplan, R. N., Webber, B. N. (1972). *The lunar sciences natural language information system: Final Report*. BBN Report 2378, Bolt Beranek and Newman Inc., Cambridge, MA.
152. Yankelovich, N. (1996). How do users know what to say? *Interactions*, 3 (6), 32–43.
153. Young, S. L., Hauptmann, A. G., Ward, W. H., Smith, E. T., Werner, P. (1989). High-level knowledge sources in usable speech recognition systems, *Commun. ACM*, 32 (2), 183–194.
154. Zock, M., Sabah, G. (eds) (1988). *Advances in Natural Language Generation: An Interdisciplinary Perspective*. Pinter Publishers, London.

Chapter 4

The Industry of Spoken-Dialog Systems and the Third Generation of Interactive Applications

Roberto Pieraccini

4.1 Introduction

One of the first speech recognition systems was built in 1952 at Bell Laboratories [1]. The system could recognize sequences of digits spoken with pauses between them. Davis, Biddulph, and Balashek, the three inventors, reported that

[...] an accuracy varying between 97 and 99 percent is obtained when a single male speaker repeats any random series of digits. However, the system required to be adjusted for each talker [...] if no adjustment is made, accuracy may fall to as low as 50 or 60 percent in a random digit series [1].

The automatic digit recognition machine, dubbed AUDREY, was completely built with analog electronic circuits, and although automated voice dialing was an attractive solution for AT&T toward cost reduction in the long-distance call business, it was never deployed commercially.

Given these early successes, why were they not exploited? They were not economically attractive. [...] AUDREY occupied a six-foot high relay rack, was expensive, consumed substantial power and exhibited the myriad maintenance problems associated with complex vacuum-tube circuitry. More important, its reliable operation was limited to accurate recognition of digits spoken by designated talkers. It could therefore be used for voice dialing by, say, toll operators, or especially affluent telephone customers, but this accomplishment was poorly competitive with manual dialing of numbers. In most cases, digit recognition is faster and cheaper by push-button dialing, rather than by speaking the successive digits [2].

As described in Chapter 1, it took more than three decades for the algorithms and the technology of speech recognition to find a stable setting within the framework of statistical modeling. And it took nearly two more decades of incremental improvement for it to reach an acceptable level of performance. Only toward the beginning of the new millennium, almost 50 years after AUDREY was built, did we witness the emergence of a more mature market: a structured industry based on human-machine communication through speech.

R. Pieraccini (✉)
SpeechCycle, Inc., 26 Broadway, 11th Floor, New York, NY 10004, USA
e-mail: roberto@speechcycle.com

The principles of modern speech recognition technology today are not dissimilar to those introduced in the late 1970s and early 1980s. However, faster and cheaper computers, and the increased availability of transcribed speech data have enabled a steady, incremental improvement of accuracy and performance. Although automatic speech recognition is far from perfect, many applications benefit from it, particularly spoken dialog. This is where much of the speech technology industries have placed their bets during the last decade.

As speech recognition went through several generations of algorithms and technology (see Chapter 2), in a similar fashion, the young spoken-dialog industry moved from the first generation of simple interactive applications, in the mid 1990s, to the most complex and sophisticated third-generation systems of 2005 and later years. The improvement of the speech recognition technology during the past decades in terms of accuracy, vocabulary size, and real-time performance is certainly one of the main enablers of today's spoken-dialog industry. But, at the same time, and not less importantly, the advancement in voice user interface (VUI) design, dialog management, standards, and the ever increasing computer power are playing a major role for the industry of conversational systems.

4.2 A Change of Perspective

For many years, the goal of speech recognition research was to build machines that matched human language capability. Spoken natural-language understanding was the holy grail of much speech recognition research.

In the mid-1990s a large part of the research community was working to improve the speech recognition performance in tasks of increased complexity. In the United States, DARPA—the Defense Advanced Research Project Agency—had launched several speech research initiatives such as the Resource Management task [3], ATIS [4], the Communicator [5], and Switchboard, which was based on a corpus of free-form telephone conversations. At the same time, similar research programs were launched in Europe and Japan by different funding agencies. The goal of this research was to promote technological improvement in a field considered strategic. At the time, two start-ups appeared on this empty market landscape: Corona, later renamed Nuance, a spin-off of the Stanford Research Institute (SRI), and Altech, which became SpeechWorks (now Nuance, after the merger with Nuance in 2005), a spin-off of MIT.

While the research community was focusing on complex human-like speech-related capabilities, such as natural-language understanding of spontaneous speech, SpeechWorks and Nuance took a different approach [6]. Rather than striving for human-like performance, they focused on relatively simple tasks that could be addressed with simple word and phrase recognition. Simultaneously, they realized that *user experience* was fundamental to the achievement of commercial success. Thus, they invested most of their efforts in optimizing the engineering of the voice user interface. At the time, the user experience standard of automated telephony

interaction was set by the conventional touch-tone interactive voice response (IVR) systems (e.g., *dial one for billing; dial two for technical support*). The VUI engineered by the emerging industry was set to prove that speech recognition-based applications would provide a better experience than would traditional touch-tone IVR. Simple applications, such as package tracking, where the user is only required to speak an alphanumeric sequence, and slightly more complex applications such as stock quotes and flight status information were the first large deployments of spoken-dialog systems.

User-centered design became the focus of the new industry of dialog. Users, unaware of the complexity of speech recognition, are impatient with its quirks, and interested solely in completing their task with minimal effort. Whether they can speak to machines with the same freedom of expression as in human communication, or they are directed to provide the required information, is of little concern to the majority. Their goal is to get the job done, with minimal frustration and maximum efficiency.

It was clear that in spite of the efforts of the research community, free-form natural-language speech recognition was, in the mid-1990s, still highly error prone. On the other hand, limiting the user response to well-crafted grammars provided enough accuracy to attain high levels of task completion. That was also demonstrated by a company called Wildfire (now part of France Telecom), which in the mid-1990s commercialized a personal attendant system using extremely simple word-based speech recognition with an impressively elegant vocal interface. In a way, SpeechWorks, Nuance, Wildfire, and the other companies that followed the same paradigm, pushed back on the dream of natural-language mixed-initiative conversational machines, and engaged in a more realistic proposition that was called *directed dialog*. In directed dialog a meticulously engineered user interface *directs* the user to provide answers that can be predicted with a high degree of accuracy and thus modeled by reasonably simple, hand-crafted, context-free grammars.

At the same time of the first commercial success of the directed-dialog approach, the term *mixed initiative* [7] was acquiring popularity in academic research as a characterization of dialog systems where the user is granted a high degree of control and can, to a certain extent, change the course of the interaction. However, while widely pursued in research, it was clear that the mixed-initiative paradigm could not be readily applied to commercial systems. Mixed-initiative dialogs are difficult to design, test, and maintain, and all the possible outcomes are hard to predict. It is difficult to design and develop a mixed-initiative user interface that is consistent and complete while constraining the variability of the user expressions, and thus preserving the accuracy of the speech recognizer [8]. Since the goal of the spoken-dialog industry was that of building *usable*—and not necessarily *anthropomorphic*—systems, directed dialog was the way to go.

Travel, telecom, and financial industries were the early adopters of directed-dialog systems, while natural-language mixed-initiative communication remained mostly the goal of the academic research community. The concept of properly engineered directed-dialog speech applications soon became an effective replacement for touch-tone IVR systems, and an effective enabler for phone-based customer self-service.

4.3 Beyond Directed Dialog

In the early 2000s, the pull of the market toward commercial deployment of more sophisticated systems, and the push of new technology developed at large corporate research centers prompted the industry to move cautiously from the directed-dialog paradigm toward more sophisticated interactions. A technology first developed at AT&T and known as HMIHY (How May I Help You) [9] made its way into the simpler dialog systems. HMIHY technology, known also as *call-routing* [10], or SLU (*statistical language understanding*), consists of a statistical classifier that is trained to map loosely constrained natural-language utterances into a number of predefined classes. SLU is still far from providing sophisticated language understanding capabilities, yet its application to route incoming calls to differently skilled agents proved to be effective and in many cases outperformed traditional menu-based approaches.

The availability of sophisticated technology such as SLU that can handle spontaneous natural-language utterances—as opposed to constrained prompted words or phrases—raises questions about its applicability and effectiveness when compared with simpler deterministic context-free grammars and menu-based prompts. We should consider that the commercial purpose of spoken-dialog systems is to achieve both high rates of task completion and the best user experience, which may not result from a system that grants users more naturalness and freedom of expression. In designing commercial applications one needs to be conscious of the risks of using sophisticated natural-language understanding mechanisms when a simpler directed-dialog solution would assure similar or often better performance.

In fact, there are many applications where directed dialog, operating at today's effectiveness, outperforms systems that exhibit sophisticated natural-language understanding and mixed-initiative interactions. Most of the applications in this category are characterized by a domain model that is well understood by the population of potential users. For instance, the model for ordering pizzas is known to almost everybody: a number of pies of a certain size (small, medium, or large) with a selection of crust types and toppings (mushroom, pepperoni, etc.). The same applies to flight status information: flights can be on time, late, or cancelled; they arrive and depart daily from airports which serve one or more cities, and they can be identified by their flight number or their departure or arrival time. Banking, stock trading, prescription ordering, and many other services belong to this same category.

When the domain model is simple, and familiar to the majority of users, applications with a directed-dialog strategy are extremely effective, i.e., they achieve high completion rates plus provide good user experience. While possibly more user-restrictive, a robust directed-dialog utilizing current speech recognition technology can achieve its purpose.

We must also consider that direct user guidance, besides promoting reasonable speech recognition performance, can actually improve the effectiveness of a transaction. This is observable even in human–human interactions; for example, call centers often instruct human agents to follow precompiled scripts, similar to machine-directed dialogs. Scripted interactions help reduce the agent call handling time by increasing the effectiveness of the transactions. Moreover, as discussed in [11],

guidance of users reduces speech disfluencies. Thus, the combination of user direction, strict grammars, and less disfluency is a guiding principle of today's commercial dialog systems that help attain quite high automation rates and user satisfaction.

Contrarily, more open interactions, while seemingly more natural, could increase the breadth of user expressions at each turn, resulting in reduction in recognition accuracy. Recognition errors are a primary cause of user frustration, contributing to a negative experience. Furthermore, without direct guidance, many users are lost, knowing neither what to say nor what the system capabilities and limitations are. Based on these considerations it is clear how, for a large number of applications, well-engineered directed-dialog solutions can actually outperform free-form, natural-language, mixed-initiative dialogs.

However, there is a class of applications where directed dialog can fail and natural-language understanding is likely the right solution. Applications of this type are characterized by a domain model which is complex and unknown to the majority of users. Help desk applications, for instance, fall in this class. Consider, for instance, a system designed to route callers to the appropriate computer support desk. One obvious directed-dialog design choice would be to start by prompting users with a set of choices, like

Do you need hardware, software, or networking support?

A good number of users probably would not know which of the three categories would apply to their request. Most users are not experts—that's why they call for help to start with—and thus may have difficulty categorizing what they perceive as their problem as one out of a small number of predefined, and often obscure, categories. So their choice would be most likely instinctual and quite often wrong. An alternative model might offer a more detailed menu that would include enough guidance and choices to allow a more correct categorization. Unfortunately, depending on the complexity of the application—think for instance of how many things can go wrong in a computer—such a menu would be too lengthy to enumerate and thus impractical. In such situations, the underlying domain model is largely unknown or at best vague to the users. In this and similar situations, SLU would likely outperform directed dialog.

4.4 Architectural Evolution and Standards

The spoken-dialog systems of the mid-1990s were based on proprietary architectures. Application developers integrated speech technology resources within existing IVR and telephony platforms using proprietary APIs, and they primarily used authoring tools originally designed for the development of touch-tone applications.

As the number of companies involved in the development of spoken-dialog systems increased with the number of deployed systems, the need for industrial standards emerged quickly. Standards are extremely important for the industry and

for growing the market. They enable interoperability of architectural components, allow the reuse of modules across platforms, and reduce the risks associated with single vendor dependency.

The VoiceXML [12] language introduced in 2000 was the first standard recommendation to address the authoring of spoken-dialog applications that received wide attention from the industry. The concept of *voice* browser is at the core of the VoiceXML paradigm through the basic architecture shown in Fig. 4.1.

A voice browser, analogous to a HTML browser (see Table 4.1), requests VoiceXML documents from a Web server using the HTTP protocol. Analogous with traditional HTML documents that specify the role of input/output devices such as mouse, keyboard, and display, VoiceXML documents control speech-related input/output resources such as automatic speech recognition (ASR), text-to-speech (TTS), and audio players. A VoiceXML document includes specific sets of parameters to configure each use of the ASR, TTS, and audio player, such as links to grammars and pre-recorded prompts files, and other more detailed speech recognition and TTS parameters. A telephony interface typically conveys the incoming user's speech to the ASR, which, in turn, attempts its recognition. The results are then encoded into a set of variables that are returned to the voice browser. The voice browser, in turn, embeds the results and the URL of the next document to be fetched into a HTTP request, and sends it to the Web server.

A typical VoiceXML document represents a form with individual fields to be filled by the user voice responses returned by the ASR. While the filling of individual fields from the user in a visual Web form generally does not require the control of multiple interactions, spoken input requires the handling of interaction events such as time-outs and speech recognition rejections. Re-prompting the user and handling time-outs are integral parts of the VoiceXML language. Moreover, VoiceXML allows filling the form fields in an arbitrary order, irrespective of the order in which

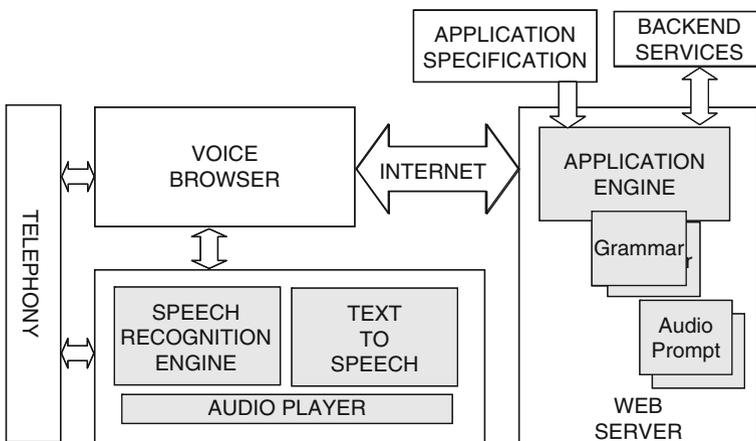
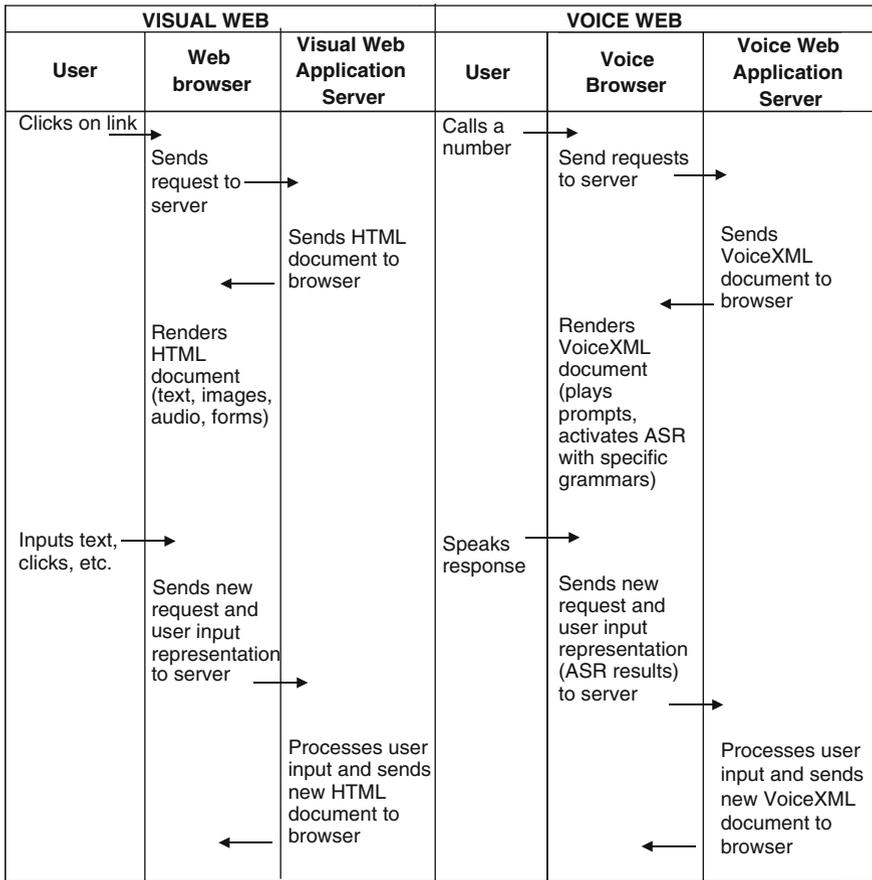


Fig. 4.1 General architecture of a modern dialog system based on the VoiceXML standard

Table 4.1 Analogy between visual and voice Web applications



they appear in the document. This procedure is controlled by the form interpretation algorithm (or FIA), the basic mechanism behind VoiceXML interpretation. The FIA was originally conceived as an attempt to allow VoiceXML developers to easily create mixed-initiative interactions. However, with the evolution of server-side application control, the mixed-initiative capability of VoicedXML is rarely used in today’s commercial applications. Developers prefer to control the dialog almost exclusively at the level of the application server.

VoiceXML language, with the FIA, the possibility of creating links to different documents, and the ability to use scripts, has the power to represent reasonably complex dialog flows. As with the early visual Web applications (also known as Web 1.0), the first VoiceXML dialog systems were implemented by sets of static documents. However, in analogy with the visual Web, developers felt that encoding the whole interaction into static documents imposed severe limitation on the more

sophisticated applications. In fact, generating dynamic VoiceXML by server-side programming—rather than by writing a set of static documents—has the advantage of providing the developer with more powerful computational and programming models than those available at the voice browser client, and the possibility of handling the dynamic nature of sophisticated interactions and business logic in a more flexible way.

Again, in analogy with the programming models of the visual Web, a second generation of VoiceXML applications emerged based on the programmatic generation of markup documents on the server, using increasingly sophisticated tools such as the J2EE and the .NET programming models. The application was encoded into programs defined by *servlets*, or sets of JSP or ASP pages. The next evolution thus prescribed the application of an MVC (model–view–controller) paradigm. The controller is typically a general dialog engine that runs on the server and interprets an application description document (the model) in order to create dynamic VoiceXML pages (the view) that are rendered by the voice browser.

While VoiceXML was the first widely adopted standards, and the spoken-dialog industry started to organize itself into different layers of competence, the need for other complementary standards became evident. For instance, as different vendors supplied core technology, such as speech recognition and text-to-speech engines, there was a need for them to communicate with the voice browser in a standard protocol. MRCP (media resource control protocol) [13] addressed that need by allowing voice browser platforms to interoperate with different ASR and TTS engines.

Another important standard that emerged after VoiceXML concerns the specification of the grammars used by the speech recognizer. Although the majority of commercial speech recognition engines used context-free grammars represented in BNF (Backus–Naur Form), their actual format was not standard across vendors, and thus prevented interoperability of the engines. Moreover, in spoken-dialog applications, there is a need to represent the semantics of all possible word strings that can be recognized. That was achieved by allowing developers to embed scripting into the grammars using an ECMAScript [14] compliant language such as JavaScript. SRGS (Speech Recognition Grammar Specification) [15], is now a widely accepted format for representing context-free grammars with semantic scripting.

Finally, it is worth mentioning other standards such as CCXML (Call Control Markup Language) [16], a language for the control of the computer-telephony layer, and SSML (Speech Synthesis Markup Language) [17] for driving the speech synthesizer.

The industry has yet to agree on a standard application language for authoring the entire spoken-dialog application. Today vendors and application integrators use their proprietary specifications. However, the W3C (the World Wide Web Consortium) created a recommendation for a state machine control language called SCXML (State Chart XML) [18] to represent a state machine execution environment based on the Harel State Chart [19] abstraction. Although SCXML is not specifically intended to represent dialog systems, it can be considered as a basic abstraction for a standard dialog application language.

4.5 The Structure of the Spoken-Dialog Industry

Figure 4.2 shows the different levels of competence present in today's spoken-dialog market. Different companies and organizations address one or several of these layers. Core technology forms the basis of the industry by providing speech recognition and text-to-speech engines, necessary for the implementation of spoken-dialog systems. Core technology is used by platform integrators to provide the necessary infrastructure for commercial dialog systems based on VoiceXML voice browsers and communication with the telephony network. The MRCP standard guarantees that the speech recognition and text-to-speech engines built by different vendors are reasonably interchangeable. Today, most platforms on the market can actually use different ASR and TTS engines providing a choice to the application builders and clients, and even allowing changing speech recognizer and text-to-speech engine at any point of the interaction.

Tools that enable application integrators to build, tune, run, and maintain full applications are the next level of competency in the industry. Again, tools can be integrated with different platforms, different flavors of VoiceXML, and adapt to different speech recognition engines.

Based on core technology and platforms, professional service organizations specialize in cost-effective processes for building applications. Initially each application was considered a new custom project, and little was shared among different systems of the same type. However, as the number of applications in each area increased, reusing segments of them for different clients became commercially attractive. Thus, a new business model emerged based on the concept of pre-packaged applications. This is justified by the fact that many applications in the same industry sector share common structures and business rules. For instance, often a self-service banking application for bank A can be adapted for bank B with minor modifications. Obviously the voice, the wording of the prompts, and the speech recognition grammars used at different stages of the dialog may be changed, but the overall structure and logic can be reused.

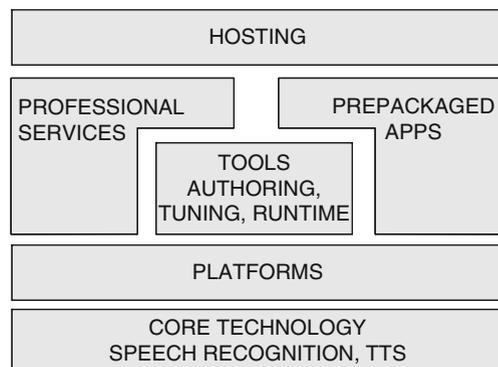


Fig. 4.2 Different levels of competence present in today's spoken-dialog market

The advantages of using pre-packaged applications are evident. The cost of adapting a basic application for each new customer rather than building it ex-novo each time is certainly smaller. But the most important advantages of the pre-packaged commercial strategy are in the reduced risk perceived by the clients and the potential continuous design improvement. Clients are more willing to adopt a tested and proved speech application rather than one that has been newly developed. Because a pre-packaged application can be continuously improved, all customers benefit from upgrades that bring better interfaces and higher task completion rates.

Hosting service companies are the top layer of the industry stack. Many purchasers of deployed spoken-dialog applications want the service to be hosted by a third party data center that assures reliable connectivity and 24/7 up time. Moreover, hosting allows for same level of service in spite of unpredicted spikes in the volume of calls, guarantees geographic redundancy during blackouts and natural disasters, allows load-balancing between the machines, and provides reliable and consolidated logging and reporting. Large companies that provided hosting for touch-tone IVR systems are today providing the same service for spoken-dialog applications with a commercial model typically based on per-minute usage.

Today several large organizations simultaneously play different vertically integrated roles. For example, core technology groups such as Nuance and IBM are also platform providers, tool suppliers, and application integrators. Similarly, hosting companies, like Convergys and Tellme, offer proprietary platforms and provide professional design, development, and application maintenance services. On the other hand, there are small companies that specialize in one or part of one level of competence, such as VUI design or tools for tuning applications. Moreover, with the current trend in mergers and acquisitions, large companies tend to acquire small and specialized companies in order to increase their offering and control the commoditization of technology.

The synergy between hosting companies and packaged application providers enabled the emergence of a new model within the paradigm of software as a service (SaaS), the leading business model of the modern Web. Companies like SpeechCycle develop key product applications for vertical industry sectors (e.g., technical support for telecommunication and internet providers). These applications are hosted by partner companies, and their services offered to client companies—the telecommunication and internet providers—who pay for their use on a per-call or per-minute pricing model.

4.6 The Speech Application Lifecycle

The pioneering companies of the spoken-dialog industry, such as SpeechWorks and Nuance, gave two clear messages to the whole community. The first is that it is indeed possible, with the available speech technology, to build spoken-dialog applications that can be deployed commercially, deliver a reasonably good user experience, and produce a compelling cost saving for the client. The second

message is that for building high-quality spoken-dialog applications one has to follow a repeatable well-defined development process based on industry-wide best practices.

Commercial applications need to be produced in a reasonable time, at a reasonable predictable cost, and with a consistent quality of the final result. Moreover, the response of the system to all possible combinations of user inputs has to be deterministically predictable, a concept that can be expressed by the idea of VUI completeness [8]. As the more general software industry has favored processes based on the experience and the skills of specialized professionals, such as designers, architect, and developers, so has the spoken-dialog industry.

Based on the general concepts of software production, the lifecycle of spoken-dialog systems thus include phases of requirement gathering, specification, development, testing, quality assurance, tuning, and phases of limited, and then full deployment. Of course, given the peculiar aspects of spoken-dialog applications, those phases need to be specialized. For instance, the specification phase requires a detailed description of the interaction that allows developers and clients to predict, with a high degree of certainty, all possible outcomes of an interaction. The spoken-dialog industry approached the specification problem by using a graphical abstraction, known as *call flow*, reminiscent of the abstractions used in touch-tone IVR applications. The development, testing and quality assurance phases typically include final tests for usability.

From the need for a clear and specialized development process based on well thought-out best practices emerged the professions of VUI designer, application speech scientist, and speech application developer. VUI designers are experts in creating usable interactions for speech applications and generally responsible for the design of the call flow, including the detailed wording of prompts and the specification of the grammars. Application design includes a specification document which represents the whole call flow as a finite state machine, with nodes described by detailed tables of properties that include prompts, grammars requirements, and all the parameters for each turn of the interaction.

All the prompts needed for the application, specified in the design document, are generally recorded by professional voice talents coached by VUI designers. Although advanced text-to-speech synthesis is available in commercial platforms, most of the applications today are deployed using high-quality recorded prompts. TTS technology is typically used when the variability of prompts is so high and unpredictable that it is problematic, or impossible, to record all the possible words or phrases.

Another responsibility of the VUI designers is to test the usability of the system once a working prototype is available. This is typically achieved by inviting a potential group of users to interact with the application within a defined scenario. The goal of usability tests is not that of assessing the accuracy of the speech recognizer or the completion rate of the system, but that of testing whether the assumptions made during design are valid, and what the main hindrances to usability are. Thorough observation and analysis of subjects' reactions can generate recommended changes in prompts, grammars, and logic of the call flow. Typically two or more usability

testing cycles using small numbers of subjects are necessary before a system can go into production.

Quality assurance (QA) is another type of test repeated every time a major update is performed on a dialog system. The goal of QA is not that of assessing usability or automation, but to guarantee that the system reflects the requirements and specifications; in other words to verify that the system does what it is supposed to do. QA is a quite complex process especially for sophisticated systems since it is hard to test all the possible paths of a call flow. QA is generally accomplished with the help of semi-automated procedures that explore the call flow, generate testing plans, and identify major development flaws.

Application speech scientists are responsible for creating, testing, and improving all the grammars required by the dialog specification. Typically grammars are tested offline using a reasonable number of samples of utterances recorded and transcribed during the early stages of deployment. The goal is to assess and optimize the accuracy of the speech recognizer in the face of the range of errors that can occur in spoken dialog (Table 4.2).

The idea of organizing a call flow into elemental objects known as dialog modules (DMs) emerged quite early in the industry. DMs encapsulate all the functionality required to capture a single piece of information—or multiple related pieces—by interacting with the user. DMs include all the actions needed in typical speech interactions, such as prompting, re-prompting, handling timeouts, disambiguating, etc. Core technology vendors such as SpeechWorks and Nuance have made libraries of pre-built configurable DMs available to developers together with the option of building custom ones. DMs for the collection of digits, natural numbers, social security numbers, money amounts, credit card numbers, dates, names, addresses, and many other types of information are available off-the-shelf. They can be configured at the level of the language (e.g., English, Spanish, Chinese), style of interaction (e.g., confirm at each turn, confirm only for low confidence), resolution of the grammars (e.g., for money amounts, whether the quantities are in dollars and cents, hundreds of dollars, thousands or tens of thousands of dollars), range of values (e.g., for dates whether the request is in the past or future). Moreover, the leading vendors integrated DMs into different programming environments, and made them

Table 4.2 Types of speech recognition errors in dialog systems

Utterance is:	Recognized string is:	Recognition result is:	Error Severity
In Grammar	Corrected	Accepted	Low
		Confirmed	
		Rejected	
	Wrong	Accepted	High
		Confirmed	Medium
		Rejected	
Out of Grammar	Wrong	Accepted	High
		Confirmed	Medium
		Rejected	

available in different forms, such as C++ objects or ActiveX controls. Most recently, after the advent of VoiceXML, DMs are distributed as Web objects, such as JSP pages, that generate dynamic VoiceXML documents. DMs and call flows based on them created a powerful model for the development of dialog systems. However, few commercial dialog systems were built in native code, such as C++, or Visual Basic. Developers primarily used tools provided by the platform companies. In the beginning they were tools evolved from the touch-tone IVR platforms, which with time and the evolution of standards such as VoiceXML, became more and more geared to spoken-dialog abstractions and the corresponding lifecycles. Today systems are generally built based on programming models which are typical of the Web development world. It is not uncommon to find dialog systems developed using Web programming abstractions such as *servlets*, Java Server Pages (JSP), and STRUTS, from the J2EE paradigm, or analog concepts from the Microsoft .NET programming environment.

4.7 Speech 3.0: The Third Generation of Spoken-Dialog Systems

Since the start of the spoken-dialog industry about a decade ago, we have seen the deployment of applications with increasing levels of complexity and sophistication. The first types of applications (Table 4.3) were mostly *informational*: the system returns information based on the user input. Package tracking and flight status applications were previously mentioned. The underlying model of this first generation of application is that of a form. Once all the fields—or a relevant subset of them—are filled, the form is used to query a database, and the result presented to the user. In fact the form-filling paradigm prompted the evolution of VoiceXML and its form interpretation algorithm.

Table 4.3 Evolution of the dialog industry

	GENERATION		
	FIRST	SECOND	THIRD
Time Period	1994–2001	2005–2005	2004–today
Type of Application	Informational	Transactional	Problem Solving
Examples	Package Tracking, Flight Status	Banking, Stock Trading, Train Reservation	Customer Care, Technical Support, Help Desk.
Architecture	Proprietary	Static VoiceXML	Dynamic VoiceXML
Complexity (Number of DMs)	10	100	1000
Interaction Turns	few	10	100
Dialog Modality	directed	directed+natural language (SLU)	directed + natural language (SLU) + limited mixed initiative

The second generation of spoken-dialog systems implemented more complex transactional applications which involve some form of negotiation, like buying a train ticket, or trading stocks. A simple form-filling algorithm may not be enough to drive the interaction. Several intermediate forms may be needed, and there may be a certain divergence from a fixed script. For instance, the user of a train reservation application may need to browse the departure times and the prices before deciding which ticket to buy.

The evolution from first- to second-generation systems corresponds to a quantifiable increase in application complexity, which can be measured by the size of the equivalent *call flow*, i.e., the graph representing the progression of the interaction. Call flows of first- and second-generation applications include a number of nodes, each node roughly corresponding to a DM, formerly in the tens, and then in the hundreds. Today, third-generation applications may reach thousands of call-flow nodes and handle sophisticated interactions that easily transcend the filling of forms. Third-generation dialog systems, or *Speech 3.0*, contain problem-solving procedures often requiring much sustained user interaction. The development, monitoring, and tuning of this new generation of applications pose interesting questions at all levels. New sets of tools and models are needed, and they are rapidly evolving to the next level of industry standards. Modularity, encapsulation, inheritance, configurability, and reusability are fundamentals that help manage complex code structures, and can be directly applied to call-flow-based dialog specifications.

Moreover, the traditional lifecycle of design specification followed by a development phase—the so called *waterfall* model—cannot be applied to complex dialog systems of the third generation, mainly because they typically require rapid modifications. So, while new abstractions are needed and explored, another trend is emerging where design, specification, and development are unified. This *iterative* development lends itself to the concept of *perpetual beta*, a key philosophy inherent in the evolution of Web 2.0 [20]. Specialized tools that address both phases of design and development by keeping a unique representation of the dialog artifacts, and allowing collaborative development, are keys to the evolution of third-generation applications.

Automated technical support [21] is a typical example of a Speech 3.0 application. Technical support spoken-dialog systems for products or services like internet, cable TV, or VoIP telephony require extensive logic to diagnose the problem based on symptom descriptions in natural language, deep integration with customer databases and diagnostic tools, extensive knowledge base of facts and resolution strategies, and a continuous improvement process for tracking new products and problems and improving the voice user interface and the resolution effectiveness. The intelligence needed to perform technical support, troubleshooting, and problem solving transcends the complications of speech recognition and dialog. That intelligence needs to be seamlessly integrated with the voice user interface in order to achieve superior user experience and high automation rates. The main challenge of today's spoken-dialog industry consists in creating technology, tools, and processes which allow building and maintaining such sophisticated and complex systems for an increasingly large number of customers and, yet, maintain a sustainable business.

4.8 Conclusions

Automatic speech recognition research, which begun more than 50 years ago, found commercial deployment within a structured and maturing market only during the past decade. The vision of building machines that mimic human speech interaction neither is yet a reality nor has been abandoned, just pushed back in favor of a more pragmatic use of the technology. What enabled the industry to move toward commercial success was the realization that users do not necessarily need a full replication of human-like speech and language skills in a system, but rather a good user experience and effective task completion. In fact, market adoption of spoken-dialog systems is driven more by success factors such as the quality of user experience and task completion rate, and not by the freedom of linguistic expression. Highly engineered solutions focused on the delivery of effective transactions proved to be instrumental to the development of the industry of spoken-dialog systems.

Today we are seeing the evolution of a third generation of dialog systems which exhibit a much higher level of performance and sophistication than that developed during the last decade. Still, the complexity did not evolve solely from the ability to fully understand rich natural language and manage user initiative in dialog. Rather, managing the intrinsic complexity of the application and its logic is at the core of this evolution.

Finally, an important consideration when discussing the use and usability of commercial dialog systems is that their success is contingent on the willing cooperation of users. Even the best-designed application, with the most advanced architecture, fails when callers refuse to use it. Today the problem is that the users of commercial dialog systems remain frustrated when faced with a computer rather than a live agent. This is true especially for those systems developed without proper consideration for usability and good caller experience. This can be proven by the existence of *cheat sheets* published on the Web, which suggest keywords and touch-tone keys to bypass automated systems and more or less instantly produce a human agent.

The fact is that most consumers are ignorant of the vast and true costs of proper service. We will reach a point in time where the need for human agent far exceeds their availability; long waiting times even now are common for the simplest requests. The cost of training and maintaining human resources at off-shore call centers—a common solution for cost reduction in the United States and other English-speaking countries—may soon produce diminishing returns, unbearable even by the largest companies. Non-English-speaking countries present larger challenges and costs. The situation is reminiscent of the early 1900s, before the introduction of the automatic telephone switching, when AT&T predicted that the rate of telephone use would exceed the required number of switch operators, more than the country's population

The switchboards were something to behold, with many, many operators sitting in long rows plugging countless plugs into countless jacks. The cost of adding new subscribers had risen to the point foreseen in the earlier days, and that cost was continuing to rise, not in a direct, but in a geometric ratio. One large city general manager wrote that he could see

the day coming soon when he would go broke merely by adding a few more subscribers (http://www.bellsystemmemorial.com/capsule_bell_system.html)

Automated self-service dialog systems enabled by speech recognition constitute a solution to the need for service, especially if cost alternatives mean no service at all. Public rejection of automated self-service technology is not new; witness early resistance to the answering machine and the ATM banking center. The convenience these now represent should be viewed as complementary rather than as replacements, since people have been reutilized within these industries in other capacities. Similarly, dialog systems are tools that, if used properly, can be a positive force in an evolving service-oriented society.

References

1. Davis, K., Biddulph, R., Balashek, S. (1952). Automatic recognition of spoken digits. *Soc. Am.*, 637–642.
2. Flanagan, J. L., Levinson, S. E., Rabiner, L. R., Rosenberg, A. E. (1980). Techniques for expanding the capabilities of practical speech recognizers. In: *Trends in Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ.
3. Price, P., Fisher, W. M., Bernstein, J., Pallet, D. S. (1988). The DARPA 1000 word Resource Management database for continuous speech recognition. In: *IEEE Conf. on Acoustics Speech and Signal Processing*.
4. Hirschmann, L. (1992). Multi-site data collection for a spoken language corpus. In: *Proc. 5th DARPA Speech and Natural Language Workshop*. Defense Advanced Research Projects Agency.
5. Walker, M., Rudnicky, A., Aberdeen, J., Bratt, E. O., Carofolo, J., Hastie, H., Le Audrey Pellow, B., Potamianos, A., Passonneau, R., Prasad, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D. (2002). DARPA communicator: Cross system results for the 2001 evaluation. In: *ICSLP 2002*.
6. Barnard, E., Halberstadt, A., Kotelly, C., Phillips, M. (1999). A consistent approach to designing spoken-dialogue systems. In: *IEEE Workshop*. Keystone, CO.
7. Zue, V. (1997). Conversational interfaces: Advances and challenges. In: *Eurospeech 97*. Rhodes, Greece.
8. Pieraccini, R., Huerta, J. (2005). Where do we go from here? Research and commercial spoken dialogue systems. *SIGdial*, 1–10.
9. Gorin, A. L., Riccardi, G., Wright, J. H. (1997). How may i help you? *Speech Commun.*, 113–127.
10. Chu-Carroll, J., Carpenter, B. (1999). Vector-based natural language call routing. *Comput. Linguist.*, 361–388.
11. Oviatt, S. L. (1995). Predicting spoken disfluencies during human–computer interaction. *Comp. Speech Lang.*, 19–35.
12. Voice Extensible Markup Language (VoiceXML) 2.1. (2005). W3C Candidate Recommendation 13 June 2005.
13. Media Resource Control Protocol (MRCP) Introduction.
14. Standard ECMA-262 (1999). ECMAScript Language Specification, 3rd Edition.
15. Speech Recognition Grammar Specification Version 1.0. (2004). W3C Recommendation.
16. Voice Browser Call Control: CCXML Version 1.0. (2005). W3C Working Draft.
17. Speech Synthesis Markup Language (SSML), Version 1.0. (2004).
18. State Chart XML (SCXML) State Machine Notation for Control Abstraction. (2006). W3C Working Draft.

19. Harel, D., Politi, M. (1998). Modeling Reactive Systems with Statecharts: The STATEMATE Approach. McGraw-Hill, New York, NY.
20. O'Reilly, T. (2004). What is Web 2.0. In: Design Patterns and Business Models for the Next Generation of Software. W3C Recommendation.
21. Acomb, K., Bloom, J., Dayanidhi, K., Hunter, P., Krogh, P., Levin, E., Pieraccini, R. (2007). Technical support dialog systems, issues, problems, and solutions. In: Bridging the Gap, Academic and Industrial Research in Dialog Technology. Rochester, NY.

Chapter 5

Deceptive Speech: Clues from Spoken Language

Julia Hirschberg

5.1 Introduction

Deception is generally defined as ‘a deliberate attempt to mislead others’ [1]; thus, deceivers are those trying to convince others that something the deceivers know to be false is in fact true. Distinguishing the behavior of deceivers from that of truth-tellers has long been a topic of interest, both to behavioral scientists and to law enforcement personnel. To the latter, research into deception may help to improve human performance in deception detection by pointing out reliable cues to deceptive behavior [2]. However, numerous studies have found that, even when subjects are trained to look for cues highly correlated with deception, many still perform very poorly at deception detection [3, 4]. Alternatively, research into the objectively observable correlates of detection may help us to develop new technologies to identify such behavior automatically, augmenting current technology such as the polygraph as a tool for deception detection.

To date, most practical and scientific deception studies have focused on visual cues to deception, such as facial expressions (e.g., [5]) or body gestures (e.g., [6]), or on traditional biometric cues used currently in polygraphy (e.g., [7]). While studies associating the detection of vocal indicators of stress with deception have been popular in the recent past [8, 9], current implementations of voice stress analysis technologies have not proven useful in and of themselves in separating deceivers from truth-tellers.

In recent years, there has been considerable interest in the speech community in the automatic identification of affective speech [10]. Promising research is currently underway on the application of corpus-based machine-learning approaches to the detection of certain emotions, such as anger or uncertainty, in spoken dialogue systems in such venues as call centers or tutoring systems [11–15]. Such research has sparked new interest in using similar techniques to identify other types

J. Hirschberg (✉)

Department of Computer Science, Columbia University, New York, NY, USA
e-mail: julia@cs.columbia.edu

of speaker state which appear to be related to the classic emotions, such as deception [16–20]. Deception has been associated in the psychological literature [21] with manifestations of both fear of detection and elation at not being detected.

A major difficulty of studying deception in any channel is the variety of factors which may influence any act of deception. ‘White’ lies in social settings, where the consequences of detection are small, differ from ‘serious’ lies, which themselves may also vary in terms of whether the deceiver is hiding a transgression or lying for a cause deemed to be worthy. Studies of ‘high-stakes’ lies today suffer from the difficulty of creating realistic scenarios in the laboratory. Due to concerns for the protection of human subjects, most laboratory research is conducted with subjects motivated to lie via financial or ‘self-presentational’ incentives. In the latter, subjects are persuaded that their ability to deceive is a positive quality [1]. One popular scenario currently is the ‘mock-theft’ paradigm, in which subjects are given the option of taking a check or not, and then interrogated about this [22]. Other scenarios may involve asking subjects to lie about the content of a movie they are watching or about the number they see on a card.

Age and culture also play an important role in the feelings subjects have about lying and thus, in the auditory and visual manifestations of their deceiving. Even within an age and cultural group, there appears to be wide variation in the cues people exhibit or recognize as indicators to deception. However, such differences, while recognized, have as yet been little studied.

A primary obstacle to studying deception in speech is the lack of cleanly recorded corpora of deceptive and non-deceptive speech to use for training and testing. Speech collected during previous experiments or in the field during actual interviews is difficult to analyze due to poor recording conditions. While early studies were better able to utilize scenarios with ‘high-stakes’ deception (in which subjects could be motivated by fear or shame) in the laboratory [23], more recent studies are limited to less stressful scenarios by human subjects’ protocols and privacy considerations.

Most researchers and practitioners would agree that there is no single cue to deception, but that multiple indicators should be examined. While few studies have focused on spoken cues, there has been considerable work on lexical and semantic indicators of deception, generally coded by trained annotators or otherwise subjectively labeled. Below we examine some of the previous literature on deceptive speech and language, including the features and data that have been examined and findings from these, and describe current computational work in the area of deception detection from spoken cues and text-based cues.

5.2 Perceptual and Descriptive Studies of Deception

Studies of deceptive speech and language by behavioral scientists have focused primarily upon human perception of deception and descriptive analyses of variation in a number of dimensions manifest in the deceiver’s speech or written statements. These studies provide valuable information on human perception of cues

to deception and occasionally on how human judgments correlate with more objective measures of cues in the speech signal. However, many of the hypotheses and findings are contradictory, perhaps due to variation in the motivation of the deceivers studied, to the amount of prior preparation spent devising the lie, to individual differences among speakers, or to the way in which particular features have been defined in different studies.

Deceivers have been hypothesized to speak more than truth-tellers or to speak less [23, 24], depending perhaps upon the care with which the lie has been prepared in advance of the telling or the desire of the deceiver to ‘hold back’ information. They have also been thought to exhibit more response latency or less, for similar reasons [25–27]; over-rehearsed deceivers may give themselves away by answering particular questions too quickly, while under-rehearsed deceivers may need to spend more time thinking about the lie they are concocting. Deceivers have been observed to speak louder or softer when lying, to speak with higher or lower pitch [28, 29] or with a faster or slower speaking rate [23, 27], and to exhibit more vocal tension and less vocal ‘pleasantness.’ Studies have found that deceivers exhibit fewer disfluencies or more than truth-tellers, again perhaps depending upon the amount of rehearsal of their stories [23, 26, 27, 30]. On similar grounds that rehearsed lies differ from normal truth-telling, deceivers are thought to make fewer admissions of forgetfulness than truth-tellers. Less well-rehearsed deceivers are said to appear less confident, to provide fewer details and scene descriptions, to be less plausible and logical in their stories, to produce more repetitions, to use more passives, negations, and ‘indirect’ speech (e.g., attributing actions and opinions to we or they), to provide fewer details, to exhibit less cognitive complexity in their speech, and to stray from the topic more frequently by mentioning peripheral events or relationships [26, 31–33]. Many of these features are captured in various coding schemes, such as Vrij et al.’s [26] NVB coding of nonverbal behaviors of gaze, gesture, disfluencies, response latency, and speaking rate; CBCA (criteria-based content analysis), which encodes lexical content [34]; and RM (reality monitoring), which codes perceptual, cognitive, and affective information identified in subjects’ statements [32, 35].

While some similarities have been found across studies, it is not clear how a number of these features can be objectively measured; even cues which are objectively measurable must be calibrated against a speaker-dependent baseline, which may be difficult to obtain in practice. Practitioners typically explain that they spend a good portion of an initial interview determining whether a speaker normally exhibits such behaviors as avoiding eye gaze; for these speakers, making eye contact may arouse suspicion in subsequent interrogation, while for those who do not avoid eye contact normally gaze avoidance during interrogation might be seen as suspicious [36]. And most features involving what is said must be coded or otherwise interpreted by a human agent with some skill.

DePaulo et al.’s [1] meta-study of cues to deception provides an excellent survey of 158 hypothesized indicators and 1,338 separate estimates from previous studies. This useful study compiles results from within subject experiments in which adult subjects were observed both lying and telling the truth, where potential cues to deception were either measured objectively in some way or were rated

impressionistically by humans, in an attempt to determine which cues represent statistically significant discriminators of deceptive from non-deceptive behavior when examined across all studies which include them as factors. DePaulo et al. [1] examine the significance of individual cues in support of five basic hypotheses about deceivers: (1) Deceivers are less forthcoming than truth-tellers (they 'hold something back'). (2) Deceivers' stories are less compelling in terms of the fluency and plausibility of their narrative; they tend to be less convincing than truth-tellers overall. (3) Deceivers appear less positive and pleasant than truth-tellers, in terms of what they say and how they say it. (4) Deceivers appear tense, due to the cognitive load of maintaining a consistent lie or to fear of discovery. (5) For similar reasons, deceivers may include more imperfections in their tales, or they may include fewer, due to prior rehearsal of what they plan to say. While many of the cues examined in these categories are facial and body gestures, a number of possible speech and language cues are included, so it is instructive to note which of these cues are borne out across studies.

With respect to acoustic and prosodic cues to deception, DePaulo et al. [1] found that, across the studies they examined, there was evidence of a significant difference between deceivers and truth-tellers in the proportion of overall talking time deceivers spoke versus their conversational partner, with deceivers speaking significantly less than truth-tellers. Deception was also negatively correlated with observer impressions of 'verbal and vocal involvement' and with observer ratings of vocal pleasantness (e.g., [6]), while it was positively correlated with impressions of 'verbal and vocal uncertainty.' Overall rater impressions of tenseness were positively correlated with deception, with both vocal tension and higher pitch being positively correlated. Note that Streeter et al. [29] found stronger correlations between high pitch and deception for subjects more highly motivated to deceive.

However, factors such as overall response length, length of interaction, response latency, loudness, and speaking rate, which have also been proposed as potential cues to discriminating deceptive from non-deceptive speech, did not show significant differences in this meta-study. Note that Baskett [25] reports that listeners were more likely to judge speakers to be liars if they answered 'too quickly' or 'too slowly,' which may wash out differences in this cue. Mehrabian [23] found similar convincing evidence for speaking rate across studies, with rate generally increasing as the speaker's comfort level increased. So these features may require more sophisticated modeling, perhaps based upon individual differences in normal production, to prove useful. Note also that Gozna and Babooram [27] found that whether subjects were seated or standing during an interview affected differences between deceptive and non-deceptive behaviors, such that speaking rate increased during deception in the standing condition but not in the seated condition, while stutters decreased only in liars who were standing; in this study response latency decreased for deceivers in both conditions. So the context of the deceptive situation appears to play an important role in the behavioral cues deceivers exhibit.

With respect to speech disfluencies (including filled and silent pauses and hesitations), often thought to mark the speech of at least the less-rehearsed deceiver, DePaulo et al. [1] did not find evidence for this across studies; in fact, they found

that deceivers tended to make significantly fewer ‘spontaneous self corrections.’ Note also more recent work on filled and silent pauses as cues to deception by Benus [30], which shows a positive correlation between these pauses and truth-telling.

Examining lexical and semantic cues to deception, as coded by human raters, DePaulo et al. [1] found support across studies for claims that deceivers’ productions are less plausible and fluent than those of truth-tellers in a number of categories hypothesized in the literature: Deceivers did provide significantly fewer details than truth-tellers and tended to make significantly more negative statements and complaints. They also tended to repeat words and phrases more often than truth-tellers did. Deceivers made fewer admissions of lack of memory and fewer expressions of self-doubt. They were significantly more likely to mention extraneous material in their speech than truth-tellers. In general, there were significant negative correlations between deception and observer ratings of the plausibility of deceivers’ stories and their logical structure, and there were significantly more discrepancies and ambivalent statements in their narratives.

For other hypothesized cues to deception in this category, DePaulo et al.’s [1] study found no significant correlations with deception. These included the proportion of unique words used by deceivers, their use of generalizing terms, self-references or mutual or group references, the use of tentative constructions (e.g., ‘I think’), the amount of unusual or superfluous detail they provided, their discussions of speaker’s or listener’s mental state, the amount of sensory information they provided (coded using RM), and the cognitive complexity of their output.

However, it is important to note that even though DePaulo et al. [1] found no significant correlations of many hypothesized cues across the studies they included, individual studies have found these features to be useful cues to deception, either alone or in combination with other features. And more recent work has been done on some of them, which of course was not included in this meta-study. It is also difficult to combine studies of individual cues which may be subject to different definitions and interpretations, particularly when these cues are measured perceptually rather than objectively. So, while DePaulo et al.’s [1] results are useful, they clearly do not rule out potential cues to deception.

5.3 Practitioners’ Lore

There is also some literature and much lore among members of law enforcement agencies and the military to identify various practical auditory and lexical cues to deception for use by interviewers and interrogators. The most commonly cited oral cues for these practitioners include longer or shorter response latency, filled pauses and other disfluencies, and repetitions [36]. In most cases these cues are intended to be calibrated against a ‘norm’ for an individual being questioned; such norms are typically established while asking interviewees questions they are likely to answer truthfully, depending upon the purpose of the interview, such as *What is your name?* or *What is today’s date?* Considerable weight is also given to detection of deception

by a close analysis of the lexical and syntactic choices of suspects' oral (transcribed) or written statements, calibrated here against a general 'normal' usage developed by practitioners over years of experience. Statement analysis [37] is one of the best-documented versions of this approach. Designed as a tool for interrogators, this approach looks for deviation from 'normal' use of pronouns and verb tense as well as hedges (e.g., *I think*) and memory lapses in critical positions in the narratives elicited. For example, explicit use of the first person pronoun rather than a more general attribution or the absence of any subject (e.g., *Went to the bank*) is deemed normal in narrative; failure to pronominalize on subsequent mention (e.g., repeating *My wife and I* rather than using *we*) is deemed abnormal. Changes in tense during a narrative, as from past to present, are also seen as suspect, indicating a place in the statement where subjects may not be telling the truth. Truthful subjects are believed to recount events chronologically and concisely, while liars will not. Such analyses must currently be performed by trained interviewers.

5.4 Computational Approaches to Deceptive Speech

Corpus-based, machine-learning approaches to detecting deception via automatically extractable objective features have been rare, in part due to the absence of corpora recorded under suitable conditions and labeled for truth or lie. One exception is work on voice stress analysis, which assumes that indicators of vocal stress also indicate deception, but this hypothesis has not been supported in experimental testing, although features examined for VSA analysis may eventually prove to be useful in combination with other features.

5.4.1 Lexical and Semantic Analysis

There has been some attempt to automate a simple form of lexical analysis of deceptive text in a program called Linguistic Inquiry and Word Count (LIWC), developed in the 1990s [38, 39]. LIWC computes the percentage of words in a text that fall in one of 72 different categories, to capture 'negative' emotion, degree of self-reference, and indicators of cognitive complexity, under the hypothesis that liars exhibit more of the first and less of the second two. Using this keyword-based analysis, Newman et al. [39] report classifying liars versus truth-tellers at an overall accuracy rate of 61%.

5.4.2 Voice Stress Analysis

Voice stress analysis (VSA) approaches rely upon low-level indicators of stress such as microtremors or vocal jitter, as indirect indicators of deception. There has been little evidence that VSA systems can effectively discriminate deception from non-deceptive speech [8], although Hopkins et al. [9] have found that such systems might

be useful tools for a skilled examiner. Liu [17] recently tested the utility of jitter versus other features as discriminators for deception and found that, while jitter did not discriminate, pitch did, although only in a speaker-dependent manner. However, VSA systems continue to be marketed widely to law enforcement agencies as the answer to their deception detection problems.

5.5 Machine-Learning Approaches

Recently, there has been interest in applying machine-learning techniques to the problem of deception detection from speech, seeking to test which of the many features proposed in the behavioral literature might be (a) objectively measurable and (b) useful discriminators.

Qin et al. [40] have described preliminary studies using decision trees trained on lexical information to predict deception. Cues included numbers of syllables, words, sentences, short sentences, and ‘simple’ sentences; measures of word and sentence complexity; indicators of specificity and expressiveness; and an ‘informality’ measure based on errors that were automatically detectable. Results for the best-performing decision trees examined from 20 cross-validation runs on a very small data set are reported in the mid-high 70% range.

Work has also been underway to apply speech technologies and machine-learning techniques to a new, cleanly recorded corpus of deceptive speech, the Columbia-SRI-Colorado (CSC) Corpus [16, 20, 30]. This corpus was designed to elicit within-speaker deceptive and non-deceptive speech [16]. The corpus includes interviews with 32 native speakers of Standard American English. Subjects performed tasks in six areas, where the difficulty of tasks was manipulated so that interviewees scored higher than an artificial profile in two areas, lower in two, and identically in another two. Subjects received financial and self-presentational incentives to convince an interviewer that they had in fact performed the same as the target profile. Subjects were instructed to press one of two pedals hidden from the interviewer after each statement, one pedal for truth and one for lie to capture ground truth. The interviews lasted between 25 and 50 min, and comprised approximately 15.2 h of dialogue; they yielded approximately 7 h of subject speech. Data was recorded using head-worn microphones in a sound booth and was subsequently orthographically transcribed. Several segmentations were created from the data: the implicit segmentation of the pedal presses, which was hand-corrected to align with corresponding sets of statements; word segments, from the automatic alignment of the transcription using an SRI ASR engine; hand-labeled sentence-like units [41]; and ‘breath groups’ which were identified from ASR word alignments plus intensity and pauses, and subsequently hand-corrected. The corpus thus consists of lexical transcription, global and local lie labels, segmentations, and the speech itself.

A series of machine-learning experiments employing different learning algorithms and a variety of features sets and segmentations on this corpus have achieved classification accuracies of 66.4% [16], using a combination of acoustic/prosodic, lexical, and speaker-dependent features, and 64.0% using acoustic/prosodic features

alone [20]. A human perception study performed on this data found that human judges asked to determine whether each statement was truth or lie scored on average worse than chance. Thus, the automatically produced results are quite encouraging.

5.6 Conclusion

The current state of deception studies from speech and language cues remains largely the domain of behavioral scientists conducting laboratory studies, which peripherally include vocal cues, and of practitioner proponents of various types of text-based statement analysis. Larger machine-learning studies combining speech and text-based cues with potential facial, gestural, and biometric cues to deception have yet to be undertaken, largely due to the lack of corpora which include clean data from each potential cue dimension and which can be reliably labeled for truth or lie. The investigation of machine-extractable rather than hand-coded or impressionistic cues also suffers from this lack, since insufficient data for training and testing of such features is lacking. Furthermore, data used in most current research on deception is collected from subjects whose motivation for deception is probably very different from that of deceivers in the real world, and scenarios closer to real life which will nonetheless be accepted by institutional review boards are hard to devise. 'Real' data, collected by law enforcement agencies, is rarely recorded under conditions sufficient to do adequate acoustic/prosodic analysis, although, when transcribed, it may suffice for those focusing on lexical information if ground truth (was the subject really lying or not) can be reliably established. Using such data, where it is available, also involves resolving serious ethical and legal issues. Investigation of the importance of individual and cultural differences in deception, another major area that importance is generally acknowledged, has rarely been undertaken.

In sum, the field of deception studies presents abundant open questions for research. Answering these questions, however, requires the resolution of some very difficult data collection and annotation questions, involving both technical and ethical/legal issues. It is likely that current security concerns will provide powerful incentives for finding solutions to these issues, but it is also likely that solutions to the problem of detecting deception will be proposed that have not been scientifically tested, due to the difficulty of such testing. For these reasons, it is important for behavioral scientists and speech and language technologists to work together to ensure that deception detection itself is not deceptive.

References

1. DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., Cooper, H. (2003). Cues to deception. *Psychol. Bull.*, 129, 74–118.
2. Frank, M. G., Feeley, T. H. (2003). To catch a liar: Challenges for research in lie detection training. *J. Appl. Commun. Res.*, 31(1), 58–75.

3. Vrij, A. (1994). The impact of information and setting on detection of deception by police detectives. *J. Nonverbal Behav.*, 18(2), 117–136.
4. Aamodt, M. G., Mitchell, H. (2004). Who can best detect deception: A meta-analysis. Paper presented at the Annual Meeting of the Society for Police and Criminal Psychology, Rome.
5. Ekman, P., Friesen, W. V. (1976). *Pictures of Facial Affect*. Consulting Psychologists Press, Palo Alto, CA.
6. Burgoon, J. K., Buller, D. B. (1994). Interpersonal deception: iii. Effects of deceit on perceived communication and nonverbal behavior dynamics. *J. Nonverbal Behav.*, 18(2), 155–184.
7. Horvath, F. (1973). Verbal and nonverbal clues to truth and deception during polygraph examinations. *J. Police Sci. Admin.*, 1(2), 138–152.
8. Haddad, D., Ratley, R. (2002). Investigation and evaluation of voice stress analysis technology. Technical report, National Criminal Justice Reference Service.
9. Hopkins, C. S., Ratley, R. J., Benincasa, D. S., Grieco, J. J. (2005). Evaluation of voice stress analysis technology. In: Proc. 38th Hawaii Int. Conf. on System Sciences, Hilton Waikoloa Village Island of Hawaii.
10. Cowie, R., Douglas-Cowie, E., Campbell, N., eds. (2003). *Speech communication: Special issue on speech and emotion*.
11. Lee, C. M., Narayanan, S., Pieraccini, R. (2002). Combining acoustic and language information for emotion recognition. In: Proc. Int. Conf. on Spoken Language Processing 2002, Denver, 873–876.
12. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proc. Int. Conf. on Spoken Language Processing, Denver, 2037–2039.
13. Batliner, A., Fischer, R., Huber, R., Spilker, J., Nöth, E. (2003). How to find trouble in communication. *Speech Commun.*, 40(1-2), 117–143.
14. Litman, D., Forbes-Riley, K. (2004). Predicting student emotions in computer-human dialogues. In: Proc. ACL-2004, Barcelona.
15. Liscombe, J., Venditti, J., Hirschberg, J. (2005). Detecting certainty in spoken tutorial dialogues. In: Proc. INTERSPEECH 2005, Lisbon.
16. Hirschberg, J., Benus, S., Brenier, J. M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L., Pellom, B., Shriberg, E., Stolcke, A. (2005). Distinguishing deceptive from non-deceptive speech. In: Proc. INTERSPEECH 2005, Lisbon.
17. Liu, X. (2005). Voice stress analysis: Detection of deception. Master's Thesis at the University of Sheffield, Department of Computer Science.
18. Fadden, L. (2006). The prosody of suspects' responses during police interviews. In: *Speech Prosody 2006*, Dresden.
19. Enos, F., Benus, S., Cautin, R. L., Graciarena, M., Hirschberg, J., Shriberg, E. (2006). Personality factors in human deception detection: Comparing human to machine performance. In: Proc. INTERSPEECH 2006, Pittsburgh, PA.
20. Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., Kajarekar, S., Haddad, D., Ratley, R. (2006). Combining prosodic lexical and cepstral systems for deceptive speech detection Investigation and evaluation of voice stress analysis technology. In: Proc. ICASSP–2006 toulouse.
21. Ekman, P. (1992). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. Norton, New York, NY.
22. Frank, M. G. Ekman, P. (1997). The ability to detect deceit generalizes across different types of high stake lies. *J. Personality Social Psychol.*, 72, 1429–1439.
23. Mehrabian, A. (1971). Nonverbal betrayal of feeling. *J. Exp. Res. Personality*, 5, 64–73.
24. Harrison, A. A., Hwalek, M., Raney, D. F., Fritz, J. G. (1978). Cues to deception in an interview situation. *Social Psychol.*, 41, 156–161.
25. Baskett, G. D. a R. O. F. (1974). Aspects of language pragmatics and the social perception of lying. *J. Psycholinguist. Res.*, 117–130.
26. Vrij, A., Edward, K., Roberts, K. P., Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *J. Nonverbal Behav.*, 24(4), 239–263.

27. Gozna, L. F., Babooram, N. (2004). Nontraditional interviews: Deception in a simulated customs baggage search. Paper presented at the 14th European Conference of Psychology and Law, Krakow, Poland, July 7–10.
28. Ekman, P., Friesen, W. V., Scherer, K. R. (1976). Body movement and voice pitch in deceptive interaction. *Semiotica*, 16(1), 23–77.
29. Streeter, L. A., Krauss, R. M., Geller, V., Olson, C., Apple, W. (1977). Pitch changes during attempted deception. *J. Personality Social Psychol.*, 35(5), 345–350.
30. Benus, S., Enos, F., Hirschberg, J., Shriberg, E. (2006). Pauses in deceptive speech. In: *Speech Prosody 2006*, Dresden.
31. Wiener, M. Mehrabian, A. (1968). *Language within Language: Immediacy, a Channel in Verbal Communication*. Appleton-Century-Crofts, New York, NY.
32. Zuckerman, M., DePaulo, B. M., Rosenthal, R. (1981). *Verbal and Nonverbal Communication of Deception*. Academic Press, New York, NY, 1–59.
33. Zapamiuk, J., Yuille, J. C., Taylor, S. (1995). Assessing the credibility of true and false statements. *Int. J. Law Psychiatry*, 18, 343–352.
34. Steller, M. Koehnken, G. (1989). *Criteria Based Content Analysis*. Springer-Verlag, New York, NY, 217–245.
35. Masip, J., Sporer, S. L., Garrido, E., Herrero, C. (2005). The detection of deception with the Reality Monitoring approach: A review of the empirical evidence. *Psychology, Crime, Law*, 11(1), 99–122.
36. Reid, J. E. and Associates (2000). *The Reid Technique of Interviewing and Interrogation*. Chicago: John E. Reid and Associates, Inc.
37. Adams, S. H. (1996). Statement analysis: What do suspects' words really reveal? *FBI Law Enforcement Bull.* October, 1996.
38. Pennebaker, J. W., Francis, M. E., Booth, R. J. (2001). *Linguistic Inquiry and Word Count*. Erlbaum Publishers, Mahwah, NJ.
39. Newman, M. L., Pennebaker, J. W., Berry, D. S., Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. *Personality Social Psychol. Bull.*, 29, 665–675.
40. Qin, T., Burgoon, J. K., Nunamaker, J. F. (2004). An exploratory study on promising cues in deception detection and application of decision tree. In: *Proc. 37th Annual Hawaii Int. Conf. on System Sciences*, 23–32. Big Island, Hawaii, USA.
41. NIST (2004). Fall 2004 rich transcription (rt-04f) evaluation plan.

Chapter 6

Cognitive Approaches to Spoken Language Technology

Roger K. Moore

6.1 Introduction

As evidenced by the contributions of the other authors in this volume, spoken language technology (SLT) has made great strides over the past 20 or so years. The introduction of data-driven machine-learning approaches to building statistical models for automatic speech recognition (ASR), unit selection inventories for text-to-speech synthesis (TTS) or interaction strategies for spoken language dialogue systems (SLDS) has given rise to a steady year-on-year improvement in system capabilities. Such continued incremental progress has also been underpinned by a regime of public benchmark testing sponsored by national funding agencies, such as DARPA, coupled with an ongoing increase in available computer power.

The consequence of such developments has been that spoken language technology has successfully migrated from the research laboratories into a wide range of practical applications, and a viable market for voice-based products is beginning to mature. For example, since the 1990s most mobile phones have had an inbuilt facility for hands-free name dialling using ASR. Likewise, software for dictating documents onto your PC using ‘large vocabulary continuous speech recognition’ (LVCSR) has been available in most computer stores since 1997, and has recently made an appearance as a core component of Microsoft’s Vista[®] operating system. Also, interactive voice response (IVR) systems are now becoming commonplace for handling telephone bookings and enquiries.

6.1.1 Limitations of Current Technology

These successes reflect the practical viability of the core technological processes underpinning ASR, TTS and SLDS for a range of basic functionalities. However, many attractive applications remain out of reach due to the inherent limitations of

R.K. Moore (✉)

Department of Computer Science, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK
e-mail: r.k.moore@dcs.shef.ac.uk

the current approaches [1]. For example, contemporary spoken language technology can be extremely fragile in ‘real-world’ everyday situations; a noisy environment or a non-native speaker can lead to a large drop in ASR performance and a possible failure to handle the interactive dialogue. Likewise, a naïve or uncooperative user may misunderstand what the system is saying or is capable of doing, and hence fail to complete their transaction satisfactorily. Such performance shortfalls not only lead to disappointed customers, but also render speech-based interfaces non-competitive with respect to alternative methods of human–machine interaction (such as a telephone keypad or a computer mouse).

Current spoken language technology also tends to be highly optimized for particular users and application scenarios. This is done because it is critically important to maximize system performance in any given situation, but the downside is that the resultant SLT becomes expensive to port to new applications and languages since each system more or less has to be re-built from scratch. The prospect of a high-performance general-purpose SLT still eludes the R&D community.

As two of the prominent leaders in SLT R&D have said,

The industry has yet to bridge the gap between what people want and what it can deliver. Reducing the ASR error rate remains the greatest challenge. [2]

After sixty years of concentrated research and development in speech synthesis and text-to-speech (TTS), our gadgets, gizmos, executive toys and appliances still do not speak to us intelligently. [3]

Contemporary SLT is thus characterized by a lack of robustness to changing and alternative real-world situations together with an inability to apply high-level conceptual information to its low-level behaviours. As a consequence, although performance continues to improve year-on-year, the incremental gains appear to be diminishing and performance is asymptoting well short of human-level abilities [4].

6.1.2 What Is Missing?

Clearly these shortfalls in the capability of SLT systems are exactly in the areas in which human beings excel, and a large part of this difference would seem to be concerned with higher-level ‘cognitive’ involvement in the relevant processes. A human being has wide-ranging general-purpose knowledge, expertise, learning and problem-solving abilities which they are able to bring to bear in the course of spoken language interaction. The ability of a human to ‘understand’ spoken language plays a significant role in their ability to recognize speech even under very difficult conditions, and the natural expressiveness of human speech facilitates the communication of information currently out of the reach of contemporary SLT systems. Spoken language is also highly tailored to people’s communicative needs; its production is carefully crafted for consumption by an intended listener, and the listener can not only take advantage of this, but is also aware of the processes involved. Such behaviour is currently missing from existing SLT systems.

The consequence of this situation is that, far from being the ‘natural’ means of human–machine interaction that is often quoted in advertising literature, contemporary spoken language technology is rather restrictive in terms of permitted user behaviour. The lack of higher-level cognitive processes¹ means that conversational interaction with a current SLT-based system is necessarily quite constrained, and attempts to fake such functionality (for example, by providing a system with a high-quality human-like voice) can lead to users grossly overestimating the capabilities of an automated system with potentially disastrous implications for the success of the interaction.

Of course research into advanced forms of spoken language technology has attempted to incorporate high-level constraints, usually by integrating ASR and TTS with computational techniques derived from the field of natural language processing (NLP). However, the prevalence of uncertainty and errors in the speech input components has meant that such approaches have not met with any great success. Indeed the most successful solutions have resulted from ignoring traditional linguistic structures and extending the stochastic modelling paradigm to encompass high-level representations directly for both understanding [5] and dialogue [6].

There is relatively little research that draws directly on models of human cognition or that exploits the cognitive basis of human spoken language, yet such behaviour appears to be a key missing element in contemporary SLT systems. This chapter attempts to redress the balance by offering some modest insights into how this might be achieved. Section 6.2 provides background on a number of areas of research that are concerned with natural cognition and which may have a bearing on future approaches to spoken language technology. Section 6.3 then presents a short overview of artificial cognitive systems, with special attention being given to research into the situated and embodied grounding of language. Finally, Section 6.4 attempts to formulate a roadmap for future cognitively inspired approaches to spoken language technology.

6.2 Models of ‘Natural’ Cognition

The term ‘cognition’ has many different definitions, but the core concepts are concerned with

1. *learning and memory*: the acquisition and retention of abstract representations of conceptual knowledge
2. *reasoning*: the manipulation of such representations for generalizing, action planning and problem solving
3. *awareness*: including attention, perception and consciousness/self-monitoring

¹Current spoken language technology is literally ‘mindless’!

It is hypothesized that an individual's behaviour is conditioned on a set of basic needs and goals [7] and that attributes such as emotion arise in response to appraisal mechanisms that assesses any given situation with regard to an individual's needs and goals [8]. Perception and action are seen as core behaviours in cognition since they mediate the relationship between a living organism and the physical world, and speech and language have received much attention as the prime mechanism for abstract communicative interaction between one human being and another.

6.2.1 Cognitive Science

Studies of the cognitive aspects of human behaviour fall into the domains of cognitive science and cognitive psychology, and for more than 30 years these two fields have investigated the human mind through a combination of empirical observation and computational modelling. The experimental paradigms involved include the measurement of reaction times and sensory thresholds, the eliciting of similarity/difference judgments, and the implementation of experiments involving categorization, memory tasks, problem solving and decision making.

Since the late 1950s, the research agenda has been dominated by Broadbent's information processing model of cognition [9] in which mental processes operating within the brain are analogized metaphorically to software programs running on some form of computer. Issues concerning the relationship between symbolic/discrete and sub-symbolic/continuous representations of mental states have been to the fore, and both cognitive science and cognitive psychology have drawn inspiration from the field of artificial intelligence (AI), and especially from the connectionist, or artificial neural network, approaches to simulating cognitive phenomena within a computational framework.

Traditional models in cognitive psychology typically decompose cognition into a hierarchy of broad functional components [10], inspired by the neural structures induced through the results of studies into the psychological effects caused by mental illness and brain damage. Such models typically adopt a stimulus-response view of behaviour and capture the observed variability in such relationships in a modelling approach known as probabilistic functionalism or the 'Brunswik lens model' [11]. The Brunswikian stimulus-response view of behaviour has had a significant impact on studies of human cognitive abilities (e.g. [12]) as well as on aspects of spoken language (e.g. [13]). However, more recent research – particularly the discovery of 'mirror neurons' [14] – reveals a more complex picture in which cognitive behaviour is modelled in terms of multiple hierarchies of control feedback processes [15] that involve significant sensorimotor overlap linking perceptual and motor behaviours [16, 17] and which are capable of predicting and anticipating future events using memory-based emulation mechanisms learnt through imitative behaviour [18, 19].

6.2.2 Hierarchical Control

The notion that cognitive behaviour is essentially hierarchical in nature is based on the observation that evolutionary development tends to add rather than replace functionality [10] in an organism. This means that higher-order processing is overlaid on lower-order processing, and hence behaviour associated with consciousness—or ‘meta-cognition’ – is taken to be operating at the highest levels and behaviour associated with basic emotions at the lower levels [20].

Such a perspective is in general agreement with the principles of perceptual control theory [15, 21]. Powers proposed that much of the variability in human behaviour was not simply stochastic (as viewed within the traditional Brunswikian framework) but could be explained using a hierarchy of closed-loop negative feedback systems. Perceptual control theory claims that the behaviour of a living organism is actively directed towards maintaining desired perceptual consequences of its actions, and this approach has the rather radical outcome that perception is viewed not as a process for a detailed analysis of the world in order to figure out what is going on, but as a process for checking that the world is as an organism would want. If it is not as desired, then action can be taken to correct the imbalance.

The structure of a simple single-layered perceptual control system is shown in Fig. 6.1. An organism’s intention/need defines a reference signal that is realized through the actuators of its motor system as an action. The effects of the action are sensed, and the perceptual interpretation of the effects is compared with the original intention. If the perceived consequences do not match the original intention, further action is automatically triggered to correct the difference. The key outcome of such a negative feedback process is that behaviour can be constantly modified in the face of varying levels and types of disturbance, thereby providing an explanation to the conditional dependencies underlying what would otherwise be observed as random stimulus–response variation.

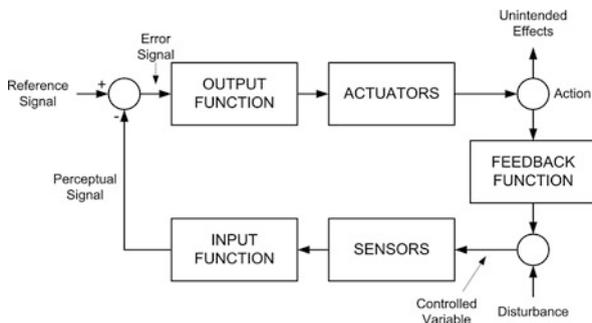


Fig. 6.1 Architecture of a single-layered perceptual control system

6.2.3 Emulation Mechanisms

Control feedback systems provide a powerful framework for modelling complex cognitive behaviour. However, studies of low-level neural control of muscle movements reveal that the external loop delays are too great to facilitate fine motor control over fast, goal-directed movements [22]. This has led to the introduction of theories that invoke internal models – emulators – that are capable of simulating proprioceptive and kinaesthetic feedback in response to efferent copies of the relevant motor commands. Such a ‘pseudo-closed-loop control system’ receives feedback, not from the target system, but from the output of the emulator [18] – see Fig. 6.2.

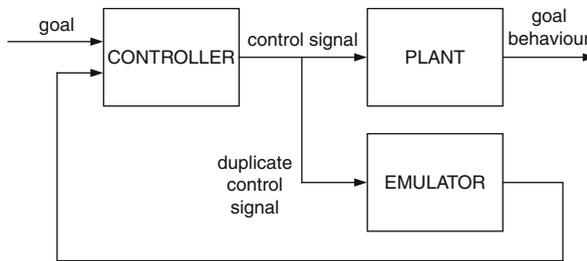


Fig. 6.2 Architecture of a pseudo-closed-loop control system [18]

What is interesting about such theories is that the notion of an internal forward model not only solves the feedback timing issue, but if the motor commands going to the target system were to be inhibited, then it would also provide a mechanism for motor imagery. Indeed it has been suggested that such emulators could provide a general means for both covert simulation and overt imitation of others’ behaviour [23] – both of which are implicated in the acquisition and development of spoken language [24–26].

A key feature of the forward models implied by such emulation mechanisms is that they provide a means for predicting future outcomes, and the general notion of perceptual prediction has been cited as the prime function of the neocortex and the basis for intelligent behaviour [19, 27]. Prompted by the observation that the neocortex is surprisingly uniform in structure [28], Hawkins’ has proposed an architecture known as ‘hierarchical temporal memory’ [29] that reflects the six-layer columnar organization of the cortex, and which performs prediction based on information stored in memory – see Fig. 6.3.

6.2.4 Mirror Neurons

An area of cognitive science that is receiving an increasing amount of attention and interest – mainly as a consequence of the emergence of neuroimaging as a powerful investigative tool for unravelling the patterns of activation in living

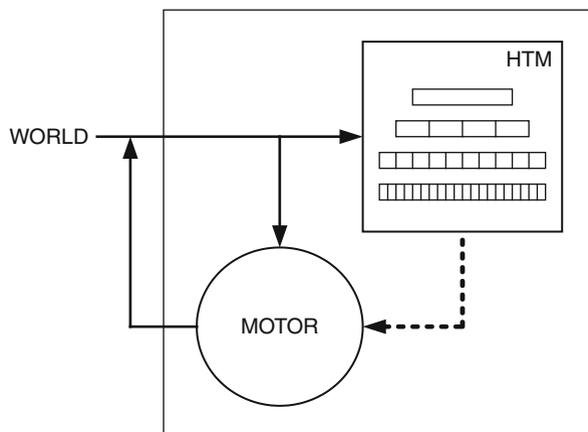


Fig. 6.3 Illustration of the generation of motor behaviour using ‘hierarchical temporal memory’ (after [29])

brains – is a growing appreciation of the intimate relationship between sensor and motor behaviour for action understanding and control. Cognitive psychologists had already described key sensorimotor interactions such as the ‘chameleon effect’ [30] in which people have a tendency to unconsciously mimic the behaviour of others or in which neonates show spontaneous imitation of facial gestures [31]. It had also been found that subjects were able to produce faster finger movements in response to seeing a video of that same finger movement [32] and faster pronunciation of a syllable in response to hearing that syllable or to seeing the same syllable pronounced than to hearing a tone or seeing a different syllable [33]. However in the mid-1990s, cognitive neuroscientists uncovered a more general mechanism in which imitative motor activation occurs in pre-motor cortex even when overt behaviour does not result – mirror neurons [14, 34].

The discovery of mirror neurons in the F5 pre-motor cortex area of a macaque monkey’s brain occurred during an experiment that involved measuring neural activation resulting from grasping movements. During the course of the study, it was found that such activation not only occurred when the monkey grasped an object such as a raisin, but also occurred when the monkey observed such a grasping action by the human experimenter. The activation did not occur when the experimenter grasped the raisin with a tool. The implication is that perceiving an action activates the neural circuitry involved in preparing to perform that action.

Mirror neurons have given new impetus to cognitive models of interactive behaviour. It is now accepted that not only is there sensory involvement in motor behaviour that provides feedback processes for controlling movement (as described above), but there is also motor involvement in sensory behaviour that facilitates the understanding of perceived actions [35]. In other words, it is thought that an organism possessing such mirror mechanisms is able to interpret the actions, intentions and emotions of another organism by referencing its own capabilities [17, 36].

Mirror mechanisms have now been posited as the neural basis for a wide range of cognitive functions such as imitation, learning by imitation and empathy [16]. They have also been linked to the general principle of ‘theory of mind’ which posits an ability of an organism to view the world from the perspective of another – a faculty that seems to be lacking in autistic individuals [37, 38]. Also, audiovisual mirror neurons have been shown to code actions independently of whether they are performed, heard or seen [39].

Of particular interest is that mirror mechanisms have shown to be involved in speech and language processing. For example, a degree of overlap in activation is elicited when a subject reads action words or performs the corresponding actions [40]. Similarly, it has been found that the speech production mechanism is activated whilst listening to speech [23]. Indeed the F5 area of pre-motor cortex in monkeys is thought to be the homolog of Broca’s area in humans [41], and hence mirror neurons have been implicated in the evolution of social cooperation [42], gestural communication, speech and language [43–45].

6.3 Artificial Cognitive Systems

In parallel with research into ‘natural’ cognition, scientists in the field of artificial intelligence (AI) have been investigating computational approaches to developing ‘artificial’ cognitive systems – computer-based systems that perform some practical function and that exhibit some degree of autonomous behaviour. Traditional approaches have been based on a ‘sense–plan–act’ cycle of behaviour [46], supported by symbolic cognitive processing architectures such as Soar [47], ACT-R [48] and BDI [49, 50]. Such architectures typically utilize procedural and declarative memory in which production rules fire in response to specific conditions thereby initiating a chain of events that can solve a particular problem by searching the corresponding problem state space.

Cognitive architectures such as Soar, ACT-R and BDI (beliefs, desires and intentions) are goal oriented and are assumed to require large amounts of knowledge in order to be able to plan an intelligent behaviour that satisfies the goals. Processing is essentially symbolic, and abstraction is seen as a key to generalizing to novel solutions. Learning takes place through experience. However, whilst such approaches – often referred to as GOF AI (‘good old fashioned AI’) – have been reasonably successful in making quantitative predictions about a range of psychological phenomena, recent years have seen a shift away from the execution of logical constructs and the manipulation of rules towards data-driven statistical approaches [51] and towards situated and embodied agents incorporating real-world interactions [48].

6.3.1 Embodied Cognition

One of the most significant influences on the development of ‘new AI’ is the work of [52–54]. Brooks argued that the high-level sense–model–plan–act model

of cognitive behaviour was insufficient to facilitate the type of dynamic real-world interaction required by physical devices such as robots. Drawing inspiration from the evolution of natural systems, Brooks proposed a layered architecture employing tight feedback loops between sensing and action in which each layer adds higher levels of competence that augment, modulate and/or subsume the lower levels. Not only has Brooks' approach been applied successfully to practical systems, but the results compare favourably with research in the neurosciences on the architecture of the vertebrate nervous system [55].

One of the key problems that Brooks set out to solve was how to 'ground' the behaviour of an artificial cognitive system in the physical world. In other words, how is it possible to connect the meanings of internal abstract representations to external objects and actions? This question is particularly relevant to language, spoken language interaction and the acquisition of speech [56].

6.3.2 Grounding Language

The most comprehensive attempt to ground spoken language in a situated context is the work of [57] in which they investigated the dynamic acquisition of a lexicon through cross-model correlation using a robot called Toco. Roy and Pentland's approach was based on a model of cross-channel early lexical learning (CELL) that used information theoretic measures to associate visual and auditory information – see Fig. 6.4. The result was that Toco was able to perform word unsupervised discovery and visual categorization through spoken language interaction.

Roy [58] has subsequently generalized his approach towards a general framework for grounding language in which 'schemas act as structured beliefs that are grounded in an agent's physical environment through a causal-predictive cycle of action and perception'.

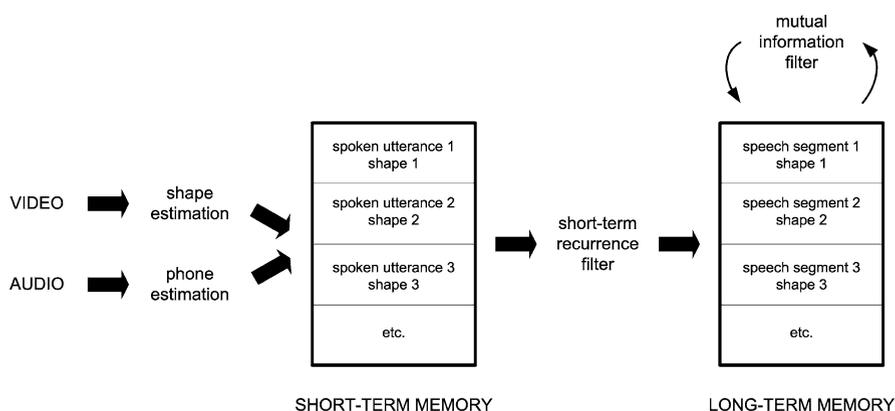


Fig. 6.4 Overview of the CELL model of cross-channel early lexical learning [57]

6.4 Roadmap for the Future

All of the foregoing supports the view that, in order to make progress in developing spoken language technology with the advanced capabilities that are normally associated with human spoken language interaction, it is necessary to pursue a radically new approach in which SLT is viewed as an intimate part of an artificial cognitive system and not simply as a peripheral interface technology. Contemporary R&D tends to treat ASR, TTS and SLD as independent components of an SLT system, and this has the consequence that the wider communicative function of speech is very difficult (if not impossible) to model because any systematic behaviour that results from speaker–listener interaction is observed (and is thus obliged to be modelled) as random variation.

6.4.1 The Way Forward?

What is required is a more integrative approach in which spoken language interaction is modelled as a cognitive behaviour that is conditioned on communicative context involving significant sensorimotor overlap. For example, the recently proposed PRESENCE – ‘PREdictive SENsorimotor Control and Emulation’ – theory of spoken language processing invokes a computational framework based on interlocked control feedback processes operating between speakers and listeners in which a talker has in mind the *needs* of the listener(s) and a listener has in mind the *intentions* of the talker(s) – see Fig. 6.5. PRESENCE draws inspiration from a number of different fields, including cognitive neuroscience, and is an attempt to accommodate within a single unified framework hitherto disparate aspects of spoken language behaviour such as speaking effort (hyper vs. hypo), listening effort (attention), empathy, alignment, imitation, emotion and individuality.

As illustrated in Fig. 6.5, the PRESENCE architecture is organized into four layers. The top layer is the primary route for motor behaviour such as speaking. An organism’s needs [7] – N_s – modulated by motivation conditions a communicative intention – I_s – that would satisfy those needs (determined by a process of search – as indicated by the diagonal arrow running through the module), which then drives both motor actions – M_s – and an emulation of possible motor actions – $E_s(M_s)$ – on the second layer. Sensory input feeds back into this second layer, thereby facilitating a check as to whether the desired intention has been met. If there is a mismatch between intended behaviour and the perceived outcome, then the resulting error signal will cause the system to alter its behaviour so as to reduce the mismatch.

The third layer of the model represents a feedback path on the behaviour of ‘self’ based on emulating the effect of self’s behaviour on ‘other’. In other words, $E_o(I_s)$ represents the emulation by ‘other’ of the intentions of ‘self’, and $E_s(E_o(I_s))$ represents the emulation of that function by ‘self’. A similar arrangement applies to $E_s(E_o(M_s))$. It is this layer that captures the empathetic relationship between speaker and listener in conditioning the behaviour of a speaker.

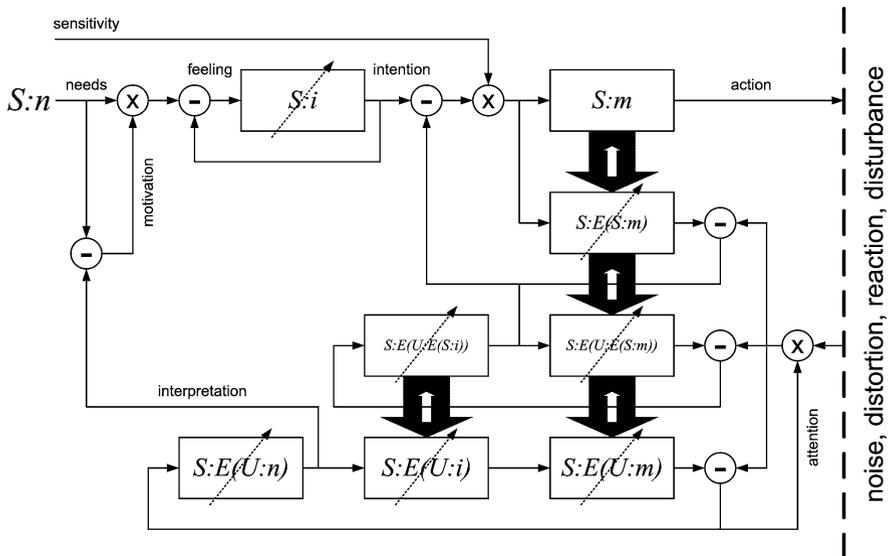


Fig. 6.5 Architecture of the PRESENCE model (where S: ‘self’, O: ‘other’, N: needs, I: intentions, M: motor activity, E(): emulation)

The fourth layer in the model represents self’s means for interpreting the needs, intentions and behaviour of others through a process of emulating others’ needs, intentions and behaviour based on others’ emulation of self’s own needs, intentions and behaviour. Like the second and third layers, this layer exploits the information embedded in the previous layers, and this process of parameter sharing or learning is indicated by the large block arrows. The small block arrows indicate a flow of information in the opposite direction.

The implications of cognitive-inspired approaches like PRESENCE for future generations of spoken language technology are that speech-based human–machine interaction would be founded on significant sensorimotor integration, overlap and control, providing an intimate link between low-level physical behaviours and high-level conceptual representations. In particular, it is possible to envisage

- new approaches to automatic speech generation that select their characteristics appropriate to the needs of the listener(s), monitor the effect of their own output and modify their behaviour according to their internal model of the listener.
- new approaches to automatic speech recognition that extend the generative model of the talker to incorporate the features listed above, and which adapt their generative model to the behaviour of the talker based on knowledge of their own vocal and linguistic capabilities.
- new approaches to spoken language dialog that model interaction as multiple synchronized/phase-locked behaviours rather than as simple turn-taking.

6.4.2 A New Scientific Discipline: Cognitive Informatics

Finally, it is clear from the topics covered in this chapter that future progress in understanding and modelling spoken language behaviour, and in implementing and exploiting a viable spoken language technology, depends on significant interdisciplinary exchanges between a wide variety of research fields. Spoken language is the most sophisticated behaviour of the most complex organism in the known universe, and this places spoken language technology at the heart of artificial cognitive systems and therefore firmly within the newly emerging transdisciplinary field of ‘cognitive informatics’ [59–61] – see Fig. 6.6.

Cognitive informatics aims to forge links between a diverse range of disciplines spanning the natural and life sciences, informatics and computer science, and it is founded on the conviction that many fundamental questions of human knowledge share a common basis – an understanding of the mechanisms of natural intelligence and the cognitive processes of the brain. The appearance of CI presents a unique opportunity to reverse the fragmentation and dispersion that has occurred in spoken language processing R&D [62]. Indeed, it can be argued that spoken language processing (by mind or machine) should constitute a core ‘grand challenge’ in

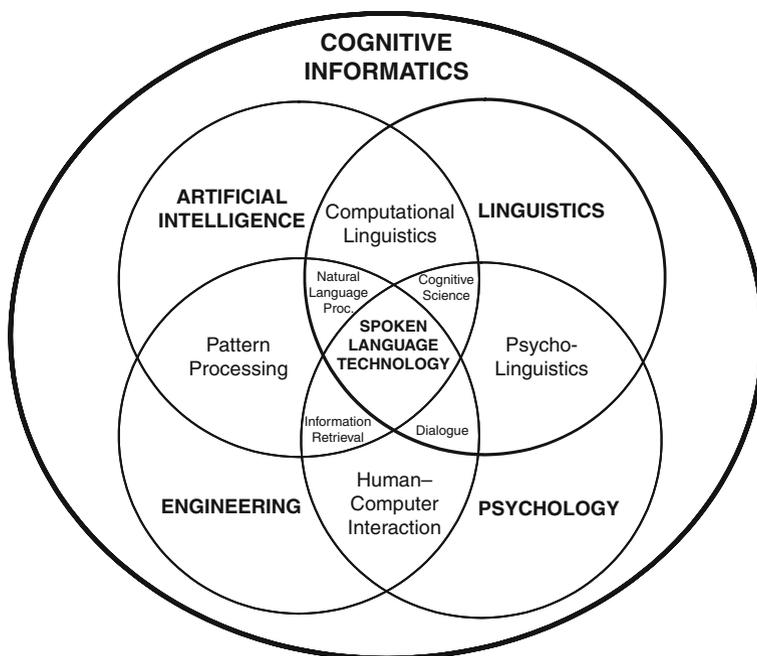


Fig. 6.6 Spoken language technology sits at the intersection of many disciplines, but it is at the heart of ‘cognitive informatics’

cognitive informatics [63], and this could pave the way for not only a deeper understanding of the mechanisms underpinning spoken language behaviour, but also lead to a step change in performance of spoken language technology systems.

References

1. Moore, R. K. (2005). Research challenges in the automation of spoken language interaction. In: Proc. COST278 and ISCA Tutorial and Research Workshop on Applied Spoken Language Interaction in Distributed Environments (ASIDE 2005): Aalborg University, Denmark, 10–11.
2. Huang, X. D. (2002). Making speech mainstream. Microsoft Speech Technologies Group.
3. Henton, C. (2002). Fiction and reality of TTS, *Speech Technology Magazine* 7(1).
4. Moore, R. K. (2003). A comparison of the data requirements of automatic speech recognition systems and human listeners. In: Proc. EUROSPEECH'03, Geneva, Switzerland, September 1–4, 2582–2584.
5. Gorin, A., Riccardi, G., Wright, J. (1997). How may I help you? *Speech Commun.*, 23, 113–127.
6. Young, S. J. (2006). Using POMDPs for dialog management. In: Proc. IEEE/ACL Workshop on Spoken Language Technology, Aruba Marriott, Palm Beach, Aruba, December 10–13, 8–13.
7. Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.*, 50, 370–396.
8. Scherer, K. R., Schorr, A., Johnstone, T. (2001). *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, New York and Oxford.
9. Broadbent, D. E. (1958). *Perception and Communication*. Pergamon Press, London.
10. Toates, F. (2006). A model of the hierarchy of behaviour, cognition and consciousness. *Consciousness Cogn.*, 15, 75–118.
11. Brunswik, E. (1952). The conceptual framework of psychology. *International Encyclopaedia of Unified Science*, vol. 1, University of Chicago Press, Chicago.
12. Figueredo, A. J., Hammond, K. R., McKierman, E. C. (2006). A Brunswikian evolutionary developmental theory of preparedness and plasticity. *Intelligence*, 34, 211–227.
13. Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Commun.*, 40, 227–256.
14. Rizzolatti, G., Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.*, 27, 169–192.
15. Powers, W. T. (1973). *Behaviour: The Control of Perception*. Aldine, Hawthorne, NY.
16. Wilson, M., Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychol. Bull.*, 131, 460–473.
17. Becchio, C., Adenzato, M., Bara, B. G. (2006). How the brain understands intention: Different neural circuits identify the componential features of motor and prior intentions. *Consciousness Cogn.*, 15, 64–74.
18. Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behav. Brain Sci.*, 27, 377–442.
19. Hawkins, J. (2004). *On Intelligence*. Times Books, New York, NY.
20. Lexandrov, Y. I., Sams, M. E. (2005). Emotion and consciousness: End of a continuum. *Cogn. Brain Res.*, 25, 387–405.
21. Taylor, M. M. (1992). Strategies for speech recognition and understanding using layered protocols. *Speech Recognition and Understanding – Recent Advances*. NATO ASI Series F75, Springer-Verlag, Berlin, Heidelberg.
22. Gerdes, V. G. J., Happee, R. (1994). The use of an internal representation in fast goal-directed movements: A modeling approach. *Biol. Cybernet.*, 70, 513–524.
23. Wilson, S. M., Saygin, A. P., Sereno, M. I., Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.*, 7, 701–702.

24. Gopnik, A., Meltzoff, A. N., Kuhl, P. K. (2001). *The Scientist in the Crib*. Perennial, New York, NY.
25. Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nat. Rev.: Neurosci.*, 5, 831–843.
26. Cowley, S. J. (2004). Simulating others: The basis of human cognition. *Lang. Sci.*, 26, 273–299.
27. Weber, C., Wermter, S., Elshaw, M. (2006). A hybrid generative and predictive model of the motor cortex. *Neural Netw.*, 19, 339–353.
28. Mountcastle, V. B. (1978). An organizing principle for cerebral function: The unit model and the distributed system. In: Edelman, G. M., Mountcastle, V. B. (eds) *The Mindful Brain*, MIT Press, Cambridge, MA.
29. Hawkins, J., George, D. (2006). *Hierarchical Temporal Memory: Concepts, Theory, and Terminology*. Numenta Inc., Redwood City, CA.
30. Chartrand, T. L., Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Social Psychol.*, 76, 893–910.
31. Meltzoff, M., Moore, K. (1997). Explaining facial imitation: A theoretical model. *Early Dev. Parenting*, 6, 179–192.
32. Brass, M., Bekkering, H., Wohlschlagel, A., Prinz, W. (2000). Compatibility between observed and executed finger movements: Comparing symbolic, spatial, and imitative cues. *Brain Cogn.*, 44, 124–143.
33. Kerzel, D., Bekkering, H. (2000). Motor activation from visible speech: Evidence from stimulus response compatibility. *J. Exp. Psychol. [Hum. Percept.]*, 26, 634–647.
34. Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Res.*, 3, 131–141.
35. Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., Rizzolatti, G. (2005). Grasping the intentions of others with one’s own mirror system. *PLoS Biol.*, 3, 529–535.
36. Gallese, V., Keysers, C., Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends Cogn. Sci.*, 8(9), 396–403.
37. Baron-Cohen, S., Leslie, A. M., Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition*, 21, 37–46.
38. Baron-Cohen, S. (1997). *Mindblindness: Essay on Autism and the Theory of Mind*. MIT Press, Cambridge, MA.
39. Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, 297, 846–848.
40. Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nat. Neurosci. Rev.*, 6, 576–582.
41. Rizzolatti, G., Arbib, M. A. (1998). Language within our grasp. *Trends Neurosci.*, 21, 188–194.
42. Pacherie, E., Dokic, J. (2006). From mirror neurons to joint actions. *Cogn. Syst. Res.*, 7, 101–112.
43. Studdart-Kennedy, M. (2002). Mirror neurons, vocal imitation, and the evolution of articulate speech. In: *Mirror Neurons and the Evolution of Brain and Language*. M.I. Stamenov, V. Gallese (Eds.), Philadelphia: Benjamins, 207–227.
44. Arbib, M. A. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguists. *Behav. Brain Sci.*, 28, 105–167.
45. Aboitiz, F., Garcia, R. R., Bosman, C., Brunetti, E. (2006). Cortical memory mechanisms and language origins. *Brain Lang.*, 40–56.
46. Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.
47. Rosenbloom, P. S., Laird, J. E., Newell, A. (1993). *The SOAR Papers: Research on Integrated Intelligence*. MIT Press, Cambridge, MA.
48. Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychol.*, 51(4), 355–365.

49. Bratman, M. E. (1987). *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge, MA.
50. Rao, A., Georgoff, M. (1995). *BDI agents: From theory to practice*. Technical Report TR-56. Australian Artificial Intelligence Institute, Melbourne.
51. Winograd, T. (2006). Shifting viewpoints: Artificial intelligence and human-computer interaction. *Artif. Intell.*, 170, 1256–1258.
52. Brooks, R. A. (1991). Intelligence without representation. *Artif. Intell.*, 47, 139–159.
53. Brooks, R. A. (1991). Intelligence without reason. In: *Proc. 12th Int. Joint Conf. on Artificial Intelligence*, Sydney, Australia, 569–595.
54. Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE J. Rob. Autom.* 2, 4–23.
55. Prescott, T. J., Redgrave, P., Gurney, K. (1999). Layered control architectures in robots and vertebrates. *Adaptive Behav.*, 7, 99–127.
56. Roy, D., Reiter E. (2005). Connecting language to the world. *Artif. Intell.*, 167, 1–12.
57. Roy, D. K., Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cogn. Sci.*, 26, 113–146.
58. Roy, D. (2005). Semiotic schemas: A framework for grounding language in action and perception. *Artif. Intell.*, 167, 170–205.
59. Wang, Y. (2003). Cognitive informatics: A new transdisciplinary research field. *Brain Mind*, 4, 115–127.
60. Wang, Y. (2003). On cognitive informatics. *Brain Mind*, 4, 151–167.
61. Moore, R. K. (2005). Cognitive informatics: The future of spoken language processing? In: *Proc. SPECOM – 10th Int. Conf. on Speech and Computer*, Patras, Greece, October 17–19.
62. Moore, R. K. (2007). Spoken language processing: Piecing together the puzzle. *J. Speech Commun.* 49:418–43.
63. Moore, R. K. (2005). Towards a unified theory of spoken language processing. In: *Proc. 4th IEEE Int. Conf. on Cognitive Informatics*, Irvine, CA, USA, 8–10 August, 167–172.

Chapter 7

Expressive Speech Processing and Prosody Engineering: An Illustrated Essay on the Fragmented Nature of Real Interactive Speech

Nick Campbell

7.1 Introduction

This chapter addresses the issue of expressive speech processing. It attempts to explain a mechanism for expressiveness in speech, and proposes a novel dimension of spoken language processing for speech technology applications, showing that although great progress has already been made, there is still much to be done before we can consider speech processing to be a truly mature technology.

There have been considerable and rapid advances made in the various component technologies over the past 10 years, and we now see functioning speech translation devices that are capable of mediating a conversation between people who do not even speak the same language. For fixed-domain applications such as travel or shopping assistance, these devices are capable of recognizing speech input in several languages, converting the speech to text, translating the text, and then converting the translated text into speech in a different output language. This successful integration of three separate speech-related technologies, recognition, translation, and synthesis, proves that each has independently reached a degree of maturity in itself, and that all can be used together to model spoken dialogue processes.

However, although the component technologies have been employed successfully within an integrated application, we cannot yet claim them to be fully integrated in a way that models all aspects of spoken interaction. Each has been developed independently of the other, and the implicit assumption behind each component technology is that there is some form of one-to-one mapping between text and speech; i.e., that speech can be rendered as text, text can be manipulated preserving the original content, and that new speech can be generated from existing text. Furthermore there is the underlying assumption that this mapping is sufficient for the processing of spoken language.

N. Campbell (✉)

Centre for Language and Communication Studies (CLCS), The University of Dublin,
College Green, Dublin 2, Ireland
e-mail: nick@tcd.ie

In the sections that follow we show that while the mapping may be adequate for the conversion of linguistic or propositional aspects of spoken interaction, it is not capable of processing a large part of the social or interpersonal information exchange that takes place in human speech communication, or of recognizing and generating the discourse control signals that speakers use in a conversation. We examine the role of prosody in spoken language interactions, not from its function as an indicator of syntactic and semantic relationships, but more from the point of view of its role as a social lubricant in mediating human spoken interactions.

Section 7.2 considers the role of prosody in speech communication from a theoretical standpoint, presenting a broader view of prosodic information exchange. Section 7.3 presents some acoustic evidence for the ideas put forward in Section 7.2, and finally, Section 7.4 suggests some technological applications that might arise from this broader view of spoken language interaction and its related speech processing.

7.2 Prosodic Information Exchange

The user of a current speech translation system can input a sentence, wait briefly while it is translated, and then hear it reproduced in a foreign language. His or her partner will then be able to reply similarly, producing an utterance in their own language, waiting briefly while it is translated, and then watch the original speaker's reaction while it is synthesized in that person's own language. The processing is in near real-time, so the delays are not long, but the interaction itself is thereby very strained. The partners have to wait for their turn to speak, and there are long silences in the conversation.

A naturally interactive dialogue is not like a tennis match, where there is only one ball that can only be in one half of the court at any given time. Rather it is like a volley of balls being thrown in several directions at once. The speaker does not usually wait silently while the listener parses and reacts to an utterance; there is a constant exchange of speech and gesture, resulting in a gradual process of mutual understanding wherein a true 'meeting of the minds' can take place.

7.2.1 Natural Interactive Speech

As part of the JST/CREST Expressive Speech Processing (ESP) project [1], we recorded a series of conversations between 10 people who were not initially familiar with each other, and who had little or no face-to-face contact, but who were paid to meet once a week to talk to each other over the telephone for 30 min each, over a period of 10 weeks. The content of the conversations was completely unconstrained and left up to the initiative of the participants.

The volunteer speakers were paired with each other as shown in Fig. 7.1 so that each conversed with a different combination of partners to maximize the different

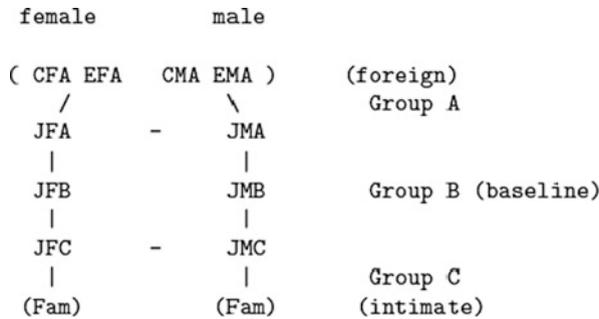


Fig. 7.1 Showing the form of interactions between participants in the ESP_C corpus. Here, the first letter of each three-letter participant identifier indicates the mother tongue (Japanese/Chinese/English) of the speaker, the second letter indicates the speaker’s sex (female or male), and the third letter is the group identifier, A, B, or C. (Fam) is short for ‘family’; indicating intimate conversations with relatives

types of expressiveness in the dialogues without placing the speakers under any requirement to self-monitor their speech or to produce different speaking styles ‘on-demand.’ The 10 speakers were all recorded in Osaka, Japan, and all conversations were in Japanese. Since the speakers were not familiar with each other initially, little use was made of the local dialect and conversations were largely carried out in the so-called ‘standard’ Japanese. No constraints on type of language use were imposed, since the goal of this data collection was to observe the types of speech and the variety of speaking styles that ‘normal’ people use in different everyday conversational situations.

Figure 7.2 shows the speech activity patterns of two Japanese speakers, one male (JMB) and one female (JFC) for the first 11 min of their sixth 30-min telephone conversation. We can see that even though it is usually quite clear who is the dominant speaker at any point in the conversation, neither speaker stays quiet for long, and that a gap of even 5 s in the speech could be considered as a long pause. In this example, the female speaker was older than the male and she tended to lead the conversation.

Table 7.1 gives details of speech activity time per speaker. It shows that for a 30-min conversation between two people, median speaking time is approximately 18 min per speaker. There is approximately 3 min when no one is speaking (10% of the total time) and 7 min (i.e., more than 20% of the conversation) when both speakers are speaking at once. Since time of non-overlapping speech is approximately 14 min per speaker, we can conclude that people overlap their speech, or talk simultaneously, one third of the time. These data were calculated from time-aligned transcriptions of 100 telephone conversations.

If we compare this ‘natural’ form of speech activity to that required for use of a speech translation system, we find that the waiting time imposed by the ‘ping-pong’ type of speech interaction assumed in that technology is excessively long.

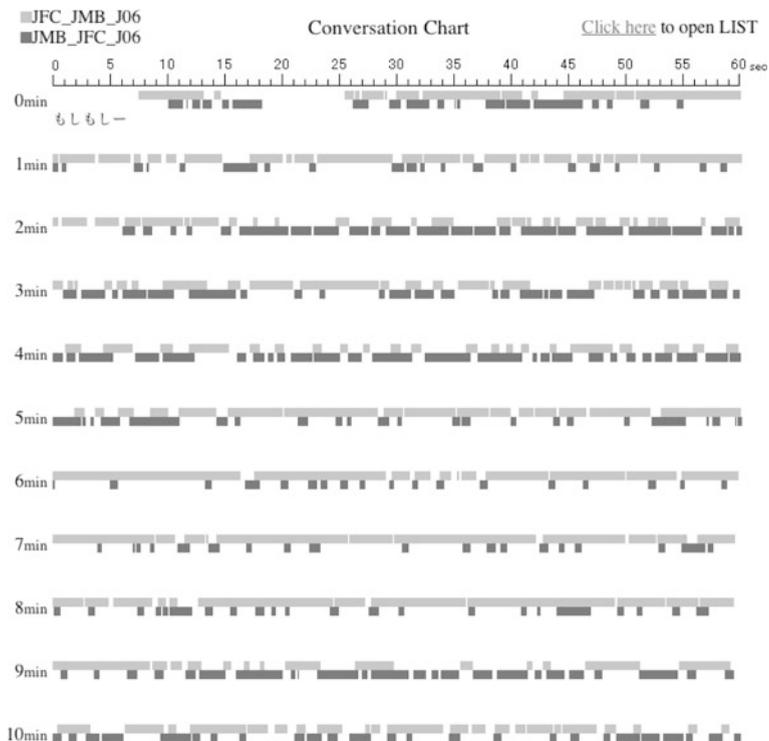


Fig. 7.2 Speech and silence plots for the first 11 min of conversation #6 between two Japanese speakers, JFC and JMB, showing fragmentation of the discourse and progressive but not absolute alternations of speaker dominance. Each line shows 1 min of speech, with speaker JFC's speech activity plotted above and that of speaker JMB plotted below. White space indicates lack of speech activity. The figure is taken from a screen capture of an interactive web page (see www.speech-data.jp)

7.2.2 Two-Way Interactive Speech

The careful and controlled speech of professionals, such as broadcasters, newsreaders, and announcers is typically much closer to written text in form, since they are (a) usually practiced and rehearsed and (b) remote from their listeners. The speech of two people in face-to-face or telephone-based interaction, on the other hand, is neither practiced nor remote. The interaction requires a constant to-and-fro of information exchange as the listener confirms, questions, and embellishes the speaker's propositional fragments. Being a very two-way interactive process, it also necessarily requires some form of discourse management control. Much of this is done through the use of non-verbal utterances and tone-of-voice.

It is common to speak of disfluencies in natural speech, and of fillers and hesitations as if they are performance errors, with the assumption that 'perfect' speech would be very similar in form to a written text, like that of a professional, well

Table 7.1 Showing quantiles of speech activity time per speaker. ‘Silence’ is when neither is speaking, ‘overlap’ when both are speaking at the same time. ‘Sil’ shows the time each speaker individually (A or B) was quiet. ‘Solo’ shows the total duration of non-overlapping speech per speaker, and ‘talk’ the total overall speech time including overlaps. ‘Duration’ shows timing statistics for the entire conversation (assumed to be 30 min by default). All times are shown in minutes. Data are calculated from the time-aligned transcriptions of 100 30-min conversations

	Min	25%	Median	75%	Max
Silence	0.99	2.08	2.85	3.81	7.03
silA	6.73	10.68	14.02	16.91	22.46
silB	5.72	13.09	14.68	17.68	21.58
soloA	4.14	9.51	11.66	14.68	18.17
soloB	4.55	8.39	10.64	13.32	18.90
overlap	2.66	5.53	7.01	9.04	12.80
talkA	10.80	16.04	18.75	22.44	28.52
talkB	12.20	15.66	17.93	20.15	27.15
Duration	28.57	32.00	32.93	33.96	37.98

formed, clear, concise, and precise. However, we might also consider an alternative point of view, as proposed here, that this so-called ‘ill-formed’ speech is in fact the product of natural evolution of the spoken language so that it can transmit interpersonal, affective, and discourse-related information at the same time as, and in parallel to, the transmission of propositional content.

To account for this supposedly ‘broken’ form of natural conversational speech we have suggested a structure of ‘wrappers’ and ‘fillers’ wherein the propositional content, here called a ‘filler’ (the term is used here as if describing the contents of a box of chocolates, with each wrapped distinctively and all having different fillers) is ‘wrapped’ in affect-bearing prefix and suffix fragments.

In this hypothesis, the speaker forms a complex utterance through a sequence of smaller and simple fragments. These are not presented in a concise linear sequence as they might be in writing, but are interspersed with ‘wrappers’ that indicate how they should be perceived. The speaker typically has a large repertoire of semantically ‘empty’ but affectively marked words or phrases (such as fillers in the conventional sense) that can be added at the beginning or end of an utterance fragment to embellish it or to show affect-related information. Some examples for Japanese, with their counts, are given in Table 7.2.

For an example in English, we might consider the speech of a typical Londoner who might produce the following sequence:

... erm, anyway, you know what I mean, ..., **it’s like**, er, sort of **a stream of** ... er ... **words** ... **and**, you know, **phrases** ... **all strung together**, you know what I mean, ...

where the words in bold font form the content (or the filling of the utterance) and the other words form the wrapping or decoration around the content.

This (mis-)usage of the term filler is in (deliberate) contrast to its usual interpretation as something which ‘occupies a gap’ or a supposed empty space in a discourse. On the contrary, we suggest here that these are not gaps in the discourse

Table 7.2 Counts of the hundred most common utterances of Japanese, as found in the ESP corpus of natural conversations. All function to display affect. While direct glosses are not provided here, most would be transcribed as variants of *ummm*, *aah*, *uhuh*, *yeah-yeah-yeah*, etc., @S, @E, and @K are symbols used to indicate breath-related sounds such as a hiss or a sharp intake of breath to show surprise or displeasure. Dashes indicate vowel lengthening

10073	うん	467	ズー	228	ううん	134	へー
9692	@S	455	スー	227	えっ	134	はいはいはいはい
8607	はい	450	んー	226	へー	134	そう、です
4216	laugh	446	うー	226	ハハハ	133	@E
3487	うーん	396	ねー	225	うんー	133	あ、そう、なん、ですか
2906	ええ	395	あ、あー	200	そうですね	130	そう、なん、ですか
1702	はい	393	はいはいはい	199	ほー	129	はー
1573	うーん	387	あ、はい	193	ハ	129	い
1348	ズー	372	ねえ	192	その	127	ほー
1139	ふん	369	ふー	190	え、えー	125	ハハハハハ
1098	あのー	369	だから	188	あ、あー	119	はい、はい
1084	あっ	368	あーん	187	ね	119	は、ー
981	はあい	366	ああ	180	ん、はい	114	ハハ
942	あの	345	あの、ー	180	あの、ー	113	は
941	ふーん	337	なんか	173	んん	113	で、ー
910	そう	335	え	172	アハハハ	113	て
749	えー	311	でも	168	はい、ー	112	は、あー
714	あー	305	スー	164	う、うーん	110	フフフ
701	あ	274	うん、うん、うん	161	は、ー	110	そのー
630	あー	266	ハハハハ	160	@K	110	もう
613	あ、はい	266	て、ー	159	そう、です、ねー	109	ふー
592	うん、うん	266	え、ー	151	あ、ー	108	は、あ、ー
555	あー	258	で	143	だから、ー	106	そうですね、え
500	んー	248	う	139	アハハハハ	105	んーん
469	ん	242	へー	137	そう、そう、そう	104	いや

but *essential* markers for a parallel tier of information. By their very frequency, these non-propositional and often non-verbal speech sounds provide not just time for processing the linguistic content of the spoken utterance but also a regular base for the comparison of fluctuations in voice quality and speaking style that indicate how the content is to be understood and how it relates to the discourse.

These fragments allow the speaker to express information related to mood and emotion, to interpersonal stance, and to discourse management. By being effectively transparent (i.e., they would not be transcribed when recording the speech in the minutes of a meeting, for example) they do not interfere with the transmission of linguistic or propositional content, but by being simple, frequent, and often repeated sounds, they allow easy comparison, like with like throughout the utterance so that the listener can sense the speaker's intentions through subtle variation in their usage and prosody.

7.2.3 *Speech Fragments*

From an analysis of 150,000 transcribed conversational utterances in a separate section of the JST-CREST ESP corpus, recorded from one female speaker over a period of about 4 years, we found that almost 50% of the utterances are ‘non-lexical’; i.e., they could not be adequately understood from their text alone. (Table 7.2 shows some examples, Table 7.3 provides detailed figures). Very few of these utterance types can be found as an entry in a standard language dictionary, yet it was confirmed that the intended meanings of many of these non-verbal utterances (or conversational ‘grunts’) can be perceived consistently by listeners even when presented in isolation without any discourse context information. In many cases, the intentions underlying the utterances can be appropriately and consistently paraphrased, even by listeners of completely different cultural and linguistic backgrounds [2].

Table 7.3 Counts of non-verbal utterances from the transcriptions for conversations produced by one female speaker from the ESP corpus. Utterances labeled ‘non-lexical’ consist mainly of sound sequences and combinations not typically found in the dictionary, but may also include common words such as ‘yeah,’ ‘oh,’ ‘uhuh,’ etc.

Total number of utterances transcribed	148,772
Number of unique ‘lexical’ utterances	75,242
Number of ‘non-lexical’ utterances	73,480
Number of ‘non-lexical’ utterance types	4492
Percent of ‘non-lexical’ utterances	49.4%

In the following section, we see how these affect-bearing fragments, which are effectively transparent in the discourse and do not appear at all intrusive to an observer, can carry significant interpersonal information through tone-of-voice and other such prosodic variation.

7.3 Acoustic Correlates of Discourse-Related Non-verbal Speech Sounds

In previous work we have found from an analysis of the speech of a single female speaker that her voice quality changes significantly according to type of interlocutor, familiarity with the interlocutor, pragmatic force of the utterance, etc. In this chapter we add further evidence to show that this is a general phenomenon, using speech data taken from a series of recorded telephone conversations between a small number of Japanese men and women over a period of several months.

7.3.1 Voice Quality, Prosody, and Affect

The earlier study, based on analysis of the ESP corpus of conversational speech, showed that voice quality, or laryngeal phonation style, varied consistently and in much the same way as (but independent of) fundamental frequency, to signal paralinguistic information [3]. We showed that the factors ‘interlocutor,’ ‘politeness,’ and ‘speech-act’ all had significant interactions with this variation.

The mode of laryngeal phonation can be measured from an estimate of the glottal speech waveform derivative (a result of inverse filtering of the speech using time-varying optimized formants to remove vocal tract influences) by calculating the ratio of the largest peak-to-peak amplitude and the largest amplitude of the cycle-to-cycle minimum derivative [4]. In its raw form it is weakly correlated with the fundamental period of the speech waveform ($r = -0.406$), but this can be greatly reduced by $NAQ = \log(AQ) + \log(F_0)$, yielding a Normalized Amplitude Quotient (henceforth NAQ) which has only a very small correlation of ($r = 0.182$).

We analyzed data from one female Japanese speaker, who wore a small head-mounted, studio-quality microphone and recorded her day-to-day spoken interactions onto a MiniDisk over a period of more than 2 years. The data comprise 13,604 utterances, being the subset of the speech for which we had satisfactory acoustic and perceptual labels. Here, an ‘utterance’ is loosely defined as the shortest section of speech having no audible break, and perhaps best corresponds to an ‘intonational phrase.’ These utterances vary in complexity from a simple single syllable up to a 35-syllable stretch of speech.

The factor ‘interlocutor’ was analyzed for NAQ and F_0 , grouped into the following classes: child ($n=139$), family ($n=3623$), friends ($n=9044$), others ($n=632$), and self ($n=116$). It is clear that F_0 and breathiness are being controlled independently for each class of interlocutor. Repeated t -tests confirm all but the child-directed ($n=139$) voice-quality differences to be highly significant.

Figure 7.3 (left part) shows median NAQ and F_0 for the five categories of interlocutors. The values are shown as z -scores, representing difference from the mean in SD units. NAQ is highest (i.e., the voice is breathiest) when addressing ‘others’ (talking politely), and second highest when talking to children (softly). Self-directed speech shows the lowest values for NAQ , and speech with family members exhibits a higher degree of breathiness (i.e., it is softer) than that with friends. F_0 is highest for child-directed speech and lowest for speech with family members (excluding children). Figure 7.3 (right part) shows the values for ‘family’ speech in more detail. It reveals some very interesting tendencies. Family members can be ordered according to breathiness as follows: daughter > father > nephew > mother = older sister > aunt > husband. Thus, it seems that the ordering reflects the degree of ‘care’ taken in the speech to each family member. In traditional Japanese families, the father is perhaps a slightly remote figure, but deserves respect. The mother (and older sister) comes next in ranking, and husband comes last – not indicating a lack of respect, but an almost total lack of need to display it in the speech. We can infer from the data here that this speaker also has a very close relationship with her aunt, a detail that was subsequently confirmed by her in person.

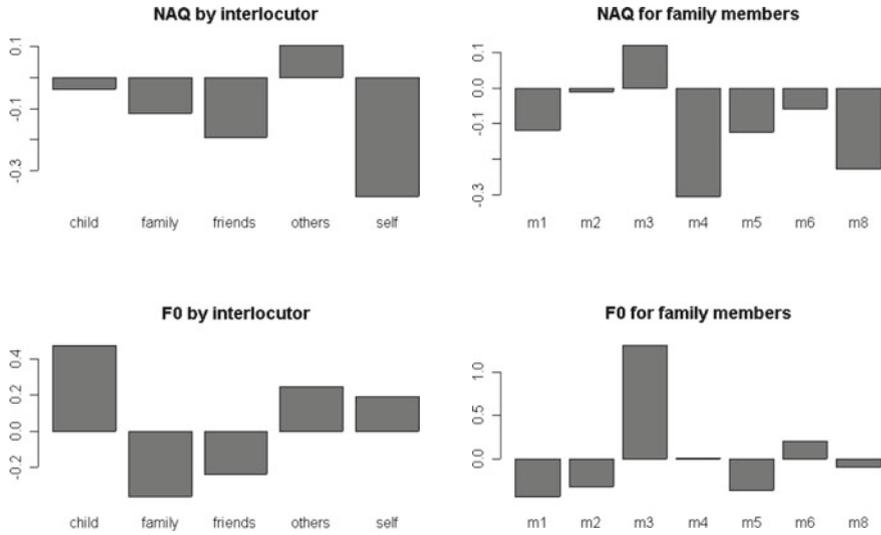


Fig. 7.3 Median values of NAQ and F_0 plotted for interlocutor (left) and for family members (right). m1: mother, m2: father, m3: daughter, m4: husband, m5: older sister, m6: sister's son, m8: aunt. Data are (z-score) scaled, so values are in SD units. 0 represents the mean of the distribution

7.3.2 Multi-speaker Variation in Prosody and Tone-of-Voice

To further validate this finding, we recently processed the data from the ESP_C corpus of telephone conversations between people who were strangers at first but then gradually became friends over the period of the recordings. These Japanese adults used head-mounted microphones and recorded their speech directly to DAT while they spoke to each other over the telephone from different locations with no face-to-face contact.

At the beginning of the recordings, they were all strangers to each other, but over the period of 10 weekly conversations they gradually became familiar to differing degrees. They spoke over the telephone to each other, to family members, and to foreign visitors to Japan who were capable of holding a simple conversation but not yet fluent in the language. In this way, we were able to control for 'ease of communication' without constraining the conversations in any artificial way. They were paid to talk to each other and, from the transcriptions of the dialogues, appeared to enjoy doing so.

Because the calculation of NAQ requires a degree of hand intervention for setting up specific initial speaker-related parameters, for this study, we opted to use a combination of several measures of prosodic information that could all be extracted automatically, without manual intervention, from the speech waveform. We extracted acoustic data from the recordings of both speakers in a series of 100 30-min conversations.

A combination of 14 different acoustic features was used in this experiment; specifically, the mean, maximum, minimum of power (rms, amplitude) and pitch (F_0), the position of the F_0 peak of each utterance, measured as a percentage distance from 0 (beginning) to 100 (end of utterance), the amount of voicing throughout the utterance, the values of the first and second harmonics, the third formant, and the spectral tilt (after Hanson 1994) [5], as well as a measure of speaking rate or normalized duration of the utterance. These measures were averaged across the whole of each utterance, giving only a general indication of prosodic settings for longer utterances but allowing a very precise comparison of the more frequent shorter utterances when comparing like with like throughout the progress of a discourse.

We performed a Principal Component Analysis (PCA) of these data to reduce the number of factors in the measure, and then plotted the first three principal components, which account for about half of the variance observed in the acoustic data, categorized by conversation number. In this way we can show how the prosodic settings vary with time. In the default case we would expect them to remain the same over time. For example, a person's voice pitch may change a little from day to day, according to health, smoking, and alcohol intake, as well as according to mood and emotion, but we would expect to see a steady average over a period of several weeks.

Table 7.4 gives details of the principal component analysis, showing how much of the variation was covered by each component. Table 7.5 shows how the individual acoustic measures were mapped by the components in the PCA reduction. We can see that approximately half of the variance is covered by the first three components alone, and that more than 80% is accounted for by the first seven.

From Table 7.5 we can see that the first principal component maps well onto F_0 mean and maximum, while the second maps onto $h1$ (power at the first harmonic) and $h1a3$ (the ratio of first harmonic to amplitude of the third formant) which is a measure of spectral tilt, related to breathiness and tension in the voice. The third component has a broader scope but appears related to degree of voicing and changes in signal amplitude.

Table 7.4 Results of the Principal Component Analysis importance of components

Importance of components	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.65	1.53	1.38	1.32	1.12	0.96	0.89	0.83
Proportion of variance	0.19	0.16	0.13	0.12	0.08	0.06	0.05	0.04
Cumulative proportion	0.19	0.36	0.49	0.62	0.71	0.78	0.83	0.88
	PC9	PC10	PC11	PC12	PC13	PC14		
Standard deviation	0.74	0.71	0.61	0.29	0.23	0.0004		
Proportion of variance	0.03	0.03	0.026	0.006	0.004	0.0001		
Cumulative proportion	0.92	0.96	0.98	0.99	1.00	1.00		

Table 7.5 Showing the precise relationship between each principal component and each prosodic factor derived automatically from the acoustic speech signal

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
fmean	-50	14	25	-20	11	-3	4	-7	8	3	14	27	71	0
fmax	-48	10	23	0	9	1	30	-26	18	18	36	-37	-46	0
Fmin	-29	11	17	-46	12	-26	-27	15	7	-14	-62	-9	-2	0
Fpct	-6	20	-13	-22	-40	46	-62	-25	11	18	12	-9	-2	0
Fvcd	-8	-23	-39	-29	-7	6	36	-61	-15	9	-28	29	-7	0
pmean	-36	-26	-31	29	-16	-7	-7	-7	-17	-28	-16	-59	30	0
Pmax	-43	-12	0	42	-5	2	-27	1	-17	-30	8	56	-34	0
Pmin	-20	-26	-37	7	-12	-34	-10	32	29	64	2	11	-1	0
ppct	-16	16	-15	-14	-51	30	45	46	23	-27	-6	8	-5	0
h1h2	-13	-30	32	4	6	53	8	25	-44	41	-26	-5	1	0
h1a3	9	-50	34	-12	-28	-9	-4	-9	20	-12	7	1	1	-67
h1	5	-57	11	-14	14	19	-7	-1	43	-21	8	0	0	60
a3	-8	0	-39	-1	63	40	-5	12	27	-10	0	-1	-1	-44
dn	5	18	22	54	-6	12	7	-27	49	14	-51	5	9	0

It is encouraging that these automatically derived measures match well to our intuitions mentioned above about the usefulness of measures of spectral tilt as a prosodic feature. At the interpersonal level of spoken interaction, tone-of-voice is perhaps more important than, e.g., pitch patterns, which form the core of traditional prosodic research and have a closer relation to syntactic and semantic structures within the linguistic component of the utterance.

Also of great interest is the finding shown in Fig. 7.4 that the first three components (at least) vary in a consistent way with progression of the conversations through the series. We can see a clear increase in values of each component, going from negative in the earlier conversations to positive in the later ones. This correlates well with the increase in familiarity between the participants and shows that their basic phonatory settings change. The discrepancy seen in the final conversation may well arise as a result of that conversation being recorded (as an afterthought) after a longer break, to make up for a missed conversation earlier in the series.

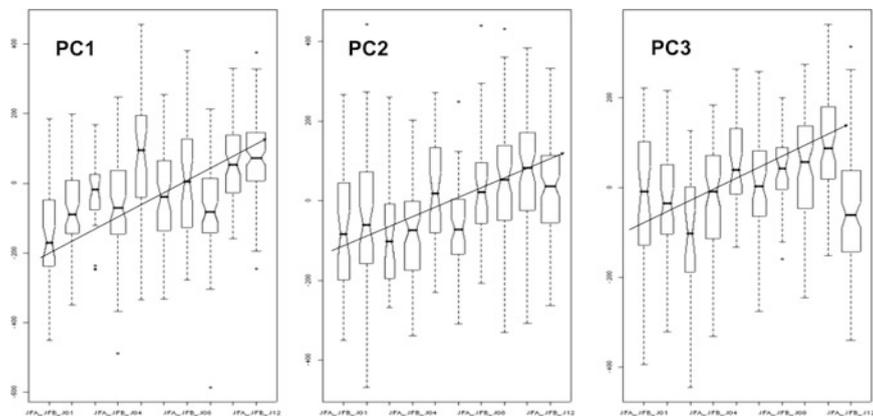


Fig. 7.4 The first three principal components plotted by conversation number for speaker JMC. We can see clear indication of an increasing trend that correlates with familiarity of the participants through the series of conversations

7.4 Technological Applications

At this point we consider how these findings can be made use of in speech technology applications. We can immediately consider two aspects of future development: one concerned with discourse flow, the other with affect sensing. The first allows people easier access to machine-mediated speech; the second allows machines access to aspects of interpersonal human-related information that may not be immediately discernible from the linguistic output of a speech recognizer.

7.4.1 Discourse Flow and Prosody Engineering

Currently, the users of a speech translation system have to wait in patient silence until their utterance and its subsequent reply have both been processed. We have already seen from the above data that this form of interaction is actually quite unlike that of normal human–human discourse.

However, just as airplanes don't flap their wings in flight, it may also be the case that this slow and 'un-natural' mode of interaction is indeed the optimal mode of usage for such a translation device, and that emulation of natural human speech habits may turn out to be inappropriate for such a technology. On the other hand, it might actually feel more natural for a user if the machine gave encouraging feedback while the conversation was in progress, or if there was some mechanism for the speaker to communicate in fragments rather than in considered and well-formed whole sentences.

Since the machine often has some knowledge of its domain, whether through ontologies, dictionaries, or example corpora, it should be feasible to generate a dialogue interactively by the mutual exchange of fragments. As in a human conversation, where both partners echo, repeat, check, suggest, and challenge, in order to show and reinforce their mutual understanding of the present state of the dialogue. Accordingly, a 'prosody-sensor' should be able to use tone-of-voice information, in addition to the recognized text input, to add fragments appropriately onto the 'understanding stack.'

Since the number of wrapper-type fragments is small (on the order of a hundred or so) they can easily be stored as a dictionary. For each entry a further set of codebook entries detailing the acoustic characteristics of the common prosodic and voice-quality variants can then be stored as a sub-dictionary listing. We have found a codebook size of 16 to be optimal here. As each is recognized, by simple pattern matching, its sub-variant is selected and a flag indicating supposed speaker intention and state added to the discourse stack. Integrating this component into an existing translation system, however, remains as work in progress.

7.4.2 Sensing Affect; Detecting Changes in People from Variation in Their Speaking Style and Tone-of-Voice

There is increasing interest nowadays in the areas related to Affective Computing [6–8], particularly with respect to sensing human states and conditions from external physical cues. Since it is likely that people sense and respond intuitively to the small affect-related changes in prosodic settings when conversing with a human partner, it would be socially beneficial if a machine could also be made sensitive to these cues from the voice.

There has recently been a call in Japan for research into such *proactive* devices for use in an advanced media society. Currently, most mechanical devices work reactively, responding to a command from a user, but certain funding agencies in this

country are hoping that future machines will be able to anticipate the user requirements and perform an appropriate function proactively, without explicit prior control from the user. For these technologies, a degree of quite sophisticated human sensing will be required. However, although the technology itself will be very sophisticated, the information that is being sensed may be quite low-level and primitive.

In the recent meetings-related research (see the ICSI and AMI projects, for example) [9, 10], sensor devices have been invented that detect different degrees of human participation in a multi-party dialogue from simple cues such as amount of bodily movement and coincidences in the timing of simple actions such as nodding. Similar cues can be detected from tone-of voice, laughter, and non-verbal speech sounds that are currently regarded as insignificant.

Machines can be trained to produce a given response when more than one person laughs or when one person makes a given sound (such as a disapproving grunt). By processing differences in the timing, prosody, and frequency of these cues, much information can be gained into the mental states and discourse intentions of the participants.

7.4.3 *Toward the Synthesis of Expressive Speech*

If we are ultimately to produce speech synthesis that resembles human speech in a conversational setting, then we will need a formal grammar of such non-verbal utterances and language models that predict how often, when and which non-verbal speech sounds should be generated in a discourse. Much of this remains as future work, though several proposals have already been made (see, e.g., [11, 12]).

Because non-verbal fragments are typically short single utterances (or discrete groups of repeated syllables), they can be reused verbatim, and there is no longer any need to calculate a join-cost when using concatenative methods of synthesis. Samples of these non-verbal speech sounds can be inserted easily into a stream of synthesized speech. However, because their target prosody can vary not just in pitch but also in voice quality, there is need for a much more precise and sensitive target-cost instead.

There is already considerable research being carried out into the generation of synthetic speech with emotion, but very little into the generation of speech signaling subtle differences in speaker intentions and relationships. Interestingly, in our analysis of a corpus of 5 years of conversational speech recorded in ordinary everyday environments, the amount that was markedly ‘emotional’ accounts for less than 1% of the total, whereas the amount marked for socially related ‘affect’ is probably more than half (see also [13]).

This difference may be a result of volunteers hesitating to give us recordings of their speech that was openly emotional. If they happened to have a blazing argument with their partner on a given day, for example, they may have deleted the recording out of embarrassment or a sense of privacy. Yet the amount of potentially embarrassing personal information that they *did* give us, without hesitation, leads

the author to believe that this is not the case. It is more likely that as socially responsible adults, we moderate our speech so as not to reveal personal emotional details most of the time.

We make more frequent use of subtle prosodic variations to show interest, enthusiasm, boredom, concern, care, relief, etc., i.e. to appear bright, cheerful, intelligent, interested, etc., than we do to openly reveal our actual inner feelings and emotions in everyday conversation.

Since there is already an exhaustive literature on the relations between prosody and syntactic structure, prosody and semantics, and the use of prosody in the expression of contrastive focus, etc., we do not address those issues further here, but instead we claim that the role of affective prosody in interactive speech is just as much to show the partner the speaker's intentions, to clarify stages of the discourse, and to manage turn-taking. This functional interpersonal role of prosody leaves plenty of scope for future research.

There are many applications, apart from human-to-human speech translation, where a natural-sounding voice is required in speech synthesis. This chapter has argued that for the voice to be completely natural sounding, a new level of language structure and discourse control will need to be incorporated into future speech synthesis research.

7.5 Discussion

Being of the UNIX persuasion since the early 1980s, the author has recently found it necessary to make use of Windows software due to the demands of publishers and conference organizers. Since disk access can sometimes be very slow when the data files become fragmented under this operating system, a Windows user soon learns the benefit of frequent use of the MS-Dos command 'Disk-Defrag.' While first considering this as a design weakness in the operating system itself, we now consider that it may indeed represent the 'natural way of things.' Natural speech appears to be similarly fragmented. It seems that when we listen to natural interactive or conversational speech we also perform considerable 'cleanup,' to remove hesitations and 'wrappers,' and then defrag the segments to produce intelligible chunks from the speech sequence.

Accordingly, it is suggested in the present chapter that the evolution of this supposedly 'broken' form of spontaneous speech is not just a side-effect of poor performance in real-time speech generation processes, but that the inclusion of frequently repeated non-verbal speech segments naturally enables the speaker to use them as carriers for affective information such as is signaled by differences in voice quality and speech prosody. Their high frequency and relative transparency with respect to the propositional content allows small changes or contrasts in phonation style to be readily perceived by the listener as carrying significant interpersonal information relevant to the discourse, even if he or she is at first unfamiliar with the speaker.

7.6 Conclusion

This chapter has presented some acoustic findings related to speech prosody and has shown how the voice is used to signal affective information in normal conversational speech. The chapter has shown that natural conversational speech is usually highly fragmented, that these fragments carry discourse and affect-related information, and that by being very frequent, and effectively transparent to the discourse, they function as an efficient carrier for this second channel of information in interactive conversational speech.

The chapter has argued that although modern speech processing technology has come a long way, and appears now to have achieved many of its original goals, it is perhaps time to ‘shift the goalposts’ and become more aware of this secondary channel of information carried in the speech signal which is currently not being processed at all as part of the human communication system.

Acknowledgments This work is partly supported by the Ministry of Public Management, Home Affairs, Posts, and Telecommunications, Japan under the SCOPE funding initiative. The ESP corpus was collected over a period of 5 years with support from the Japan Science & Technology Corporation (JST/CREST) Core Research for Evolutional Science & Technology funding initiative. The author also wishes to thank the management of the Spoken Language Communication Research Laboratory and the Advanced Telecommunications Research Institute International for their continuing support and encouragement of this work. The chapter was written while the author was employed by NiCT, the National Institute of Information and Communications Technology. He is currently employed by Trinity College, the University of Dublin, Ireland, as Stokes Professor of Speech & Communication Technology.

References

1. The Japan Science & Technology Agency. (2000–2005). Core Research for Evolutional Science & Technology.
2. Campbell, N. (2007). On the use of nonverbal speech sounds in human communication. In: *Verbal and Nonverbal Communication Behaviors*, Berlin, Heidelberg, Springer, 2007, LNAI Vol. 4775, 117–128.
3. Campbell, N., Mokhtari, P. (2003). Voice quality is the 4th prosodic parameter. In: *Proc. 15th ICPHS, Barcelona*, 203–206.
4. Alku, P., Bäckström, T., Vilkman, E. (2002). Normalized amplitude quotient for parametrization of the glottal flow. *J Acoust Soc Am*, 112(2), 701–710.
5. Hanson, H. M. (1995). Glottal characteristics of female speakers. Ph.D. dissertation, Harvard University.
6. Cahn, J. (1989). The generation of affect in synthesised speech. *J. Am. Voice I/O Soc.*, 8, 251–256. SSML, The Speech Synthesis Markup Language, www.w3.org/TR/speech-synthesis/
7. Campbell, N. (2005). Getting to the heart of the matter; speech as expression of affect rather than just text or language, *Lang. Res. Eval.*, 39 (1), 109–118.
8. Calzolari, N. (2006). Introduction of the Conference Chair. In: *Proc. 5th Int. Conf. on Language Resources and Evaluation*, Genoa, I–IV.
9. ICSI meeting corpus web page, <http://www.icsi.berkeley.edu/speech/mr>. As of May 2010.
10. AMI: Augmented Multi-party Interaction (<http://www.amiproject.org>). As of May 2010.

11. Schroeder, M. (2004). Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In: Proc. Workshop on Affective Dialogue Systems: Lecture Notes in Computer Science, Kloster Irsee, Germany, 209–220.
12. Campbell, N. (2006). Conversational Speech Synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), 1171–1178.
13. Cowie, R., Douglas-Cowie, E., Cox, C. (2005). Beyond emotion archetypes: Databases for emotion modeling using neural networks. *Neural Netw.*, 18, 371–388.

Chapter 8

Interacting with Embodied Conversational Agents

Elisabeth André and Catherine Pelachaud

8.1 Introduction

The objective to develop more human-centred, personalized and at the same time more engaging speech-based interactive systems immediately leads to the metaphor of an embodied conversational agent (ECA) that employs gestures, mimics and speech to communicate with the human user. During the last decade research groups as well as a number of commercial software developers have started to deploy embodied conversational characters in the user interface especially in those application areas where a close emulation of multimodal human–human communication is needed. This trend is motivated by a number of supporting arguments. First, virtual characters allow for communication styles common in human–human dialogue and thus can release users unaccustomed to technology from the burden to learn and familiarize with less native interaction techniques. Then, a personification of the interface can contribute to a feeling of trust in the system by removing anonymity from the interaction. Furthermore, well-designed characters show great potential for making interfacing with a computer system more enjoyable.

Simply adding a pretty face to an interface is not enough to be an interactive partner. The agent needs to have means to perceive and understand what the user is saying and doing. It should be able to provide appropriate information and answer user’s queries and remarks. The agents have a human-like appearance; they ought to be endowed with human-like communicative, emotional and social capabilities, to be able to display appropriate facial expressions and gestures with their speech. Interaction implicitly creates a social environment where norms and cultural rules are expected to be followed. ECAs may be required to take different roles such as to be a tutor, an information provider, a companion or an advisor. Their goals may be to make students learn their lessons better, to give advice, to show empathy with users’ emotional states, etc. Their behaviours and words ought to be in adequacy with their role and goals. In the last decades, researchers have tackled these complex issues and

E. André (✉)

Multimedia Concepts and Applications, University of Augsburg, Augsburg, Germany
e-mail: andre@informatik.uni-augsburg.de

developed several computational models of agents. In this chapter we review some of the major projects and ECA creations that have happened in the last years. Our aim is to cover some of the most complex areas in the design of ECAs. Much has been done but there are still core issues to be solved. We highlight these as we go along in the chapter.

By means of selected sample applications, Section 8.2 recalls the yet ongoing development of ECAs starting with TV-style information presenters to multi-party multi-threaded conversations between several ECAs and human interlocutors. The subsequent sections then report on mechanisms to realize verbal and non-verbal ECA behaviours. Most of the current systems with ECAs distinguish between an ECA's embodiment and a behaviour control component. Some relate this distinction to the biological body/brain dichotomy. Others take a more technically oriented view and associate embodiment with an animation engine while behaviour control is related to automated multimodal dialogue act generation. Following this distinction, Section 8.3 discusses approaches to determine multimodal dialogue acts for a single agent or a team of agents conversing with one or several human users. After that, we move to the realization of multimodal dialogue acts including not only conversational (Section 8.4), but also emotional signals (Section 8.5) as well as methods to generate expressive behaviour variants (Section 8.6) based on declarative parameter settings, for instance, to reflect an agent's emotional state. To emulate human face-to-face dialogue more closely, it is desirable to avoid asymmetries in communication channels. That is, the ECA has to not only include sophisticated mechanisms for behaviour generation and rendering but also be equipped with perceptive capabilities. Section 8.7 reports on first attempts towards the development of perceptive agents which are able to perceive communicative signals from the human conversational partner, for example, to monitor his or her level of interest in continuing the conversation. The question arises of how to inform the design of ECAs. Are there findings from related disciplines that could be used as a basis for our implementation or should we follow our own intuitions if no concrete guidelines can be obtained? Section 8.8 reports on a design methodology for ECAs that is based on observations of human dialogue behaviours. As new tools for collecting and analysing data of humans have become available, more and more ECA researchers have adopted a data-driven approach which enables them to ground ECA design in empirical data. The proposed design methodology includes four phases: the collection and analysis of data, the building of models, the implementation of ECAs based on the models and finally an evaluation of the resulting ECAs which may require a revision of the other phases. Studies on the evaluation of ECAs are discussed in more detail in Section 8.9 where we focus on the dialogue behaviours of ECAs.

8.2 Types of Conversational Settings

In the area of embodied conversational agents, we can observe an ongoing and manifold evolution that is characterized by an increased complexity of

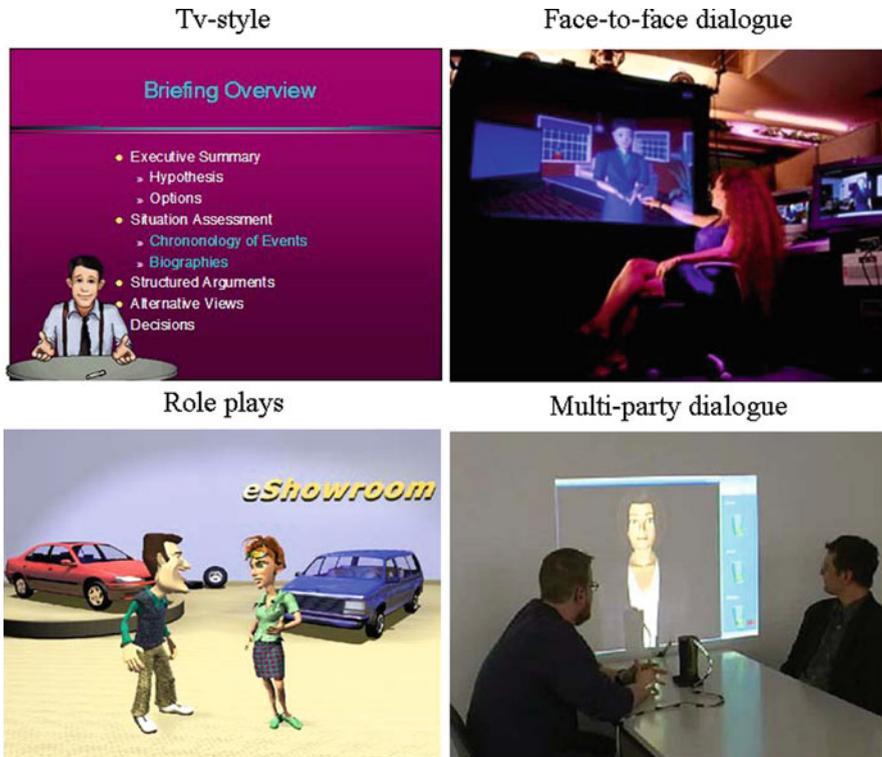


Fig. 8.1 Different conversational styles

conversational styles. As shown in Fig. 8.1, four classes of conversational styles may be distinguished.

8.2.1 TV-Style Presenters

The first type of conversational style is employed in applications in which a single character is deployed to present information. A popular example is the virtual news-reader Ananova (www.ananova.com) that reads news scripts live from ITN, a British broadcaster producing news. Further examples include the virtual weather reporter [1], DFKI's PPP persona [2], the AutoBriefer agent which presents multimedia briefings generated from high level outlines [3], as well as virtual product presenters found on commercial web pages. From the point of view of the user, viewing a presentation appears quite similar to watching a TV news speaker or to watching a video clip because the flow of information is unidirectional from the agent to the user and no user-agent interaction is foreseen at display time. Therefore, the character's style of presentation is similar to a monologue, though multiple modalities may be deployed.

8.2.2 Virtual Dialogue Partners

A great number of contemporary systems aim to emulate aspects of a face-to-face dialogue with a character that is able to converse with the user. Differences among systems concern both available input modalities as well as output modalities of the virtual conversation partner.

Quite a number of commercial sites try to boost their web presence by means of virtual sales personnel that provide customers with a more personalized online shopping experience 24 h a day, 7 days a week. In most cases the user can “talk” to the character by typing NL expressions into a text-input widget while the character talks to the user either by voice output or likewise through speech bubbles. However, the conversational skills of these characters are often quite limited to that of a pattern-based chat robot that works similar to Weizenbaum’s early Eliza system [4]. A collection of web-based chatter bots may be found under <http://www.knyetrypper.com/>. In the best case, such systems manage to map user input onto related contents available at the web site. In the worst case, a conversation with such a character is neither informative nor entertaining. Needless to say that such characters are likely to be perceived by a user as useless if not annoying. In contrast, most research prototypes of embodied conversational characters are instances of complex multimodal dialogue systems, though the focus is usually on the generation of synchronized multimodal expression. Prominent examples include the museum guide August [5], the real estate agent REA [6], the GRETA Medical Advisor [7], the agent MAX [8], and the animated interface character Smartakus that is developed in the SmartKom project [9]. Most of these systems rely on sophisticated models for multimodal conversation. For instance, Smartakus incorporates a sophisticated spoken dialogue subsystem and has a “visual sense” that enables it to recognize and understand pointing gestures of the user.

8.2.3 Role-Plays and Simulated Conversations

There are situations in which direct agent–user communication is not necessarily the most effective and most convenient way to present information. Inspired by the evolution of TV commercials over the past 40 years, André and colleagues [10] have studied role-plays with virtual characters as a promising format for presenting information. A typical commercial of the early days of TV featured a sales person who presented a product by enumerating its positive features – quite similar to what synthetic characters do on web pages today. On TV, however, this format has been almost completely replaced by formats that draw on the concept of short entertaining scenes. Typically, such performances embed product information into a narrative context that involves two or more human actors. Episodic formats offer a much richer basis compared to the plain enumeration of product features, and thus meet the commercial industry’s high demand for originality.

André and colleagues [10] suggest the use of presentation teams to convey information about products, such as cars, by performing role-plays. Using this presentation style, the user receives information about cars by watching generated sales dialogues among virtual seller and buyer agents. The eShowroom allows the user to specify prior to a presentation (a) the agents' roles, (b) their attitude towards the product, (c) some personality traits (extrovert vs. introvert, agreeable vs. not agreeable) and (d) their interests about certain aspects relevant for cars (for example, the car's relation to prestige, comfort, sportiness, friendliness to the environment, costs, etc.). Based on these settings, a variety of different sales dialogues can be generated for the same product.

Using such a setting actually means a shift from a face-to-face character-user setting to a user-as-observer setting. The shift is motivated by a number of supporting arguments: First of all, they enrich the repertoire of modalities to convey information. For example, they allow a system to convey certain rhetorical relationships, such as pros and cons, in a more canonical manner. Furthermore, they can serve as a rhetorical device to reinforce beliefs. For instance, the same piece of information can be repeated in a less monotonous and perhaps more convincing manner simply by employing different agents to convey it. In addition, the single members of a presentation team can serve as indices, which help the user to organize the conveyed information. For instance, characters can convey meta-information, such as the origin of information, or they can present information from different points of view, for example, from the point of view of a businessman or the point of view of a traveller. Furthermore, scenarios with multiple characters allow us to model interpersonal social relationships (see [11, 12]). For example, Rehm and colleagues [13] modelled a virtual beer garden where agents wander around to meet friends or to build up new relationships.

8.2.4 Multi-threaded Multi-party Conversation

Casting role-plays with characters that can interact with both other characters and the user results in an open multi-party dialogue setting, which supports reactive as well as proactive user participation. One basic idea is to provide the user with the option of taking an active role in the dialogue if she or he wishes to do so. If not, however, the characters will continue the conversation on their own – maybe encouraging the user to give feedback from time to time.

Traum and Rickel [14] have addressed the issue of automatically generated multi-party dialogues in immersive virtual environments. In the context of a military mission rehearsal application, they address dialogue management comprising human–character and character–character dialogues. The characters are based on the Steve architecture [15] which has been enhanced by a multimodal dialogue model to handle turn taking in such a challenging scenario.

A number of approaches to multi-party conversation have been inspired by research on interactive drama that aims at integrating a user in a scenario – either as

an audience member or as an active participant. An example includes the interactive installation *CrossTalk* [16] that has been designed for set-up in a public space, for example, at a booth of a trade fair. *CrossTalk* takes Brenda Laurel's [17] paradigm of "computers as theatre" a step further and introduces a combination of theatre and meta-theatre. In *CrossTalk*, characters permanently switch between two modes: (1) a presentation mode where a performance is given to an audience and (2) an off-duty mode where the characters engage in small talk among each other. In this way, the small talk between the characters becomes another performance that serves to catch the interest of passers-by and may be called "meta-theatre". The *VicTec* system (see [18]) realizes a multiagent learning environment to teach children strategies against bullying relying on a Forum Theatre metaphor in which spectators play a role in the unfolding of the narrative. In particular, the user may influence the storyline by interacting with one of the agents and by suggesting plans of action.

Hardly any work so far has been conducted on the realization of scenarios with multiple users and synthetic agents. An exception includes the work by Isbister and colleagues [19] who concentrate on social interaction between several humans in a video chat environment which is supported by a so-called Helper Agent. Helper Agent is an animated, dog-faced avatar that tracks audio from two-person conversations and intervenes if it detects longer silences. Rehm and colleagues [13] focus on a game scenario in which the agent takes on the role of a co-player. By projecting the agent on a screen at the end of the table, they convey the impression that the agent is sitting together with the other players at the table. To allow for a more natural simulation of the traditional game of dice, the users do not interact via a keyboard, but use a tangible interface to toss the dice and communicate via voice with the other players and the agent. In this case, the dialogue flow is not driven by theatrical principles to structure the conversation, but controlled by the rules of the game.

8.3 Dialogue Management

Depending on the conversational settings discussed in Section 8.2, different approaches to dialogue management have been developed.

Scenarios with one or several embodied conversational agents that do not allow for any user interaction often rely on scripts. Thereby, a script is a temporally ordered sequence of dialogue actions that may be realized using not only speech but also body gestures, facial expressions and verbal utterances. A straightforward approach is to equip the character with a library of manually authored scripts that determine what the character might do in a certain situation. At runtime, the remaining task is to choose from the library a suitable script that meets the constraints of the current situation and at the same time, helps to accomplish a given task.

A particular problem with manually authored scripts and script libraries is that the author has to anticipate scripts for all possible situations and tasks, and that the scripts must allow for sufficient variations in order to avoid characters that behave

in a monotonous and too predictable way. Furthermore, the manual scripting of characters can become quite complex and error-prone since synchronization issues have to be considered. Creating scripts manually is, however, not feasible for many applications since it would require anticipating the needs of all potential users and preparing presentations for them.

To automate the scripting process, planning approaches have been proven useful. An example includes the Inhabited Market Place that creates dialogues between several ECAs automatically by employing a centralized planner. Character-specific dialogue contributions (for example elementary speech acts) constitute leaf nodes in the decomposed hierarchical planner. Systems, such as Miao [20], employ a self-scripting approach and assign each agent its own reactive planner. Dialogue contributions then result from autonomous characters trying to achieve their individual goals. In order to control the flow of dialogue to a certain extent, the Miao system, also foresees a director component. This component may intervene in the course of a conversation, for example, to increase coherence or to bring in new topics. A great challenge of de-centralized multi-party dialogue is to determine who is addressed by whom and to decide who is supposed to get the next turn. Rist and colleagues [20] present an approach that selects the next speaker based on his or her motivation which depends, among other things, on his or her personality.

While the approaches discussed above start from a communicative goal that is decomposed into elementary dialogue acts, the BEAT system [21] starts from text that is automatically converted into embodied speech using a generate-and-filter approach. After a shallow analysis of the text, for example, to identify the rheme and the theme, the text is annotated with plausible gestures based on rules that are derived from studies of human–human dialogue. Modifiable filters are then applied to trim the gestures down to a set appropriate for a particular character. For instance, specific filters may be used for an introvert character resulting into communicative behaviours with fewer gestures. Furthermore, it may happen that the initially proposed gestures cannot co-occur physically so that some of them have to be filtered out.

A number of ECA implementations rely on dialogue managers that have been originally developed for pure speech-based human–computer communication, but may be applied in a broader multimodal context as well. For instance, both the MagiCster [7] and the MRE projects [14] developed conversational settings with ECAs that follow the Trindi dialogue manager [22]. Due to the flexibility of description, the Trindi dialogue manager allows choosing dialogue moves that may be realized by multiple modalities and updating the information state of the dialogue based on declaratively stated selection and update rules.

For the realization of collaborative human–computer dialogue, the Collagen framework [23] has been proven useful which relies on a framework of collaborative discourse theory to generate multimodal discourse acts. Examples include the generation of posture shifts in the REA agent [6], the determination of tutorial strategies for the Steve agent [24] and the creation of head nods for the robotic Mel agent [25].

While the approaches above focus on sophisticated behaviour mechanisms for fully animated conversational agents, there may be scenarios with a large number of characters in which it is no longer feasible to equip each character with fully fledged conversational behaviours. Jan and Traum [26] present an approach to dialogue simulation for background characters that allows for dynamically starting, joining or leaving a conversation. A similar approach has been presented by Rehm and colleagues [13] who developed a model for group dynamics that simulates changes in the social relationships between agents as a side-effect of social interactions. Based on the model, they designed and implemented a game-like scenario which has been employed as a test bed to study social navigation behaviours of humans. Their system does not plan the contents of an utterance. Instead they just determine elementary interaction types for the single agents based on Interaction Process Analysis (IPA) by [27], such as ask for suggestion, show tension or show solidarity, which are then reflected by the agents' posture and gestures. IPA has also been successfully employed in other system of social group dynamics, see for example [28] or [29].

8.4 Communicative Signals

We communicate through our choice of words, facial expressions, body postures, gaze, gestures etc. Non-verbal behaviours accompany the flow of speech and are synchronized at the verbal level, punctuating accented phonemic segments and pauses. They may have several communicative functions [30, 31]. They are used to control the flow of conversation; that is they help in regulating the exchange of speaking turns, keeping the floor or asking for it. Actions such as smiling, raising the eyebrows and wrinkling the nose often co-occur with a verbal message. They may substitute for a word or string of words, or emphasize what is being said. Gestures may indicate a point of interest in space or describe an object. Facial expressions are the primary channel to express emotions. They can also express the attitude towards one's own speech (such as irony) or towards the interlocutor (like showing submission). Non-verbal behaviours do not occur randomly, but rather are synchronized to one's own speech, or to the speech of others [32–34]. Raised eyebrows go along with emphasized words [35–37]; the stroke of a gesture, that is the most forceful part of a gesture, happens also on emphasized words or just before [38]. Most of the gesturing happens when speaking. Hands and faces come to a rest when speech ends [39].

Isabella Poggi proposes a taxonomy of communicative functions based on the meaning they convey [30]:

1. information about speaker's beliefs: behaviours that provide information on the speaker's beliefs such as the degree of certainty on what she is talking.
2. information about speaker's intentions: the speaker may provide information on her goal through, for example, her choice of performative or the focus of her sentence.

3. information about speaker's affective state: the speaker may show her emotion state through particular facial expressions.
4. meta-cognitive information about speaker's mental state: the speaker may try to remember or recall information.

A communicative function is defined as a pair (meaning, signal). Each function may be associated with different signals; that is for a given meaning, there may be several ways to communicate it. For example, the meaning "emphasis" (of a word) may co-occur with a raised eyebrow, a head nod, a combination of both signals or even a beat gesture. Vice versa, a same signal may be used to convey different meanings; for example, a raised eyebrow may be a sign of surprise, of emphasis or even of suggestion.

Pelachaud and colleagues [40, 41] have created an Embodied Conversational Agent, Greta, that incorporates communicative conversational and emotional qualities. The agent's behaviour is synchronized with her speech. It is consistent with the meaning of the sentences she pronounces. To determine speech-accompanying non-verbal behaviours the system relies on the taxonomy of communicative functions just presented [30]. To control the agent's behaviour they are using a representation language, called Affective Presentation Markup Language (APML), where the tags of this language are these communicative functions [42]. An example of APML and the corresponding greeting gesture are shown in Fig. 8.2.

Several models have been proposed for agent's behaviour selection and agent's behaviour animation. Work in behaviour selection has mostly been concerned with semantic aspects of human gesturing. McNeill's classification of gestures [38] is often used. He defined several communicative gesture types: iconic (refers to some physical/spatial/temporal properties of the speech referents), metaphoric (refers to abstract property of the speech referents), deictic (indicates a concrete or abstract point in the world), beat (follows the rhythmic structure of the speech) and emblem (has well-specified meaning). As pointed out by several studies, humans gesture a lot while communicating [38, 33]. Gestures may have several functions such as giving information (depicting the size of a box while mentioning it), smoothing the interaction (indicating one wants to take the speaking turn), providing information on the mental state of the speaker (showing how certain one is of one's assertion), and so on. Endowing ECAs, human-like figures, with communicative gestures is therefore important to enable them to convey information in a more natural way.

REA [43] is a humanoid agent able to understand the user's behaviour and respond with appropriate speech, facial expressions, gaze, gestures and head movements. It is a real state agent that can chat with users not only about houses but also on everyday things. While talking, REA shows appropriate iconic and beat gestures that accompany her speech. In particular REA can provide information via gestures that complement her speech; she will indicate with her hands and arms the shape of a garden or if the stair of the house is ellipsoidal. Moreover, REA changes body posture when changing topics of conversation or speaking turn.

```

<APML>
<turn-allocation type="take turn">
  <performative type="greet">
    Good Morning, Angela.
  </performative>
</turn-allocation>
<affective type="happy">
  It is so
  <topic-comment type="comment"> wonderful </topic-comment>
  to see you again.
</affective>
  <certainty type="certain">I was
  <topic-comment type="comment"> Sure </topic-comment>
  we would do so, one day!
</certainty>
</APML>

```



Fig. 8.2 Example of APML text input (*top*) Greeting gesture that gets instantiated from the APML example (*bottom*)

The MAX agent [44, 45] can be either a conversational assistant in construction tasks or a virtual receptionist that engages visitors in a museum hall. It can select and generate complex iconic, beat and emblem hand gestures based on the semantic content of discourse. Gestures are described at a symbolic level as a combination of hand configuration, wrist orientation and movement in space as well as palm orientation. The timing of the gesture phases are instantiated following the agent's speech. Recently MAX has been improved to track the user's gaze and drive its own eyes/head direction accordingly.

Head movements hold an important role in conversation and researches have been done to determine their pattern in order to enrich ECAs with more believable head animation. Heylen analysed head patterns to define their properties and

functions [46] useful to implement ECAs' behaviour. "August" [5], a talking head able to give information about the Stockholm Culture Center, replies to museum visitors showing head movements and facial expressions.

8.5 Emotional Signals

We inform our addressee about the emotions we feel while talking (by affective words, gestures, intonation, facial expression, gaze and posture). They may be elicited by the evaluation of an event, an action or a person [47, 48]; some emotions are triggered by an event and are not directed towards someone (examples are emotion of fear, surprise); one can feel emotion towards another person (such as love, scorn, hate). Emotions can be represented using categories (for example, happiness, anger), dimensions (such as valence, activation) or appraisal dimensions (for example, suddenness, familiarity of the events).

Ekman and his colleagues have proposed the existence of universal facial expressions linked to six emotions, namely anger, disgust, fear, happiness, sadness and surprise [49]. These emotions are often referred to as basic emotions. For example, anger can arise from frustration, physical threat or psychological harm. In the case of anger, the eyebrows form a frown, and the mouth can either be pressed firmly together or open in a tense manner.

Of course, there exist several variants of facial expressions for each emotion. Moreover the display of the facial expression corresponding to an emotion will be modulated by one's own culture, social environment, to whom one is talking [11, 35, 50]. Many agent systems have modelled solely the expression of the six basic emotions [51–53], but later models have considered a large set of emotions [54, 55].

Rather than using a categorical representation of emotions, Tsapatsoulis et al. [56] and Albrecht et al. [57] use a dimensional representation to compute the facial expressions of non-basic emotions. In the work of Tsapatsoulis et al. [56], each emotion is represented by coordinates in the 2D space introduced by Whissel [58] and Plutchnik [59]. Emotions are described by their coordinates along the axes (positive/negative) *valence* and (active/passive) *activation*. Facial expressions are described by MPEG-4 parameters [53]. Albrecht et al. [57] adapt this model to muscular facial model and use a 3D emotional space description. The dimension *power* was added to the dimensions *activation* and *valence*.

The model called Emotion Disc [60] uses bi-linear interpolation between the two closest basic expressions and the neutral one. It is based on Schlosberg's results according to which expressions of emotions can be placed in a 2D space [61]. The disc is divided into six identical sections, one for each emotion. The neutral expression is at the centre of the disc. The intensity of an emotion increases towards the outer part of the disc. Thus, the distance of an emotion from the centre represents the intensity of the corresponding expression. The spatial relations between neighbouring emotions and the centre of this circle are used to establish the expression corresponding to any point on the Emotion Disc.

A different approach was introduced by Duy Bui [55]. It follows Ekman's concept that blends are constructed by composing different parts of basic expressions [49, 62]. Duy Bui uses a set of fuzzy rules to determine the blended expressions of the six basic emotions. In this approach a set of fuzzy rules is attributed to each emotions pair. Each rule decides, based on the emotion intensity, about the intensity of muscle contraction for the blended expression. The final expression is obtained using fuzzy inference.

In a study on deceptive agents, Rehm and André [63] show that users are able to differentiate between the agent displaying an expression of felt emotion versus an expression of fake emotion. This last expression type is differentiated by the type of action involved in the final expression. For example, a polite smile (associated here as a fake smile) does not involve raising the cheeks that in turn produces the characteristic crow feet wrinkles, characteristic of true happiness. Fake expressions are also more symmetrical than felt expressions; they arise too fast or too slow on the face. Rehm and André focused on visual cues to deception. For an analysis of language and speech cues to deception, we refer to Chapter 5.

Real life situation may trigger simultaneous emotions as an event may be evaluated along different aspects. Moreover emotions may not be displayed due to some socio-cultural norms. A model of blend of emotions in the case of super position of two emotions and of masking one emotion by another has been developed [64, 65]. This model is based on Ekman's research [49, 62] and is implemented using fuzzy logic. Faces are decomposed into several areas. Inference rules are applied to combine facial expressions of both emotions. Two different sets of inference rules have been derived to simulate both types of blending.

The Facial Action Composing Environment (FACE) tool generates facial expression [66, 67]. It is based on Scherer's component process model of emotion [48] which stipulates that facial expressions are the result of a sequence of appraisal checks. Each check may trigger specific actions in different face areas. Each emotion is linked to a different sequence of appraisal checks, thus resulting into different facial expressions. In this theory, blends of emotions can be seen as arising from several complex appraisal checks. The model has been developed using an experimental and modelling loop [67]. The tool GAME [68] is used to induce emotions in users. It is composed of a series of game scenario. For each scenario, FACE is linked to the tool GAME, a generator of games that is used to induce emotions into users [68]. GAME records the game process and the user's actions. A user's facial expression is analysed and annotated automatically. Each game scenario corresponds to a set of appraisal checks. A correspondence between appraisal checks and facial expressions can be done. FACE is used to replay the facial expression animation.

8.6 Expressive Behaviours

A behaviour is defined by a given facial expression or by a particular hand configuration and arm position. There is also another element that characterizes a behaviour:

the manner of execution of the behaviour; we call this parameter the expressivity of the behaviour. Until now the behaviour has been described statically: a facial expression is defined at its apex (i.e. the time at which the facial expression comes to its peak intensity value) [35] and the shape of a gesture is specified by the shapes it has over the various phases that composed it (for example, preparation phase, stroke) [38]. The expressivity parameter refers to the dynamic variation of the behaviour along this static description, for example, the temporal duration and strength of the behaviour.

So far, gesture animation has been mainly concerned with realistic movement generation of an agent's arm and hand. Few efforts have been done to model gesture expressivity. Earlier notable examples of generating expressive movement include Perlín and Goldberg's use of noise functions [69], and Bruderlin et al.'s use of signal processing techniques to change frequency characteristics [70]. EMOTE by Chi et al. [71] is based on dance annotation as described by Laban [72]. In particular, it implements the Laban principles of Effort and Shape to render gesture performances more expressive by varying kinematic and spatial curve properties of existing keyframed actions. EMOTE acts as a generic filter on pre-existing behaviours; it can generate a wide variety of motions by manipulating pre-existing animations.

Several perceptual studies arrived at a dimensional characterization of expressivity in human bodily movement [73, 74]. Hartmann et al. [75] implemented six of these dimensions. These six dimensions have been designed for communicative behaviours only. Each of them acts differently for each modality. For the face, the expressivity dimensions act mainly on the intensity of the muscular contraction and its temporal course (how fast a muscle contracts). On the other hand, for an arm gesture, expressivity works at the level of the phases of the gesture: for example, the preparation phase, the stroke, the hold as well as on the way two gestures are co-articulated. Three of the dimensions of expressivity, namely spatial extent, temporal extent and power, act on a given gesture and facial expression parameters (respectively, amplitude of movement corresponding to physical displacement of a facial feature or a hand position), duration of movement (linked to execution speed of the movement) and dynamics properties of the movement (acceleration value of the movement) (see Fig. 8.3). Another dimension, fluidity, acts over several behaviours of a same modality and specifies how smoothly a behaviour is mapped in another one. The last two dimensions, overall activation and repetitiveness, act, respectively, on the quantity of behaviours and on the repetition of a given behaviour. Each dimension varies from low to high values. For example, Fig. 8.3 depicts a same gesture made with different values of spatial extent: the gesture goes from contracted (low value) to very extended (high value).

Egges et al. [76] present a system that generates automatically idle movements for a virtual agent depending on its emotional state. The method is based on a Principal Component Analysis of motion capture data. Idle movements are rarely considered in animation. The idle animation model by Egges consists of balance shifts and small variations in posture. These idle motions are added to conversational gesture animation by a blending engine.

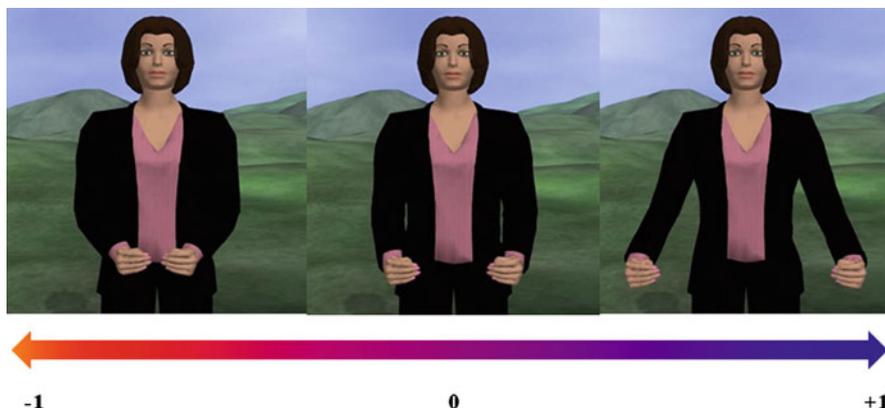


Fig. 8.3 Variation of the spatial extent parameters for a given gesture. From *left to right*: gesture with low spatial extent value, neutral (“normal”) gesture and gesture with high spatial extent value

8.7 Perceptive Behaviours

Face-to-face communication does not take place in a vacuum, but should be linked to the surroundings of the conversational partners. An embodied agent may acquire knowledge about its surroundings by a sensory mechanism and/or by consulting a model of the interaction space that might be mentioned during the interaction.

Earlier agents, such as the PPP Persona, were usually tied to the desktop and had only limited perceptual abilities to access the user’s physical surroundings [2]. The agent’s knowledge about the users and their physical environment was primarily based on pre-stored information and simple text-based interactions. More recent projects equip animated agents with a set of sensors to detect and track people in front of the screen. Examples include the kiosk agents Mack [77] and Max [8].

In order to come across as believable, an agent should not just be equipped with a sensory mechanism. Instead it should reflect psychologically plausible attending behaviours that combine locomotion, visual search and response to peripheral events. For instance, an agent that continuously stares at a target while moving towards it may appear rather unnatural. A first attempt to automate attending behaviours of a synthetic agent has been proposed by Chopra and Badler [78]. Examples of behaviours include spontaneous looking, visual search and monitoring of events in a 3D environment.

Humans usually rely on a variety of conversational grounding cues including head pose, gestures and eye gaze to indicate which information they consider to be common ground. The establishment of common ground is an important prerequisite to conduct a successful interaction. However, only a few characters are able to perceive non-verbal grounding cues from the user. An exception includes the work by Nakano and colleagues [79] who developed a model of grounding for embodied

conversational characters. This has been tested within the Mack agent that analyses the user's head movements and gaze to establish a common understanding between user and agent. Sidner and colleagues [25] implemented a conversational robot that is able to track the face of the conversational partner to find out which objects he or she is paying attention to. Depending on the recognized non-verbal conversational cues of engagement, the conversation topic is adjusted to the presumed interests of the human interlocutor.

In order to enable an agent to reason about another agent's attentive behaviours, Peters [80] proposes to rely on a theory of mind. The model by Baron-Cohen [81] seems to offer a promising approach since it relates gaze perception to higher level cognitive processes. Based on this model, Peters proposes the implementation of a synthetic vision module that considers to what extent the orientation of an agent's eyes, head and body are directed towards the conversational partner in order to determine its attention towards the conversational partner.

The SmartKom agent is able to recognize and understand the user's pointing gestures on a display and tries to read a small number of emotional expressions from his or her face [9]. The hypotheses from the vision component are then merged with the hypotheses from a recognizer for emotional speech [82] relying on adaptive confidence measures. For instance, facial expressions are recognized from different parts of the face, such as the eye and the mouth region. However, when the user is speaking, the mouth shape might no longer provide reliable information on the user's emotion so that the confidence values of the results obtained for the mouth region have to be decreased.

In order to simulate natural behaviours of a listening agent, Maatman and colleagues [83] analyse features of a speaker's behaviour that are available in real-time including posture shifts, head movements and acoustic features. Among other things, their agent is able to mimic the speaker's behaviour or to use backchannel continuers, such as *mmhmm*, based on the detected speaker's behaviour.

8.8 Social Talk

The concept of a virtual character promotes the idea that humans, rather than interacting with tools prefer to interact with an artefact that possesses some human-like qualities at least in a large number of application domains. That is, a virtual character is not just another interface gadget. It may become a companion and even a friend to the user.

To build up socio-emotional relationships between humans and agents, we need to move away from pure task-based dialogue to social dialogue including aspects such as politeness and small talk. The emulation of social dialogue places new demands on discourse planning resulting from the fact that we do not only have to include task-based goals, such as informing the user how to operate a certain device, but also have to account for social goals, such as building up relationships of trust and likeability. Social goals are persistent in the sense that they have to be taken into account over the whole period of the conversation or even over a series of

conversations and realized in parallel with potential task goals. For instance, a tutor agent may have the objective to tell the student how to solve a particular equation and at the same time pursue the social goal of increasing her self-confidence.

While most approaches to social dialogue concentrate on short-term interactions, Bickmore and Cassell [84] focus on the question of how to build up trust between an agent and a human user over a series of interactions. To achieve this goal, they enhance their REA agent by the ability to engage in small talk. Dynamically changing socio-emotional relationships between a human and an agent are represented by a relational model that includes three dimensions: solidarity, familiarity and affect. In order to interleave the realization of task-based and social goals, they developed a computational model of mixed task and social dialogue by adapting an activation network-based approach to discourse planning. The objective of the system is to find conversational moves that pursue task goals as efficiently as possible, but at the same time minimize the expected face threat to the user and maximize trust (as a function of solidarity, familiarity and affect). Depending on the expected face threat and the state of the relational model, the agent decides whether or not to engage in small talk and what kind of small talk to choose. The work is grounded in a theory of social talk and in addition informed by a number of pilot experiments.

When humans interact with each other, they continuously risk threatening the face of their conversational partners, for example by showing disapproval or by putting the other person under pressure. To avoid this, humans usually rely on various face threat mitigation strategies. Bickmore and Cassell [84] focused on small talk as one means to mitigate face threats. Other researchers developed discourse planners incorporating various politeness tactics as described in Brown's and Levinson's Theory of Politeness [85] (B&L Theory). Strategies identified by Brown and Levinson include negative politeness (for example, showing approval for the addressee), positive politeness (for example, emphasizing the addressee's freedom) and off-record statements (for example, vague indications that an action is needed). For instance, instead of formulating a direct request "Solve the equation", a teacher might rely on a mixture of positive and negative politeness and say "Why don't we solve the equation together?".

Walker et al. [86] presented one of the first approaches to implement politeness strategies. They relied on the B&L Theory and described a selection mechanism that is based on the variables power, social distance and ranking of speech act according to the degree of conveyed politeness. Johnson et al. [87] investigated the potential benefits of politeness in a tutoring system. Examining the interactions between a real tutor and his students, they came up with a set of templates each of which is annotated according to the amount of redress that tactic gives to the learner's face. The templates have been employed in a tutorial tactic planner that selects different tutorial strategies automatically depending on the type of expected face threat. In Johnson et al. [88], they investigated how far politeness theory applies equally to tutorial dialogue tactics in English and in German. Rehm and André [89] conducted a corpus study with human speakers to shed light on the question of how face threats are mitigated by non-verbal means. Their study revealed that gestures are used to strengthen the effect of verbal acts of politeness. In particular, vagueness

as a means of politeness is not only reflected by verbal utterances, but also by gestures. Iconic and deictic gestures were predominantly used in more direct criticism while metaphoric gestures frequently occurred in off-record strategies. The results of the corpus analysis were employed to inform the gesture selection mechanism of an ECA. While Bickmore and Cassell [84] focus on the level of discourse planning, work on the implementation of politeness tactics focuses on the question of how to redress a face threat by appropriate verbal and non-verbal realization strategies.

8.9 Design Methodology for Modelling ECA Behaviours

The design of an ECA that emulates aspects of human–human face-to-face conversation is an interdisciplinary effort that may greatly benefit from models developed in the cognitive sciences, social psychology, conversational analysis and many other related fields. The question arises of how to inform the design of the ECA if a direct operationalization of models found in the literature is not possible or if existing models lack ECA-relevant information.

Cassell [90] describes a design loop for ECAs which includes four (potentially iterating) phases: observation, modelling, generation, evaluation. In the observation phase, data of humans that are engaged in a dialogue with other humans are collected. The easiest possibility to make use of these data would be a direct replication of human–human behaviours in an ECA. This method would require, however, collecting data for each possible conversational situation an ECA would encounter. Since such an approach is not feasible, the generation of an ECA should rather be driven by a formal model. Such models are built up by extracting relevant parameters of the collected multimodal data, such as the frequency of certain kinds of gesture. It is important to note that in most cases formal models are not built from scratch. Rather, the data analysis serves to refine existing models. For instance, Rehm and André [89] started from the B&L Theory of Politeness which focused on verbal means in politeness tactics and conducted a study to identify regularities in the use of gestures. The resulting models of human–human conversational behaviour then serve as a basis for the implementation of ECAs that replicate the behaviours addressed by the models. In order to identify gaps in the model, Cassell's design loop also includes an evaluation phase. Here, experiments are set up in which humans are confronted with ECAs following the model. Depending on the outcome of such experiments, the developers might decide to acquire new data from humans so that the existing models may be refined.

To get insight into human–human conversation, ECA researchers rely on a large variety of resources including recordings of users in “natural” or staged situations, TV interviews, Wizard of Oz studies, and motion capturing data. Various annotation schemes have been designed to extract relevant information for multimodal behaviours, such as facial expressions, gestures, postures and gaze. In addition, there has been increasing interest in the design of annotation schemes to capture

behaviours. The ECA is driven by these different levels of annotation. An evaluation study has been conducted to understand the role of these different levels of annotation in the perception of complex emotions [94].

The use of data-driven approaches provides a promising approach to the modelling of ECA behaviours since it allows us to validate design choices empirically. Nevertheless, the creation of implementable models still leaves many research issues open. One difficulty lies in the fact that an enormous amount of data is needed to derive regularities from concrete instantiations of human–human behaviour. In rare cases, we are interested in the replication of behaviours shown by individuals. Rather, we aim at the extraction of behaviour profiles that are characteristic of a group of people, for example, introverts versus extroverts. Furthermore, the resulting ECA behaviours only emulate a limited amount of phenomena of human–human behaviours. In particular, the dynamics of multimodal behaviours has been largely neglected so far. Last but not least, there is the danger that humans expect a different behaviour from an ECA than from a human conversational partner which might limit the potential benefits of a simulation-based approach.

8.10 Evaluation of Verbal and Non-verbal Dialogue Behaviours

After more than 20 years of intensive research and development, the technology for ECAs has become mature enough for evaluation. In this section, we report on empirical studies that focus on the conversational behaviours of ECAs. A more comprehensive overview on evaluation of ECAs is given by [95].

8.10.1 Studies Focusing on the Relationship Between Verbal and Non-verbal Means

A first experiment in order to test different strategies that combine multiple modalities when generating ECA behaviours (redundancy, complementarity and speech-specialization) was performed by Buisine and colleagues [96]. In the redundancy condition, relevant information was conveyed by both speech and arm gestures whereas in the complementarity condition half of the information was given by speech and the other half was given by gestures. In the speech-specialization condition, all information was presented by speech. The objective of the experiment was to find out whether individual multimodal strategies would be perceived by human users, whether human users would express a preference for any of the strategies and whether the different strategies would have an impact on the performance of the human users. Their experiment revealed that the single strategies were rarely consciously observed by the single users. Taken as a whole, males and females did not prefer a specific strategy. For male users, the authors observed a preference for redundant presentations while there was no such difference for females. Due to the small number of subjects, this result should, however, be handled with care. No effect of the type of multimodal strategy applied by the ECA was observed

for perceived expressiveness. Rather, the amount of movements, the voice and the ECA's appearance might have an influence on the perception of expressiveness as comments given by the subjects at the end of the experiment indicated.

A number of studies conducted by Nass and co-workers reveal the importance of consistencies in multimodal behaviour. Lee and Nass [97] observed that a user's feeling of social presence is positively affected if the personality that an utterance conveys is reflected by the employed synthetic voice as well. Nass and Gong [98] claim that maximizing the quality of each modality does not necessarily improve human-computer interaction. Even though recorded human voices are more natural than the output of a text-to-speech synthesizer, an interface may become more appealing when a synthetic face is accompanied by a synthetic voice.

8.10.2 Studies Investigating the Benefit of Empirically Grounded ECA Dialogue Behaviours

As noted in Section 8.9, an increasing number of approaches to modelling the behaviours of ECAs are based on a simulation of human behaviours. The question of whether the use of models that are informed by studies of human-human conversation actually leads to better human-ECA communication arises. In the following, we present a number of studies that address this issue.

Garau and colleagues [99] as well as Lee and colleagues [100] investigate the effect of informed eye gaze models on the perceived quality of communication. Both research teams observed a superiority of informed eye gaze behaviours over randomized eye gaze behaviours. A follow-up study by Vinayagamorthy and colleagues [99] focused on the correlation between visual realism and behavioural realism. They found that the model-based eye gaze model improved the quality of communication when a realistic avatar was used. For cartoonish avatars, no such effect was observed.

A study by Nakano and colleagues [79] revealed that an ECA with a grounding mechanism seems to encourage more non-verbal feedback from the user than a system without any grounding mechanism. Sidner and colleagues [25] showed that users are sensitive to a robot's conversational gestures and establish mutual gaze with it even if the set of communicative gestures of the robot is strongly limited.

Summing up, it may be said that the effort required to empirically ground ECA behaviours seems to pay off. Nevertheless, humans behave quite differently when conversing with an agent as opposed to conversing with another human as we see in the next section.

8.10.3 Studies Investigating the Dialogue Behaviours of Humans Interacting with an ECA

Kopp and colleagues [45] investigated conversations of humans with the public museum guide Max. The results indicate that humans are willing to engage in social

interactions with the agent. For instance, the participants relied on common strategies from human–human dialogue, such as greetings, farewells and commonplace phrases, when interacting with Max. In addition, the authors reported that the agent seemed to be more active than the human user both in terms of utterances and words per utterance. The authors attribute this finding to the fact that the keyboard was used as an input device. Furthermore, they detected flaming and implicit testing of the agent’s intelligence.

While Kopp and colleagues concentrated on the humans’ verbal behaviours, Rehm and André [101] investigated head movements as one of the most important predictors of conversational attention. The objective of their research was to investigate whether humans accept a synthetic agent as a genuine conversational partner that is worthy of being attended to in the same way as the human interlocutors. The authors were able to confirm a number of findings about attentive behaviours in human–human conversation. For instance, the subjects spent more time looking at an individual when listening to it than when talking to it – no matter whether the individual was a human or a synthetic agent. Furthermore, the addressee type (human vs. synthetic) did not have any impact on the duration of the speaker’s looking behaviours towards the addressee. While the users’ behaviours in the user-as-speaker condition were consistent with findings for human–human conversation, they noticed differences for the user-as-addressee condition. People spent more time looking at an agent that is addressing them than at a human speaker. Maintaining gaze for an extended period of time is usually considered as rude and impolite. The fact that humans do not conform to social norms of politeness when addressing an agent seems to indicate that they do not regard the agent as an equal conversational partner, but rather as an (somewhat astonishing) artefact that is able to communicate. This attitude towards the agent was also confirmed by the way the users addressed the agent verbally.

8.10.4 Studies Investigating Social Aspects of ECA Dialogue Behaviours

When evaluating the REA system, Bickmore and Cassell [84] found out that social talk may help to build up trust. A positive effect could, however, not be observed across all user groups (introverts vs. extroverts) and interaction styles (face-to-face vs. phone conversations). For instance, introverts perceived the agent significantly more trustworthy in a pure task-based face-to-face conversation than in conversations over the phone or conversations including social talk while extroverts trusted the agent the least in this condition.

Cowel and Stanney [102] observed that portrayed trusting non-verbal behaviours were rated as being significantly more credible than a character portraying no non-verbal behaviour, or one that portrayed non-trusting behaviours. Similar results were obtained by Rehm and André [63] who noticed that agents that fake emotions are perceived as less trustworthy even though the subjects were not able to name reasons

for their uneasiness with the deceptive agent. The results could, however, not be reproduced for scenarios in which the subjects did not fully concentrate on the agent's face but were engaged in a game-like scenario with full-body agents.

8.11 Conclusion

In this chapter we have presented concepts, models and technologies involved in the creation of embodied conversational agents. We have also introduced a design methodology that is used as reference by many researchers when elaborating ECA systems. In order for an ECA to be a user's dialogue companion, it has to share some human-like qualities. These qualities range from being able to display appropriate non-verbal communicative behaviours to having social and emotional intelligence. Some of these qualities are important to ensure that the ECA is perceived as believable and even trustworthy. ECAs have been used in different interactional settings and can take on various roles. We have described, through the presentation of existing systems, how each of these situations requires particular properties for the ECAs.

There is still a lot of work to be done to make an ECA a truly interactive and expressive dialogue companion. In particular, the ECA capabilities in several domains should be enhanced, such as to perceive its environment and its dialogue partners, to recognize and interpret what is being said in a conversation, to display subtle expressions, to be specific in terms of personality, culture, etc. and no more generic, to act like a friend or a tutor, and much more. Real-time is also a real issue. An ECA ought to perceive, plan, adapt, generate and the like in real-time. These are core issues to be dealt with when building an ECA system. Several of the approaches that have been proposed so far are very promising, but much more work needs to be done.

References

1. Noma, T., Zhao, L., Badler, N. I. (2000). Design of a virtual human presenter. *IEEE Comput. Graphics Appl.*, 20, 79–85.
2. André, E., Rist, T., Müller, J. (1999). Employing AI methods to control the behavior of animated interface agents. *Appl. Artif. Intell.*, 13, 415–448.
3. André, E., Concepcion, K., Mani, I., van Guilder, L. (2005). *Autobriefer: A system for authoring narrated briefings*. In: Stock, O., Zancanaro, M., (eds) *Multimodal Intelligent Information Presentation*. Springer, Berlin, 143–158.
4. Weizenbaum, J. (1967). Contextual understanding by computers. *Commun. ACM*, 10, 474–480.
5. Gustafson, J., Lindberg, N., Lundeberg, M. (1999). The August spoken dialog system. In: *Proc. Eurospeech'99*, Budapest, Hungary.
6. Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., Rich, C. (2001). Non-verbal cues for discourse structure. *ACL*, 106–115.
7. Pelachaud, C., Carofiglio, V., Carolis, B. D., de Rosi, F., Poggi, I. (2002). Embodied contextual agent in information delivering application. In: *AAMAS '02: Proc. 1st Int. Joint Conf. on Autonomous Agents and Multiagent Systems*, ACM Press, New York, NY, 758–765.

8. Kopp, S., Jung, B., Leßmann, N., Wachsmuth, I. (2003). Max – A multimodal assistant in virtual reality construction. *Künstliche Intelligenz*, 4(3), 11–17.
9. Wahlster, W. (2003). Towards symmetric multimodality: Fusion and fission of speech, gesture, facial expression. *KI*, 1–18.
10. André, E., Rist, T., van Mulken, S., Klesen, M., Baldes, S. (2000). The automated design of believable dialogues for animated presentation teams. In: Cassell, J., Prevost, S., Sullivan, J., Churchill, E. (eds) *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 220–255.
11. Prendinger, H., Ishizuka, M. (2001). Social role awareness in animated agents. In: *AGENTS '01: Proc. 5th Int. Conf. on Autonomous Agents*, ACM Press, New York, NY, 270–277.
12. Pynadath, D. V., Marsella, S. (2005). Psychsim: Modeling theory of mind with decision-theoretic agents. *IJCAI*, 1181–1186.
13. Rehm, M., André, E., Nischt, M. (2005). Let's come together – Social navigation behaviors of virtual and real humans. *INTETAIN*, 124–133.
14. Traum, D., Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. In: *AAMAS '02: Proc. 1st Int. Joint Conf. on Autonomous Agents and Multiagent Systems*, ACM Press, New York, NY, 766–773.
15. Rickel, J., Johnson, W. L. (1999). Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Appl. Artif. Intell.*, 13, 343–382.
16. Gebhard, P., Kipp, M., Klesen, M., Rist, T. (2003). Authoring scenes for adaptive, interactive performances. In: *AAMAS '03: Proc. 2nd Int. Joint Conf. on Autonomous Agents and Multiagent Systems*, ACM Press, New York, NY, 725–732.
17. Laurel, B. (1993). *Computers as Theatre*. Addison Wesley, Boston, MA, USA.
18. Paiva, A., Dias, J., Sobral, D., Aylett, R., Sobreperez, P., Woods, S., Zoll, C., Hall, L. (2004). Caring for agents and agents that care: Building empathic relations with synthetic agents. In: *AAMAS '04: Proc. 3rd Int. Joint Conf. on Autonomous Agents and Multiagent Systems*, IEEE Computer Society, Washington, DC, USA, 194–201.
19. Isbister, K., Nakanishi, H., Ishida, T., Nass, C. (2000). Helper agent: Designing an assistant for human-human interaction in a virtual meeting space. In: *CHI '00: Proc. SIGCHI Conf. on Human Factors in Computing Systems*, ACM Press, New York, NY, 57–64.
20. Rist, T., André, E., Baldes, S. (2003). A flexible platform for building applications with life-like characters. In: *IUI '03: Proc. 8th Int. Conf. on Intelligent User Interfaces*, ACM Press, New York, NY, 158–168.
21. Cassell, J., Vilhjálmsón, H. H., Bickmore, T. W. (2001). BEAT: the Behavior Expression Animation Toolkit. *SIGGRAPH*, 477–486.
22. Larsson, S., Traum, D. R. (2000). Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Nat. Lang. Eng.*, 6, 323–340.
23. Rich, C., Sidner, C. (1998). Collagen – A collaboration manager for software interface agents. *User Model. User-Adapted Interact.*, 8, 315–350.
24. Rickel, J., Lesh, N., Rich, C., Sidner, C. L., Gertner, A. S. (2002). Collaborative discourse theory as a foundation for tutorial dialogue. *Intell. Tutoring Syst.*, 542–551.
25. Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., Rich, C. (2005). Explorations in engagement for humans and robots. *Artif. Intell.*, 166, 140–164.
26. Jan, D., Traum, D. R. (2005). Dialog simulation for background characters. In: *Int. Conf. on Intelligent Virtual Agents*, Kos, Greece, 65–74.
27. Bales, R. F. (1951). *Interaction Process Analysis*. Chicago University Press, Chicago.
28. Guye-Vuillième, A., Thalmann, D. (2001). A high level architecture for believable social agents. *Virtual Reality J.*, 5, 95–106.
29. Prada, R., Paiva, A. (2005). Intelligent virtual agents in collaborative scenarios. In: *Int. Conf. on Intelligent Virtual Agents*, Kos, Greece, 317–328.
30. Poggi, I. (2003). Mind markers. In: Rector, I., Poggi, N. T. (ed) *Gestures. Meaning and Use*. University Fernando Pessoa Press, Oporto, Portugal.

31. Chovil, N. (1991). Social determinants of facial displays. *J. Nonverbal Behav.*, 15, 141–154.
32. Condon, W., Osgton, W. (1971). Speech and body motion synchrony of the speaker-hearer. In: Horton, D., Jenkins, J. (eds) *The Perception of Language*. Academic Press, New York, NY, 150–184.
33. Kendon, A. (1974). Movement coordination in social interaction: Some examples described. In: Weitz, S. (ed) *Nonverbal Communication*. Oxford University Press, Oxford.
34. Schefflen, A. (1964). The significance of posture in communication systems. *Psychiatry*, 27, 316–331.
35. Ekman, P. (1979). About brows: Emotional and conversational signals. In: von Cranach, M., Foppa, K., Lepenies, W., Ploog, D. (eds) *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*. Cambridge University Press, Cambridge, England; New York, 169–248.
36. Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R. (1996). About the relationship between eyebrow movements and f0-variations. In: *Proc. ICSLP'96: 4th Int. Conf. on Spoken Language Processing*, Philadelphia, PA.
37. Krahmer, E., Swerts, M. (2004). More about brows. In: Ruttkay, Z., Pelachaud, C. (eds) *From Brows till Trust: Evaluating Embodied Conversational Agents*. Kluwer, Dordrecht.
38. McNeill, D. (1992) *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago.
39. Knapp, M., Hall, J. (1997). *Nonverbal Communication in Human Interaction*, Fourth edition. Harcourt Brace, Fort Worth, TX.
40. Pelachaud, C., Bilvi, M. (2003). Computational model of believable conversational agents. In: Huget, M. P. (ed) *Communication in Multiagent Systems*. Volume 2650 of *Lecture Notes in Computer Science*. Springer, Berlin, 300–317.
41. Pelachaud, C. (2005). Multimodal Expressive Embodied Conversational Agent. *ACM Multimedia, Brave New Topics session*, Singapore.
42. DeCarolis, B., Pelachaud, C., Poggi, I., Steedman, M. (2004). APML, a mark-up language for believable behavior generation. In: Prendinger, H., Ishizuka, M. (eds) *Life-Like Characters. Tools, Affective Functions and Applications*. Springer, Berlin, 65–85.
43. Cassell, J., Bickmore, J., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., Yan, H. (1999). Embodiment in conversational interfaces: Rea. *CHI'99*, Pittsburgh, PA, 520–527.
44. Kopp, S., Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *J. Comput. Anim. Virtual Worlds*, 15, 39–52.
45. Kopp, S., Gesellensetter, L., Krämer, N. C. (2005). Wachsmuth, I.: A conversational agent as museum guide – Design and evaluation of a real-world application. In: *Int. Conf. on Intelligent Virtual Agents*, Kos, Greece, 329–343.
46. Heylen, D. (2005). Challenges ahead. Head movements and other social acts in conversation. In: *AISB – Social Presence Cues Symposium*. University of Hertfordshire, Hatfield, England.
47. Ortony, A., Clore, G., Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge.
48. Scherer, K. (2000). Emotion. In: Hewstone, M., Stroebe, W. (eds) *Introduction to Social Psychology: A European Perspective*. Oxford University Press, Oxford, 151–191.
49. Ekman, P. (2003). *The Face Revealed*. Weidenfeld & Nicolson, London.
50. DeCarolis, B., Carofiglio, V., Bilvi, M., Pelachaud, C. (2002). APML, a mark-up language for believable behavior generation. In: *Embodied Conversational Agents – Let's Specify and Evaluate Them!* *Proc. AAMAS'02 Workshop*, Bologna, Italy.
51. Ball, G., Breese, J. (2000). Emotion and personality in a conversational agent. In: Cassell, J., Sullivan, S. P., Churchill, E. (eds) *Embodied Conversational Characters*. MIT Press, Cambridge, MA, 189–219.
52. Tanguy, E., Bryson, J. J., Willis, P. J. (2006). A dynamic emotion representation model within a facial animation system. *Int. J. Humanoid Robotics*, 3, 293–300.

53. Pandzic, I., Forchheimer, R. (2002). *MPEG4 Facial Animation – The Standard, Implementations and Applications*. Wiley, New York, NY.
54. deRosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., Carolis, B. D. (2003). From Greta’s mind to her face: Modelling the dynamics of affective states in a conversational embodied agent. *Int. J. Hum. Comput. Studies*, Special Issue on Applications of Affective Computing in HCI, 59, 81–118.
55. Bui, T. D. (2004). *Creating emotions and facial expressions for embodied agents*. PhD thesis, University of Twente, Department of Computer Science, Enschede.
56. Tsapatsoulis, N., Raouzaïou, A., Kollias, S., Cowie, R., Douglas-Cowie, E. (2002). Emotion recognition and synthesis based on MPEG-4 FAPs in MPEG-4 facial animation. In: Pandzic, I. S., Forcheimer, R. (eds) *MPEG4 Facial Animation – The Standard, Implementations and Applications*. Wiley, New York, NY.
57. Albrecht, I., Schroeder, M., Haber, J., Seidel, H. P. (2005). Mixed feelings – expression of nonbasic emotions in a muscle-based talking head. *Virtual Reality – Special Issue on Language, Speech and Gesture for VR*, 8(4).
58. Whissel, C. M. (1989). The dictionary of affect in language. In: Plutchnik, R., Kellerman, H. (eds) *The measurement of Emotions. Volume Emotion: Theory, Research and Experience: Vol. 4*. Academic Press, New York.
59. Plutchnik, R. (1980). *Emotion: A Psychoevolutionary Synthesis*. Harper and Row, New York, NY.
60. Ruttkay, Z., Noot, H., ten Hagen, P. (2003). Emotion disc and emotion squares: Tools to explore the facial expression face. *Comput. Graph. Forum*, 22, 49–53.
61. Schlosberg, H. A. (1952). A description of facial expressions in terms of two dimensions. *J. Exp. Psychol.*, 44, 229–237.
62. Ekman, P., Friesen, W. (1975). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Prentice-Hall, Inc, Englewood Cliffs, NJ.
63. Rehm, M., André, E. (2005). Catch me if you can: Exploring lying agents in social settings. *AAMAS, Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Utrecht, Netherlands, ACM: New York, USA, 937–944.
64. Ochs, M., Niewiadomski, R., Pelachaud, C., Sadek, D. (2005). Intelligent expressions of emotions. In: *1st Int. Conf. on Affective Computing and Intelligent Interaction ACII*, China.
65. Martin, J. C., Niewiadomski, R., Devillers, L., Buisine, S., Pelachaud, C. (2006). Multimodal complex emotions: Gesture expressivity and blended facial expressions. *Int. J. Humanoid Robotics. Special issue on “Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids”*, 3(3).
66. Wehrle, T., Kaiser, S., Schmidt, S., Scherer, K. R. (2000). Studying the dynamics of emotional expression using synthesized facial muscle movements. *J. Pers. Social Psychol.*, 78, 105–119.
67. Kaiser, S., Wehrle, T. (2006). Modeling appraisal theory of emotion and facial expression. In: Magnenat-Thalmann, N. (ed) *Proc. 19th Int. Conf. on Computer Animation and Social Agents*, CASA 2006, Geneva, Computer Graphics Society (CGS).
68. Wehrle, T. (1996). *The Geneva Appraisal Manipulation Environment (GAME)*. University of Geneva, Switzerland. Unpublished computer software edn.
69. Perlin, K., Goldberg, A. (1996). Improv: A system for interactive actors in virtual worlds. In: *Computer Graphics Proc., Annual Conference Series, ACM SIGGRAPH*, New Orleans, Louisiana, USA, 205–216.
70. Bruderlin, A., Williams, L. (1995). Motion signal processing. In: *Proc. 22nd Annual Conf. on Computer Graphics and Interactive Techniques*, ACM Press, New York, NY, 97–104.
71. Chi, D. M., Costa, M., Zhao, L., Badler, N. I. (2000). The EMOTE model for effort and shape. In: Akeley, K. (ed) *Siggraph 2000, Computer Graphics Proc.*, ACM Press/ACM SIGGRAPH/Addison Wesley Longman, 173–182.

72. Laban, R., Lawrence, F. (1974). *Effort: Economy in Body Movement*. Plays, Inc., Boston.
73. Wallbott, H. G., Scherer, K. R. (1986). Cues and channels in emotion recognition. *J. Pers. Soc. Psychol.*, 51, 690–699.
74. Gallaher, P. E. (1992). Individual differences in nonverbal behavior: Dimensions of style. *J. Pers. Soc. Psychol.*, 63, 133–145.
75. Hartmann, B., Mancini, M., Pelachaud, C. (2005). Implementing expressive gesture synthesis for embodied conversational agents. In: *Gesture Workshop*, Vannes.
76. Egges, A., Magnenat-Thalmann, N. (2005). Emotional communicative body animation for multiple characters. In: *V-Crowds'05*, Lausanne, Switzerland, 31–40.
77. Stocky, T., Cassell, J. (2002). Shared reality: Spatial intelligence in intuitive user interfaces. In: *IUI '02: Proc. 7th Int. Conf. on Intelligent User Interfaces*, ACM Press, New York, NY, 224–225.
78. Chopra-Khullar, S., Badler, N. I. (2001). Where to look? automating attending behaviors of virtual human characters. *Autonomous Agents Multi-Agent Syst.*, 4, 9–23.
79. Nakano, Y. I., Reinstein, G., Stocky, T., Cassell, J. (2003). Towards a model of face-to-face grounding. *ACL'03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Sapporo, Japan, 553–561.
80. Peters, C. (2005). Direction of attention perception for conversation initiation in virtual environments. In: *Int. Conf. on Intelligent Virtual Agents*, Kos, Greece, 215–228.
81. Baron-Cohen, S. (1994). How to build a baby that can read minds: Cognitive Mechanisms in Mind-Reading. *Cah. Psychol. Cogn.*, 13, 513–552.
82. Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., Fischer, K. (2005). The recognition of emotion. In: Wahlster, W. (ed) *Verbmobil: Foundations of Speech-to-Speech Translations*. Springer, Berlin, 122–130.
83. Maatman, R. M., Gratch, J., Marsella, S. (2005). Natural behavior of a listening agent. In: *Int. Conf. on Intelligent Virtual Agents*, Kos, Greece, 25–36.
84. Bickmore, T., Cassel, J. (2005). Social dialogue with embodied conversational agents. In: van Kuppevelt, J., Dybkjaer, L., Bernsen, N. O. (eds) *Advances in Natural, Multimodal Dialogue Systems*. Springer, Berlin.
85. Brown, P., Levinson, S. C. (1987). *Politeness – Some Universals in Language Usage*. Cambridge University Press, Cambridge.
86. Walker, M. A., Cahn, J. E., Whittaker, S. J. (1997). Improvising linguistic style: Social and affective bases for agents, *First International Conference on Autonomous Agents*, Marina del Rey, CA, USA, ACM: New York, USA, 96–105.
87. Johnson, W. L., Rizzo, P., Bosma, W., Kole, S., Ghijsen, M., vanWelbergen, H. (2004). Generating socially appropriate tutorial dialog. *Affective Dialogue Systems, Tutorial and Research Workshop, ADS 2004*, Kloster Irsee, Germany, June 14–16, 2004, Springer, Lecture Notes in Computer Science, Vol. 3068, 254–264.
88. Johnson, L., Mayer, R., André, E., Rehm, M. (2005). Cross-cultural evaluation of politeness in tactics for pedagogical agents. In: *Proc. of the 12th Int. Conf. on Artificial Intelligence in Education (AIED)*, Amsterdam, Netherlands.
89. Rehm, M., André, E. (2006). Informing the design of embodied conversational agents by analysing multimodal politeness behaviours in human-human communication. In: Nishida, T. (ed) *Engineering Approaches to Conversational Informatics*. Wiley, Chichester, UK.
90. Cassell, J. (2006). Body language: Lessons from the near-human. In: Riskin, J. (ed) *The Sistine Gap: History and Philosophy of Artificial Intelligence*. University of Chicago, Chicago.
91. Martin, J. C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., Pelachaud, C. (2005). Levels of representation in the annotation of emotion for the specification of expressivity in ECAs. In: *Int. Conf. on Intelligent Virtual Agents*, Kos, Greece, 405–417.
92. Kipp, M. (2005). *Gesture generation by imitation: from human behavior to computer character animation*. Dissertation.com, Boca Raton, FL.

93. Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., Bregler, C. (2004): Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Trans. Graph.* 23, 506–513.
94. Buisine, S., Abrilian, S., Niewiadomski, R., MARTIN, J. C., Devillers, L., Pelachaud, C. (2006). Perception of blended emotions: From video corpus to expressive agent. In: *The 6th Int. Conf. on Intelligent Virtual Agents*, Marina del Rey, USA.
95. Ruttkay, Z., Pelachaud, C. (2004). *From Brows to Trust: Evaluating Embodied Conversational Agents (Human-Computer Interaction Series)*. Springer-Verlag, New York, Inc., Secaucus, NJ, USA.
96. Buisine, S., Abrilian, S., Martin, J. C. (2004). Evaluation of multimodal behaviour of embodied agents. In: Ruttkay, Z., Pelachaud, C. (eds) *From Brows to Trust: Evaluating Embodied Conversational Agents*. Kluwer, Norwell, MA, 217–238.
97. Lee, K. M., Nass, C. (2003). Designing social presence of social actors in human computer interaction. In: *CHI '03: Proc. SIGCHI Conf. on Human Factors in Computing Systems*, ACM Press, New York, NY, 289–296.
98. Nass, C., Gong, L. (2000). Speech interfaces from an evolutionary perspective. *Commun. ACM*, 43, 36–43.
99. Vinayagamoorthy, V., Garau, M., Steed, A., Slater, M. (2004). An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience. *Comput. Graph. Forum*, 23, 1–12.
100. Lee, S. P., Badler, J. B., Badler, N. I. (2002). Eyes alive. In: *SIGGRAPH '02: Proc. 29th Annual Conf. on Computer Graphics and Interactive Techniques*, ACM Press, New York, NY, 637–644.
101. Rehm, M., André, E. (2005). Where do they look? Gaze behaviors of multiple users interacting with an embodied conversational agent. In: *Int. Conf. on Intelligent Virtual Agents*, Kos, Greece, 241–252.
102. Cowell, A. J., Stanney, K. M. (2003) Embodiment and interaction guidelines for designing credible, trustworthy embodied conversational agents. In: *Int. Conf. on Intelligent Virtual Agents*, Kos, Greece, 301–309.

Chapter 9

Multimodal Information Processing for Affective Computing

Jianhua Tao

9.1 Introduction

Affective computing is interaction *that relates to, arises from or deliberately influences emotions* [1]; it tries to assign computers the human-like capabilities of observation, interpretation and generation of affect features. It is an important topic in human–computer interaction (HCI), because it helps increase the quality of human to computer communications.

The research on emotion has a long history and can be traced back to the 19th century [2]. Some researchers feel there are a few relatively unambiguous “basic” emotions, for example, anger and happiness, with more subtle ones “overlaid” on top of these or comprising blends. Other researchers see emotions as zones in emotional space [3]. The vector might be a cline of emotions shading into one another, with extremes at either end. In the vector approach, the degree of expression would simply be a reflection of intensity in a particular zone. The vector represents the potential for n emotion “zones” along the “emotion vector” with the relative sizes of the zones along the vector depicting the intensity of the specified emotion.

Traditionally, “affect” was seldom linked to lifeless machines. It is quite new to use affective features in HCI. Among them, multimodal affective information is most important for natural computer interface [4]; more and more researchers are interested in how to process the multimodal affective information for HCI, which has become known as “affective computing” [1]. The affective computing builds an “affect model” based on a variety of information, which results in a personalized computing system with the capability of perception, interpretation of human feelings as well as generating intelligent, sensitive and friendly responses.

The chapter briefly summaries some key technologies which have been developed in recent years, such as emotional speech processing, facial expression, body gesture and movement, affective multimodal system and affect understanding and generating. Further, the chapter also introduces some related projects, which places

J. Tao (✉)
Chinese Academy of Sciences, Beijing, China
e-mail: jhtao@nlpr.ia.ac.cn

current and past research work and applications in context. Finally, the chapter discusses some key research topics which comprise a large challenge in current research work.

9.2 Multimodal-Based Affective Human–Computer Interaction

Affective HCI consists of affect states modelling, recognition and understanding and expression. As we know, people express affects through a series of actions relating to facial expression, body movements, gestures, voice behaviour and other physiological signals, such as heart rate and sweat, etc. The following reviews the most active key technologies in these areas, for example, emotional speech processing, facial expression recognition and generating, body gesture and movement, multimodal system, affect understanding and generating.

9.2.1 Emotional Speech Processing

For emotional speech processing, it is widely known that emotional speech differs with respect to the acoustic features [5–8]. Some prosody features, such as pitch variables (F0 level, range, contour and jitter) and speaking rate have been analysed by researchers [9]. Parameters describing laryngeal processes on voice quality have also been taken into account [10]. Tato [11] carried out some experiments which showed how “quality features” are used in addition to “prosody features”.

These acoustic features are widely used in the research of emotion recognition; they have commonly used pattern recognition methods. For instance, Dellaert [12] used prosody features and compared three classifiers: the maximum likelihood Bayes classification, kernel regression and k-nearest neighbour in emotion recognition for sadness, anger, happiness and fear. Petrushin [9] employed vocal parameters and a computer agent for emotion recognition. Lee [13] used linear discriminant classification with Gaussian class-conditional probability distribution and k-nearest neighbourhood methods to classify utterances into two basic emotion states, namely, negative and non-negative. Yu [14] used support vector machines (SVMs) for emotion detection. An average accuracy of 73% for emotion classification was reported by him. Nick [15] proposed a perception model of affect speech utterances and showed that there are consistent differences in the acoustic features of same-word utterances that are perceived as having different discourse effects or displaying different affective states. This work suggested that rather than selecting one emotion to describe an utterance, a vector of activations across a range of features may be more appropriate.

For emotion generating with speech synthesis, Cahn [16] and Schroeder [17], by using acoustic parameters controlling rules, achieved outputting emotional speech with manual inferences (see Table 9.1). Recently, some efforts have been concerned with the idea of the large corpus. A typical system was created by Campbell [18],

Table 9.1 Some rules for emotional speech generation, concluded by Schroeder [17]

Emotions	Parameter settings
Joy German	<i>F0 mean</i> : +50% <i>F0 range</i> : +100% <i>Tempo</i> : +30% <i>Voice Quality</i> : modal or tense; “lip-spreading Feature”: F1/F2 +10% <i>Other</i> : “wave pitch contour model”: main stressed syllables are raised (+100%), syllables in between are lowered (-20%)
Sadness American English	<i>F0 mean</i> : “0”, reference line “-1”, less final lowering “-5” <i>F0 range</i> : “-5”, steeper accent shape “+6” <i>Tempo</i> : “-10”, more fluent pauses “+5”, hesitation pauses “+10” <i>Loudness</i> : “-5” <i>Voice Quality</i> : breathiness “+10”, brilliance “-9” <i>Other</i> : stress frequency “+1”, precision of articulation “-5”
Anger British English	<i>F0 mean</i> : +10 Hz <i>F0 range</i> : +9 s.t. <i>Tempo</i> : +30 wpm <i>Loudness</i> : +6 dB <i>Voice Quality</i> : laryngealization +78%; F4 frequency -175 Hz <i>Other</i> : increase pitch of stressed vowels (2ary: +10% of pitch range; 1ary: +20%; emphatic: +40%)
Fear German	<i>F0 mean</i> : “+150%” <i>F0 range</i> : “+20%” <i>Tempo</i> : “+30%” <i>Voice Quality</i> : falsetto

who generated an expressive speech synthesis with a large corpus which gave impressive results. The system could generate human-like spontaneous speech; however, the content is limited by the corpus. Schroeder [19], Eide [20] generated expressive Text to Speech (TTS) engines which can be directed, via an extended Speech Synthesis Markup Language (SSML), to use a variety of expressive styles with about 10 hours of “neutral” sentences. Optionally, rules translating certain expressive elements to ToBI markup are manually derived (see Fig. 9.1). Chuang [21] and Tao [22] used emotional keywords and emotion trigger words to generate the emotional TTS system. The final emotion state is determined based on the emotion outputs from a textual content module. The results were used in dialogue systems to improve the naturalness and expressiveness of the answering speech.

Most of the research on emotional speech is still focused on typical acoustic feature analysis in different languages. Some work in emotion classification systems and rule-based emotional speech synthesis systems has been done [23]; however, the lack of work on the acquirement and the analysis of more detailed and reliable

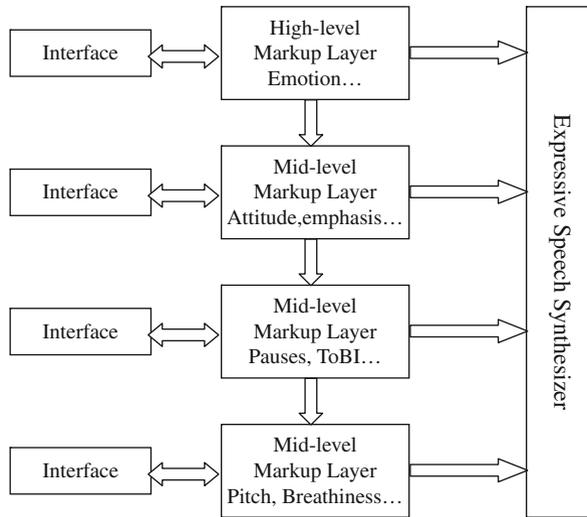


Fig. 9.1 Emotional Text to speech via Speech Synthesis Markup Language (SSML) [20]

physiological features limit further development of the work. People express their feelings not only via the acoustic features, but also with the content of what they say. Different words, phrases and syntactic structures can result in lots of different kinds of expression results and styles. Although some researches on emotional keyword detection [16, 22] and cognitive linguistics have been carried out [24], further work is still needed for the integration of these two research topics, such as the relation between emotions and semantics, etc.

9.2.2 Affect in Facial Expression

Facial expressions and movements, such as a smile or a nod, are used either to fulfil a semantic function or to communicate emotions or conversational cues. Charles Darwin wrote that facial expressions of emotions are universal, not learned differently in each culture [25]. Since then, several researchers have experimentally observed that this was true for some emotions (happiness, sadness, surprise, fear, anger, disgust), although there are cultural differences in when these expressions are shown (elicitation and display rules). Similar to speech processing, the research of facial expression consists of work on coding, recognition and generation, and have been carried out for a very long time. For instance, Etcoff [26] parameterized the structure of the major parts of human's face through a series of 37 lines; this enables people to tell the affect status of faces. Ekman [27] (http://www.face-and-emotion.com/dataface/facs/new_version.jsp) built a facial action coding system. He classified human's facial expressions into many action units. With this method, he described facial expressions with six basic emotions: joy, anger, surprise, disgust,

fear and sadness. Currently, most of the facial features can be found from the definition of MPEG-4 (Moving Picture Experts Group-4). MPEG-4 allows the user to configure and build systems for many applications by allowing flexibility in the system configurations by providing various levels of interactivity with audio-visual content [24, 28]. In this standard, both mesh [29] or muscle model are used to create three-dimension (3-D) facial models. We call these two models (mesh model and muscle model) parametrical models which generate the realistic 3-D face modelling with three important aspects: (1) recovery of geometric shapes; (2) acquisition of surface textures; (3) recovery of skin reflectance properties. The face animation parameters (FAPs) are used to control the model for the expression generation.

To do the facial expression analysis, Lyons applied the supervised Fisher linear discriminant analysis (FDA) [30] to learn an adequate linear subspace from class-specified training samples and the samples projected to this subspace can be best separated for the facial expression classification. Similar work has also been done by using principal component analysis (PCA) [31] and independent component analysis (ICA) (<http://www.cs.bham.ac.uk/%7Eaxs/cogaff.html>). Furthermore, there are many other methods, such as Gabor wavelets [30], Neural Networks [32], Hidden Markov Models (HMM) [28], Point Distribute Model (PDM), Optical Flow, Geometrical Tracking method, Elastic Graphs Matching (EGM) method. Among them, the Gabor model has been favoured by many researchers due to its good performance and lack of sensitivity to face posture and the noise of the background [30]. However, the average emotion recognition rate from facial expression is still ranged from 60% to 80%.

The pioneering work on facial animation of Frederic I. Parke in the 1970s showed a primary parameterized model for the facial animation. Over the last decade, the quality of facial animations has been improved considerably due to the rapid development of computer technologies. But, the generation of life-like animated faces still remains an open issue. Many researchers use methods based on images [33, 34], visemes [35], Facial Animation Parameters (FAPs, see Fig. 9.2) [5], Principal Components (PCs) [36, 37], Three-Dimension (3-D) coordinates [35], 3-D distance measurements [28, 36] or Optical Flows [38] to generate facial expression.

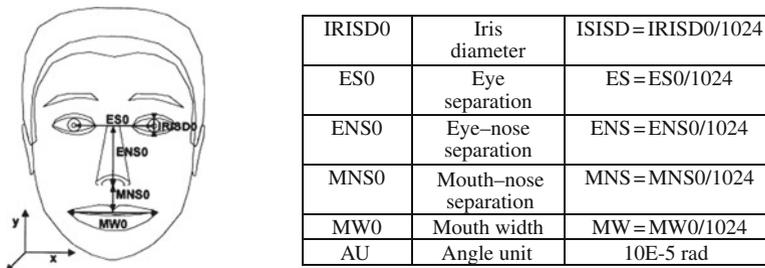


Fig. 9.2 The definition of Facial Animation Parameters (FAPs) (defined by MPEG-4) [29]

Normally, the facial expression should be synchronized with speech. While people are talking with various emotions, audio-visual mapping is always the key component to generate a vivid facial animation system. In general, there are two ways to do this mapping: using speech recognition as a bridge or driving the face expression from audio directly. The first approach uses the speech recognition engine to recognize the speech into phoneme strings, and uses the mapping model from phonemes to lip movements to generate the face expression. A typical work was finished by Yamamoto [28], who used Hidden Markov Models (HMMs) to recognize the speech into phonemes, and mapped them directly to corresponding lip shapes. The final lip movement sequence is obtained by using a smoothing algorithm. The method is simple and easy to be implemented. But it can only generate the lips moving. The second approach tries to find direct mapping models from audio to visual and drive the face animation directly from audio acoustic features. For instance, Massaro [24] trained Artificial Neural Network (ANN) to learn the mapping from Linear Predictive Coding (LPC) to face animation parameters. Many other methods have also been tried, such as, Time Delay Neural Network (TDNN) [24], Meridian Lossless Packing (MLP) [36], K-Nearest Neighbours (KNN) [36], Hidden Markov Models (HMM) [28], Gaussian Mixture Model (GMM), Vector Quantization (VQ) [28], Rule-based Model [35, 38], video-based unit selection method [34]. Till now, the lip movement is still the focus in most of the research. Full facial expression, especially the correlation between facial expression and more acoustic features, such as prosody, timbre, etc. has seldom been touched. However, even with this limitation, the facial animation technologies have been widely used in movie making, games design and virtual realities, etc.

9.2.3 Affective Multimodal System

Emotions are expressed in several modalities [39]. In face-to-face communication, the emotional state of the speaker is transmitted to the listener through a multi-channel process that involves both the verbal and the non-verbal modalities of communication. Human communicative modalities include speech, facial expressions, hand gestures (and other bodily movements), gaze (and pupil dilation), posture, spatial behaviour, bodily contact, clothes (and other aspects of appearance), non-verbal vocalizations [40]. There is a wide literature on non-verbal communication [40–43]. Among them, audio-visual emotion processing is the most important. Most of the existing methods for audio-visual emotion analysis are based on deliberately posed emotion displays [44–51]. Recently a few exceptional studies have been reported in spontaneous emotion displays [52–55] used the data collected in psychological research interview. For instance, Pal et al. [53] used recordings of infant affective displays, while Fragopanagos and Taylor [52], Caridakis et al. [54] and Karpouzis et al. [55] used the data collected in Wizard of OZ scenarios.

For emotion recognition, assuming that audio and visual channels play different roles in human perception of different emotions, the decision-level fusion method can correctly integrate audio and video data using a weighting matrix to fuse the

results of the two classifiers. For instance, Silva et al. [75] subjectively determined the weighting matrix at the decision level. Based on the idea that different modalities dominate the recognition of different emotions, [75] took the advantages of assigning the modality-adaptive weights to achieve efficient emotion recognition. Until now, however, all weights for different modalities were only decided by rules in [75], and the performance of the method was limited.

On the other hand, the feature-level fusion method is closely related to the human processing of emotion information. Fusing at the stage of feature extraction, the method is more general than the decision-level fusion method. In addition, it can also take great advantage of statistical learning methods developed recently. A typical work was reported by Chen et al. [56], in which the authors used the support vector machine (SVM) to obtain better performance in classifying emotions. However, the different time scales and metric levels of features coming from different modalities, as well as increasing feature-vector dimensions influence the performance of an emotion recognizer based on a feature-level fusion.

Till now, multimodal emotion processing has been widely used for smart room, virtual reality, etc. Among them, the ubiquitous computing (<http://www.ubiq.com/hypertext/weiser/UbiHome.html>) might be the most representative application, which encompasses a wide range of research topics, including mobile computing, sensor networks and artificial intelligence.

9.2.4 *Affective Understanding*

Affective understanding tries to make the computer understand the user's behaviours, moods and affective states, and be able to generate appropriate responses. [57]

Although there is much work to be done in this area, the OCC (Ortony, Clore & Collins, 1988) [58] model has been proved to be the most successful model in affect sensing and reaction. It classifies people's emotions as a result of events, objects and other agents under three categories. People are happy or unhappy with an event, like or dislike an object, approve or disapprove an agent. There are 22 detailed emotions under the three emotion categories in OCC model. Although the model provides three groups of emotions depending on reactions to external things, it is really hard to do that in real-time environments, where reactions are more complicated. Some reactions may involve all of the emotions in the three categories. More recent works went beyond this classification, to consider categories of emotions that occur frequently in human-computer communication. [59].

The OCC model has established itself as the standard model for emotion cognition. A large number of studies employed the OCC model to generate emotions for their embodied characters. Many developers of such characters believe that the OCC model will be all they ever need to equip their character with emotions. It has been now widely used for modelling emotions in natural language processing, user reaction of HCI, dialogue systems, etc.

9.3 Projects and Applications

The origins of the field trace back to Picard's 1997 seminal book on Affective Computing [1]. Although it is a relatively new concept, there are already some related projects or applications. A brief summary of some of them follows.

9.3.1 *Affective-Cognitive for Learning and Decision Making*

This work was developed at MIT Affective Computing research group [1]. It aims to redress many of the classic problems that most machine-learning and decision-making models encounter that are based on old, purely cognitive models and are slow, brittle and awkward to adapt, by developing new models that integrate affect with cognition. Ultimately, such improvements will allow machines to make more human-like decisions for better human-machine interactions.

9.3.2 *Affective Robot*

Several universities have tried integrating affect sensing and reaction into the robot research. The results are very successful. For instance, Carnegie Mellon University (CMU) developed a service robot which can detect different emotions and give prompt reaction with human. The initial domain for this work is used in hospitals to take care of patients. In Vanderbilt University a similar project is carried out which involves developing a novel affect-sensitive architecture for human-robot cooperation. In the project, a robot which can monitor physiological signals of human by using wearable sensors has been developed. The signals are analysed in real-time to infer the emotional states of the human who is communicating with the robot.

9.3.3 *Oz*

Oz is a computer system that allows authors to create and present interactive dramas, and was developed at Carnegie Mellon University (CMU) (<http://www.cs.cmu.edu/afs/cs.cmu.edu/project/oz/web/oz.html>). In the project, the emotional state and story context are used to provide subjective descriptions of sensory information and to generate a stream of thought. The purpose of such description is to help a user gain a deeper understanding of the role they play in the story world and help artists create high-quality interactive drama, based in part on AI technologies. A series of good results had been achieved when the project was finished in 2002.

9.3.4 Affective Facial and Vocal Expression

The work studies the relation between subjective experience (feeling), appraisal and subtle facial and vocal expressions. For instance, in the Geneva Emotion Research Group (<http://www.unige.ch/fapse/emotion/>), they have tested if the recognition of emotional expressions is improved by presenting animated facial expressions as compared to static stimuli. In UIUC, the team of Huang [47, 49] explored recognition of audio-visual spontaneous affective expressions occurring in a psychological interview named “The Adult Attachment Interview (AAI)” that is used to characterize individuals’ current state of mind with respect to past parent–child experiences.

9.3.5 Affective Face-to-Face Communication

The work is being carried out in the COST Action “Cross-Modal Analysis of Verbal and Non-verbal Communication” [60]. More than 20 European universities and institutes are involved. The main objective of the project is to develop an advanced acoustical, perceptual and psychological analysis of verbal and non-verbal communication signals originating in spontaneous face-to-face interaction in order to identify algorithms and automatic procedures capable of identifying the human emotional states. Several key aspects are considered: for example, the integration of the developed algorithms and procedures for application in telecommunication, and the recognition of emotional states, gestures, speech and facial expressions, in anticipation of the implementation of intelligent avatars and interactive dialogue systems that could be exploited to improve user access to future telecommunication services.

9.3.6 Humaine

HUMAINE (Human–Machine Interaction Network on Emotion) is a Network of Excellence in the EU’s Sixth Framework Programme (<http://emotion-research.net/>). The project aims to lay the foundations for European development of systems that can register, model and influence human emotional and emotion-related states and processes, that is, “emotion-oriented systems”. HUMAINE brings together a number of experts from the key disciplines in a programme designed to achieve intellectual integration. It identifies six thematic areas that cut across traditional groupings and offers a framework for an appropriate division: theory of emotion; signal/sign interfaces; the structure of emotionally coloured interactions; emotion in cognition and action; emotion in communication and persuasion; and usability of emotion-oriented systems. From 2007, HUMAINE has become an international association which includes more than 100 initial members.

9.4 Research Challenges

On the basis of perception, analysis and modelling of affective features, the interrelation among research contents can be illustrated as shown in Fig. 9.3. The framework involves several modules on information acquiring, recognition, understanding, generation and expression. Affective states are used in order to transfer status for the interaction. Some use agents to describe them while others use a markup language to represent them, such as eXtended Markup Language (XML) based emotion language, etc.

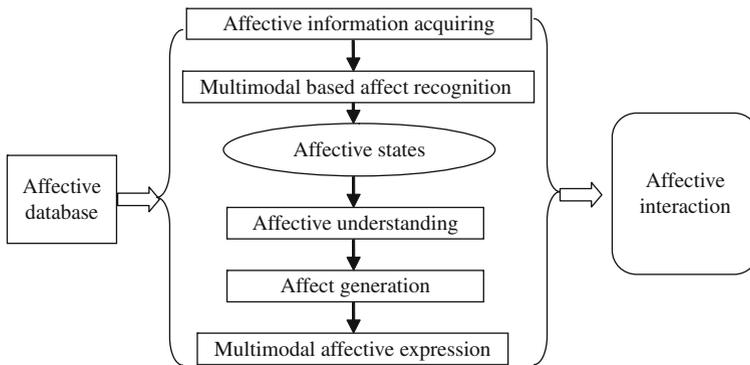


Fig. 9.3 Multimodal affective interaction system

With reference to this figure, some challenging research topics can be seen.

9.4.1 Cognitive Structure of Affects

Most of the existing affect models are based on discrete emotion states, which do not have significant correspondence to the real environment. There are lots of arguments about how best to define emotions. Some individuals think it is not possible to model affect and, thus, facilitate affective understanding. This question has been discussed by Picard in her paper [1]. “With any complex modelling problem, there are ways to handle complexity and contribute to progress in understanding. If we want to model affect at the neuropeptide signalling level, then we are limited indeed, because we know only a small part about how such molecules communicate among the organs in our body and real-time sensing of local molecular activity is not easily accomplished” [1].

With the affect model, the ultimate purpose of affective computing is to assist the computer to react accurately after it understands the user’s affect and meaning, and to be accustomed to the changes of the user’s affect. Although there is some work using the human-aided method to evaluate the user’s feelings, it is still an important issue on how to analyse the dynamic characteristics of the user’s affect and how to make the computer react according to the identification result of

affective information. Affect is closely associated with personalities, environment and cultural background; a good affect model can only be realized by combining all this information. Psychological research indicates that affect could be extended from past affect states. Moreover, the lack of dynamic affect information mechanism is another important factor restricting current affect models. Therefore, how to define and integrate this information, how to describe and integrate the dynamic affect information and how to improve the adaptation algorithm to natural scenarios will comprise the emphasis in future research.

9.4.2 Multimodal-Based Affective Information Processing

With reference to the analysis described in Section 9.2.4, the lack of the integration mechanism of affective parameters under multimodal conditions limits the affective understanding and the affective expression. The fusing of different channels involves not just the combination of them, but also finding the mutual relations among them. The mutual relation could make better integration of the different channels in interaction for both recognition and information generation as shown in Fig. 9.4.

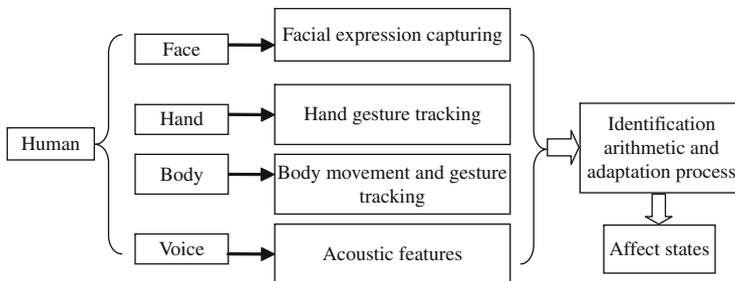


Fig. 9.4 Multimodal-based affective recognition

9.4.3 Affective Features Capturing in Real Environments

Most of current affective features capturing is still limited to laboratories or studios, which are less complicated and have less background noise than the real world. Currently available information can only be used in retrieval and common feature identification, which is not robust enough to make affective computing in real applications. Apart from developing high-quality affective interaction, there needs to be an emphasis on establishing automatic (with high robust) affective information capture within real-time environments and a need to gain more reliable descriptions of them, especially, for facial expression tracking, high robust hand/body gesture tracking and the modelling of more reliable automatic acoustic parameters capturing.

9.4.4 Affective Interaction in Multi-agent Systems

The study of agent-based systems evolved from the field of Distributed Artificial Intelligence in the early 1980s and has been given new impetus by the emergence of the Internet. Solving the problems associated with intelligence and taking advantage of the opportunities offered, this inherently distributed and unstructured environment is seen as a major application area for intelligent and multi-agent systems. Traditional affective interaction has been based on single HCI procedures. It is a challenge to facilitate affective interactions in multi-agent systems. In contrast to classical applications in Artificial Intelligence, the central ideas underlying multi-agent-based affective interaction are that

- the affect of one agent could be influenced by the other agents;
- the system exhibits goal directed behaviour;
- one agent can interact with and negotiate with other agents (possibly human) in order to achieve their goals;
- the whole system can apply intelligence in the way they react to a dynamic and unpredictable environment.

Apart from the implementation of practical and useful systems, another primary goal in the study of multi-agent-based affective interaction systems is to understand interactions among intelligent entities whether they are computational, human or both.

9.5 Conclusion

Although the research of affective computing is relatively new, it has attracted extensive attention from researchers around the world. Existing research is mainly limited to the emotion recognition or expression in separated channels such as voice, facial and body gestures. Due to the lack of an effective integration of multimodal features, the computer is not able to better recognize affect states of human, clearly understand the human meaning and make better human-like responses in human-machine interaction system. Various problems are not well solved; however, there is considerable scope for applications: for instances, adding the function of automatic human mood perception in household appliances to provide better services to people; making use of the affect analysis in information retrieval system to improve the accuracy and efficiency of the information retrieval; adding affect factors in remote education systems; processing human emotions in virtual reality applications to enhance the virtual scenario, etc. In addition, affective computing may also be applied in games, robots and intelligent toys to create objects with more personalities. The technology will give us a remarkable change in our research of human-computer interactions.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 60575032 and the 863 program under Grant 2006AA01Z138.

References

1. Picard R W. (1997). *Affective Computing*. MIT Press, Cambridge, MA.
2. James W. (1884). What is emotion? *Mind*, vol. 9(34), 188–205.
3. Oatley K. (1987). Cognitive science and the understanding of emotions. *Cogn. Emotion*, 3(1), 209–216.
4. Bigun E. S., Bigun J., Duc B., Fischer S. (1997). Expert conciliation for multimodal person authentication systems using bayesian statistics. In: *Int. Conf. on Audio and Video-Based Biometric Person Authentication (AVBPA)*, Crans-Montana, Switzerland, 291–300.
5. Scherer K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychol. Bull.*, vol. 99(2), 143–165.
6. Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Commun.*, 40, 227–256.
7. Scherer, K. R., Banse, R., Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *J. Cross-Cultural Psychol.*, 32 (1), 76–92.
8. Johnstone, T., van Reekum, C. M., Scherer, K. R. (2001). Vocal correlates of appraisal processes. In: Scherer, K. R., Schorr, A., Johnstone, T. (eds) *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, New York and Oxford, 271–284.
9. Petrushin, V. A. (2000). Emotion recognition in speech signal: Experimental study, development and application. In: *6th Int. Conf. on Spoken Language Processing, ICSLP2000*, Beijing, 222–225.
10. Gobl, C., Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Commun.*, 40(1-2), 189–212.
11. Tato, R., Santos, R., Kompe, R., Pardo, J. M. (2002). Emotional space improves emotion recognition. In: *ICSLP2002*, Denver, CO, 2029–2032.
12. Dellaert, F., Polzin, T., Waibel, A. (1996). Recognizing emotion in speech. In: *ICSLP 1996*, Philadelphia, PA, 1970–1973.
13. Lee, C. M., Narayanan, S., Pieraccini, R. (2001). Recognition of negative emotion in the human speech signals. In: *Workshop on Automatic Speech Recognition and Understanding*.
14. Yu, F., Chang, E., Xu, Y. Q., Shum H. Y. (2001). Emotion detection from speech to enrich multimedia content. In: *The 2nd IEEE Pacific-Rim Conf. on Multimedia*, Beijing, China, 550–557.
15. Campbell, N. (2004). Perception of affect in speech – towards an automatic processing of paralinguistic information in spoken conversation. In: *ICSLP2004*, Jeju, 881–884.
16. Cahn, J. E. (1990). The generation of affect in synthesized speech. *J. Am. Voice I/O Soc.*, vol. 8, 1–19.
17. Schroder, M. (2001). Emotional speech synthesis: A review. In: *Eurospeech 2001*, Aalborg, Denmark, 561–564.
18. Campbell, N. (2004). Synthesis units for conversational speech – using phrasal segments. *Autumn Meet. Acoust.: Soc. Jpn.*, vol. 2005, 337–338.
19. Schroder, M., Breuer, S. (2004). XML representation languages as a way of interconnecting TTS modules. In: *8th Int. Conf. on Spoken Language Processing, ICSLP'04*, Jeju, Korea.
20. Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., Pitrelli, J. (2002). A corpus-based approach to <ahem/> expressive speech synthesis. In: *IEEE Speech Synthesis Workshop*, Santa Monica, 79–84.
21. Chuang, Z. J., Wu, C. H. (2002). Emotion recognition from textual input using an emotional semantic network. In: *Int. Conf. on Spoken Language Processing, ICSLP 2002*, Denver, 177–180.
22. Tao, J. (2003). Emotion control of chinese speech synthesis in natural environment. In: *Eurospeech2003*, Geneva.
23. Moriyama, T., Ozawa, S. (1999). Emotion recognition and synthesis system on speech. In: *IEEE Int. Conf. on Multimedia Computing and Systems*, Florence, Italy, 840–844.

24. Massaro, D. W., Beskow, J., Cohen, M. M., Fry, C. L., Rodriguez, T. (1999). Picture my voice: Audio to visual speech synthesis using artificial neural networks. In: AVSP'99, Santa Cruz, CA, 133–138.
25. Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. University of Chicago Press, Chicago.
26. Etcoff, N. L., Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, vol. 44, 227–240.
27. Ekman, P., Friesen, W. V. (1997). *Manual for the Facial Action Coding System*. Consulting Psychologists Press, Palo Alto, CA.
28. Yamamoto, E., Nakamura, S., Shikano, K. (1998). Lip movement synthesis from speech based on Hidden Markov Models. *Speech Commun.*, vol. 26, 105–115.
29. Tekalp, A. M., Ostermann, J. (2000). Face and 2-D mesh animation in MPEG-4. *Signal Process.: Image Commun.*, vol. 15, 387–421.
30. Lyons, M. J., Akamatsu, S., Kamachi, M., Gyoba, J. (1998). Coding facial expressions with gabor wavelets. In: 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition, Nara, Japan, 200–205.
31. Calder, A. J., Burton, A. M., Miller, P., Young, A. W., Akamatsu, S. (2001). A principal component analysis of facial expression. *Vis. Res.*, vol. 41, 1179–208.
32. Kobayashi, H., Hara, F. (1992). Recognition of six basic facial expressions and their strength by neural network. In: *Intl. Workshop on Robotics and Human Communications*, New York, 381–386.
33. Bregler, C., Covell, M., Slaney, M. (1997). Video rewrite: Driving visual speech with audio. In: *ACM SIGGRAPH'97*, Los Angeles, CA, 353–360.
34. Cosatto, E., Potamianos, G., Graf, H. P. (2000). Audio-visual unit selection for the synthesis of photo-realistic talking-heads. In: *IEEE Int. Conf. on Multimedia and Expo*, New York, 619–622.
35. Ezzat, T., Poggio, T. (1998). MikeTalk: A talking facial display based on morphing visemes. In: *Computer Animation Conf.*, Philadelphia, PA, 456–459.
36. Gutierrez-Osuna, R., Rindomin, J. L. (2005). Speech-driven facial animation with realistic dynamics. *IEEE Trans. Multimedia*, vol. 7, 33–42.
37. Hong, P. Y., Wen, Z., Huang, T. S. (2002). Real-time speech-driven face animation with expressions using neural networks. *IEEE Trans. Neural Netw.*, vol. 13, 916–927.
38. Verma, A., Subramaniam, L. V., Rajput, N., Neti, C., Faruque, T. A. (2004). Animating expressive faces across languages. *IEEE Trans Multimedia*, vol. 6, 791–800.
39. Collier, G. (1985). *Emotional expression*, Lawrence Erlbaum Associates. <http://faculty.ucsb.edu/~gcollier/>
40. Argyle, M. (1988). *Bodily Communication*. Methuen & Co, New York, NY.
41. Siegman, A. W., Feldstein, S. (1985). *Multichannel Integrations of Nonverbal Behavior*, Lawrence Erlbaum Associates, Hillsdale, NJ.
42. Feldman, R. S., Philippot, P., Custrini, R. J. (1991). Social competence and nonverbal behavior. In: Rimé, R. S. F. B. (ed) *Fundamentals of Nonverbal Behavior*. Cambridge University Press, Cambridge, 329–350.
43. Knapp, M. L., Hall, J. A. (2006). *Nonverbal Communication in Human Interaction*, 6th edn. Thomson Wadsworth, Belmont, CA.
44. Go, H. J., Kwak, K. C., Lee, D. J., Chun, M. G. (2003). Emotion recognition from facial image and speech signal. In: *Int. Conf. Society of Instrument and Control Engineers*, Fukui, Japan, 2890–2895.
45. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M. et al. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In: *Int. Conf. on Multimodal Interfaces*, State College, PA, 205–211.
46. Song, M., Bu, J., Chen, C., Li, N. (2004). Audio-visual based emotion recognition – A new approach. In: *Int. Conf. on Computer Vision and Pattern Recognition*, Washington, DC, USA, 1020–1025.

47. Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T. S., Roth, D., Levinson, S. (2004). Bimodal HCI-related emotion recognition. In: *Int. Conf. on Multimodal Interfaces*, State College, PA, 137–143.
48. Zeng, Z., Tu, J., Pianfetti, B., Huang, T. S. Audio-visual affective expression recognition through multi-stream fused HMM. *IEEE Trans. Multimedia*, vol. 10(4), 570–577.
49. Zeng, Z., Tu, J., Liu, M., Huang, T. S., Pianfetti, B., Roth D., Levinson, S. (2007). Audio-visual affect recognition. *IEEE Trans. Multimedia*, 9 (2), 424–428.
50. Wang, Y., Guan, L. (2005). Recognizing human emotion from audiovisual information. In: *ICASSP, Philadelphia, PA*, Vol. II, 1125–1128.
51. Hoch, S., Althoff, F., McGlaun, G., Rigoll, G. (2005). Bimodal fusion of emotional data in an automotive environment. In: *ICASSP, Philadelphia, PA*, Vol. II, 1085–1088.
52. Fragopanagos, F., Taylor, J. G. (2005). Emotion recognition in human-computer interaction. *Neural Netw.*, 18, 389–405.
53. Pal, P., Iyer, A. N., Yantorno, R. E. (2006). Emotion detection from infant facial expressions and cries. In: *Proc. Int'l Conf. on Acoustics, Speech & Signal Processing*, Philadelphia, PA, 2, 721–724.
54. Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Paouzaoui, A., Karpouzis, K. (2006). Modeling naturalistic affective states via facial and vocal expression recognition. In: *Int. Conf. on Multimodal Interfaces*, Banff, Alberta, Canada, 146–154.
55. Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaoui, A., Malatesta, L., Kollias, S. (2007). Modeling naturalistic affective states via facial, vocal, and bodily expression recognition. In: *Lecture Notes in Artificial Intelligence*, vol. 4451, 91–112.
56. Chen, C. Y., Huang, Y. K., Cook, P. (2005). Visual/Acoustic emotion recognition. In: *Proc. Int. Conf. on Multimedia and Expo*, Amsterdam, Netherlands, 1468–1471.
57. Picard, R. W. (2003). Affective computing: Challenges. *Int. J. Hum. Comput. Studies*, vol. 59, 55–64.
58. Ortony, A., Clore, G. L., Collins, A. (1990). *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge.
59. Carberry, S., de Rosis, F. (2008). Introduction to the Special Issue of *UMUAI* on 'Affective Modeling and Adaptation', *International Journal of User Modeling and User-Adapted Interaction*, vol. 18, 1–9.
60. Esposito, A., Balodis, G., Ferreira, A., Cristea, G. (2006). Cross-Modal Analysis of Verbal and Non-verbal Communication. Proposal for a COST Action.
61. Yin, P. R., Tao J. H. (2005). Dynamic mapping method based speech driven face animation system. In: *The 1st Int. Conf. on Affective Computing and Intelligent Interaction (ACII2005)*, Beijing., 755–763.
62. O'Brien, J. F., Bodenheimer, B., Brostow, G., Hodgins, J. (2000). Automatic joint parameter estimation from magnetic motion capture data. In: *Graphics Interface 2000*, Montreal, Canada, 53–60.
63. Aggarwal, J. K., Cai, Q. (1999). Human motion analysis: A review. *Comput. Vision Image Understand.*, vol. 73(3), 428–440.
64. Gavrilu, D. M. (1999). The visual analysis of human movement: A survey. *Comput. Vision Image Understand.*, vol. 73(1), 82–98.
65. Azarbayejani, A., Wren, C., Pentland, A. (1996). Real-time 3-D tracking of the human body. In: *IMAGE'COM 96*, Bordeaux, France.
66. Camurri, A., Poli, G. D., Leman, M., Volpe, G. (2001). A multi-layered conceptual framework for expressive gesture applications. In: *Intl. EU-TMR MOSART Workshop*, Barcelona.
67. Cowie, R. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.*, vol. 18(1), 32–80.
68. Brunelli, R., Falavigna, D. (1995). Person identification using multiple cues. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17(10), 955–966.
69. Kumar, A., Wong, D. C., Shen, H. C., Jain, A. K. (2003). Personal verification using palmprint and hand geometry biometric. In: *4th Int. Conf. on Audio- and Video-based Biometric Person Authentication*, Guildford, UK, 668–678.

70. Frischholz, R. W., Dieckmann, U. (2000). Bioid: A multimodal biometric identification system. *IEEE Comput.*, vol. 33(2), 64–68.
71. Jain, A. K., Ross, A. (2002). Learning user-specific parameters in a multibiometric system. In: *Int. Conf. on Image Processing (ICIP)*, Rochester, New York, 57–60.
72. Ho, T. K., Hull, J. J., Srihari, S. N. (1994). Decision combination in multiple classifier systems. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 16(1), 66–75.
73. Kittler, J., Hatef, M., Duin, R. P. W., Matas, J. (1998). On combining classifiers. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20(3), 226–239.
74. Dieckmann, U., Plankensteiner, P., Wagner T. (1997). Sesam: A biometric person identification system using sensor fusion. *Pattern Recognit. Lett.*, vol. 18, 827–833.
75. Silva, D., Miyasato, T., Nakatsu, R. (1997). Facial emotion recognition using multi-modal information, In: *Proc. Int. Conf. on Information and Communications and Signal Processing*, Singapore, 397–401.

Chapter 10

Spoken Language Translation

Farzad Ehsani, Robert Frederking, Manny Rayner, and Pierrette Bouillon

10.1 The Dream of the Universal Translator

Researchers in the field of spoken language translation are plagued by a device from popular science fiction. Numerous television series and movies, most notably those in the “Star Trek” franchise, have assumed the existence of a Universal Translator, a device that immediately understands any language (human or alien), translates it into the other person’s language (always correctly), and speaks it fluently, with appropriate prosody. While this is a useful plot device, avoiding tedious stretches of translation and the need to invent convincing alien languages, it sets up wildly unrealistic expectations on the part of the public [1]. In contrast, anything that is actually possible can only be a disappointment.

Of course, these stories also feature many other plot devices that stretch or violate our current understandings of science but make the storyline work better: faster-than-light travel, teleportation, intelligent aliens that happen to breathe oxygen, and so forth. Yet the public is not disappointed when, for example, it takes years for an actual spacecraft to reach another planet. Perhaps because there is no obvious violation of physics associated with the Universal Translator, it is much less obvious to most people that a true Universal Translator – a device that can translate every known language on this planet (or others for that matter) – is unlikely ever to exist, just as humans are unlikely ever to travel faster than light.

A variety of approaches to spoken language translation are presented in this chapter, but the fundamental problem with the Universal Translator transcends any specific technology. This fundamental problem is the need to match the words of one language to the words of another language (ignoring for the moment all the other knowledge required, regarding syntax, phonology, etc.) In each language, the match between its words and their meanings is arbitrary. There is nothing intrinsic to the letters or sounds of the word “soap” that indicate that it is something you wash with, or of the word “soup” that indicate that it is a liquid you eat. (Worse still, “soap”

F. Ehsani (✉)
Fluential, Inc, 1153 Bordeaux Drive, Suite 211, Sunnyvale, CA 94089, USA
e-mail: farzad@fluentialinc.com

can also refer to a particular kind of television show, and “soup” can also refer to a particular kind of fog.) For each language, one simply must *know* what each word can mean. Thus the match between words of different languages is also completely arbitrary. It is worth noting that statistical systems that achieve human-level quality will need to learn this information as well, whether explicitly or implicitly.

If one takes the term “universal” seriously, this arbitrary match is an insurmountable problem. There are roughly 6,000 living languages today, so a “universal” translator would need to contain detailed knowledge about all the words in all these languages (many of which lack any significant quantity of written texts). Lowering the bar substantially, to languages with at least one million native speakers, still leaves over 300 languages to deal with. Constructing a Universal Translator to handle *just* these 300 languages would require developing speech recognition, translation, and synthesis for each of the 300. This would clearly still be a massive undertaking.

Although the Universal Translator is not on the horizon, significant progress has been made in recent years toward the more modest goal of acceptable-quality translation between single pairs of languages, given substantial development effort on each language pair. Enough progress has been made that it is clear that useful spoken language translation will be developed in the near future, although as of this writing no such systems are in mass production or use (that the authors are aware of), in contrast to speech recognition and machine translation technologies used independently.

In this chapter, we describe progress on building real spoken language translation systems, focusing primarily on ones in which the authors of the chapter have been personally involved. Even though this group represents only a small sampling of all of the different systems that are out there, we believe that they provide a good cross-section of the full range of systems. We start in Section 10.2 by briefly reviewing the component technologies. Section 10.3 describes the systems themselves, and Section 10.4 concludes.

10.2 Component Technologies

A speech translation system requires three main component technologies: Automatic Speech Recognition (ASR), Machine Translation (MT), and speech synthesis. Each of these components is obviously necessary. What is perhaps less obvious is that the integration of the individual components poses not-trivial problems: in particular, efficient integration of ASR and MT raises many interesting questions. We however postpone discussion of integration issues until Section 10.3, and for the moment discuss each technology in isolation.

10.2.1 Speech Recognition Engines

Since the early 1970s, a number of different approaches to ASR have been proposed and implemented, including Dynamic Time Warping, template matching,

knowledge-based expert systems, neural nets, and Hidden Markov Modeling (HMM) [2–4]. HMM-based modeling applies sophisticated statistical and probabilistic computations to the problem of pattern matching at the sub-word level. The generalized HMM-based approach to speech recognition has proved an effective, if not *the* most effective, method for creating high-performance speaker-independent recognition engines that can cope with large vocabularies; the vast majority of today’s commercial systems deploy this technique.

A speech recognition engine has three major components that require data collection – acoustic models, dictionaries, and language models. Acoustic models define the relationship between the sounds of a language (the phonemes) and the structure of the acoustic signal (the waveforms); these models are virtually always statistical in nature, and trained from large quantities of recorded and transcribed speech. The data used to train an acoustic model does not need to be specific to the task, although it helps if it reflects the user population and the domain [5]. Dictionaries are usually created by a combination of rules and careful transcription of recorded data. Our experience has shown that investing time in building good dictionaries typically improves word accuracy by 3–7%.

Language models constrain the range of possible utterances that the ASR component is expected to recognize, and can be either statistical or rule based in nature. Language models have a huge effect on system performance, though the actual amount depends heavily on domain vocabulary, perplexity, and word confusability. Statistical language models, which are at the moment the more popular choice, typically require large amounts (tens of thousands of utterances) of task-specific text data for training [4, 6]. Grammar-based language models are less common, and require more development work, but have the advantage of requiring far more modest quantities (thousands, or even hundreds of utterances) of domain data. The MedSLT system described in Section 10.3.1 uses a grammar-based language model; the other systems covered in Section 10.3 use statistical models.

Generally, ASR systems output not only the highest-confidence recognition hypothesis, but also lower-confidence hypothesis along with confidence scores. Some systems output a lattice of words, with their associated confidence scores. With N-best or lattice output, the result can be further processed to achieve better recognition or processed in conjunction with the MT system to achieve better translation accuracy.

Depending on the task, the accuracy of speech recognition engines has improved dramatically in the past 20 years. On average, word error rates (WER) have been improving by about 10% per year [7].

10.2.2 Translation Engines

Machine translation (MT) is a rapidly growing and increasingly important field. While English remains the primary language of the World Wide Web, non-English online text is expected to become more prevalent in the immediate future. The need for accurate translation is therefore accelerating, fed by the Internet and increasing communication among diverse language groups.

One of the very few MT systems capable of producing high-quality output was developed by the CATALYST project at Carnegie Mellon University using the KANT MT technology [8]. The system, funded by Caterpillar, the heavy equipment manufacturer, translates technical texts from Caterpillar's vast document library well enough to significantly enhance the productivity of translators who edit the final documents. This system is an example of an interlingua-based MT, an approach that builds a language-independent meaning representation of the input which is then used to generate an output sentence [9]. However, this approach is limiting in several ways. The system requires the use of highly constrained input text, limited to a small subset of English which must be manually pre-processed to reduce ambiguity. An interlingua, while attractive in theory, has never been shown to be generalizable or scalable to any significantly larger domain.

Interlingua-based MT is one category of a class of MT known as knowledge-based machine translation (KBMT), so called because it utilizes significant linguistic and domain-dependent knowledge sources. The other major category of KBMT is transfer-based MT. Transfer-based MT uses an extensive grammar of the source language to parse the input sentence, transforms the source parse tree into a generation tree using transfer rules, and finally creates the target-language string according to a set of generation rules. Whether a transfer-based system is considered to be "KBMT" or not depends on whether domain semantic information is used in the translation process, which for some systems may be a subject of dispute. One problem with such transfer-based systems is that separate transfer rules must be developed for each new language pair. When combined with the analysis and generation rule-sets, the total number of rules necessary for a developed transfer MT system can be substantial.

Clearly, KBMT and transfer-based systems can only be as good as the knowledge or rules that they contain, and that knowledge is extremely expensive to construct. Monolingual and bilingual lexicons typically absorb a large portion of the development effort. Developing correct and complete grammars and other transformational rules is difficult. Choosing the appropriate translation of a word with multiple meanings may require complex lexical selection rules to analyze the context of its use. Development of these systems can easily take many man-years of work, even if they are special purpose translators such as the CATALYST project, and especially if they are all-purpose translation engines such as SYSTRAN. Moreover, most such systems are not easily reversible. Translation in each direction usually requires the development of separate, complex components.

Since the early 1990s, researchers have investigated new approaches to MT that seek to learn appropriate translations automatically from parallel data. Grouped under the general heading of empirical approaches, these include example-based MT (EBMT) and statistical MT (SMT). The development of these MT approaches, and in particular SMT, is consistent with a general trend in natural language processing toward quantitative empirical methods, which has been made possible by the increasing availability of large electronic text corpora.

EBMT systems utilize a relatively small number of parallel source and target-language sentences as examples to generate translations in both directions [10, 11].

The idea behind EBMT is simple: look up similar examples among the translation examples, adapt them to fit the source input, and perform corresponding adaptation on the matching target-language example to create the translation output. While intuitive, EBMT has not gained wide acceptance because of the difficulty in matching new translation text against stored examples [12]. Additionally, the algorithms that combine the fragments into sentences on both the source and target sides are extremely complex.

In contrast, SMT systems attempt to compile the training data into summary statistics instead of treating it as a set of discrete examples. SMT requires a parallel corpus that is sufficiently large (millions of words) and aligned on a sentence-by-sentence basis, in order to learn accurate information about the language [13–17]. The methods of statistical MT are strongly related to those used in acoustic modeling for speech recognition; these approaches have shown increasing promise recently because of vast increases in processing power now available, as well as the relatively cheap price of that power. However, these methods suffer from the “sparse data problem,” which makes them impractical to employ without massive bilingual corpora, and the scarcity of such corpora has been a barrier to SMT for many language pairs. For these reasons, MT researchers have recently begun investigating methods of integrating linguistic knowledge with empirically based approaches [18–21].

Wang [22] tried to learn the phrase structure as well as mapping across the source and target languages. This involved constraining the search space by using part-of-speech tags and limiting the average translation span. Similarly, Och and his colleagues [23] attempted to learn phrase alignment templates from parallel data and used instantiations of the templates to segment input sentences to produce translations. These alignment templates establish a cross-language mapping of words and positions within phrases, with the words having been previously grouped into optimal classes using bilingual data. However, neither of these approaches makes much explicit use of syntax or semantics in deriving the phrase structures or the mapping. Most recently, Galley [24] and Venugopal [25] improve the state-of-the-art phrase-based statistical translation by incorporating syntactic information, and Chiang [26] by introducing a hierarchical phrase model.

10.2.3 *Synthesis*

Early approaches towards synthesis of speech that attempted to synthesize “from scratch” have been almost entirely replaced in current spoken language translation systems by *concatenative* approaches that start with a database of human speech, select appropriate snippets of sound, and smooth the resulting junctions using digital signal processing techniques. This approach is similar in spirit to empirical translation techniques that work with large corpora of parallel translated sentences.

Speech synthesis in general attempts to address a number of issues beyond simply synthesizing reasonably correct sounds from letters; these issues include

understandability, appropriateness in dialect, gender, and register, and sufficiently natural quality to avoid irritation of the listener. Speech synthesis intended for use in spoken language translation applications must also deal with the intrinsic need to provide synthesis for multiple languages. As mentioned above for recognition and translation, the languages dealt with often lack the large amounts of data available for work in English; some of them lack any significant amount of usable pre-existing data. This data scarcity has a significant effect on the research directions taken [27]. In addition to the obvious issue of acquiring language data, many languages present additional cultural or logistical problems. As one example, languages that are primarily oral in culture may have a standard writing system (or several!), but if the language is not widely written, spellings of even common words by native speakers of the language may differ frequently. Time and effort must be spent standardizing the spelling and usage of the language experts before useful databases can be constructed. Cultural issues can include, e.g., a reticence to criticize, making critical evaluation of prototypes of the system difficult to accomplish.

Other practical concerns affecting synthesis development include recent trends toward very small platforms and rapid development of systems, requiring small footprint systems and new tools for language portability. A similar but more cross-cutting issue is the need to keep the vocabularies of the different components synchronized, since the lexicons of the individual components are typically separate databases.

In the next section, we show how the individual components we have just presented have been combined to build five representative speech translation systems.

10.3 Specific Systems

The systems we look at in this section illustrate several fairly different approaches to the problem of designing a coherent speech translation architecture; to give some context, we first present a thumbnail history of the field of speech translation as a whole. Worldwide, the organization which has had the most enduring influence on the early development of speech translation technology is almost certainly Carnegie Mellon University. CMU initiated speech translation work with its JANUS project in the late 1980s [28]; shortly after this, the University of Karlsruhe, Germany, also became involved in an expansion of JANUS, beginning a long-term partnership which was later formalized as the interACT center for international collaboration (<http://www.is.cs.cmu.edu/>). In 1992, CMU and the University of Karlsruhe joined ATR in the C-STAR consortium (Consortium for Speech Translation Advanced Research), and in January 1993 gave a successful public demonstration of telephone translation between English, German, and Japanese, within the limited domain of conference registrations [29]. Since then, C-STAR has acquired several more partners and firmly established itself as the largest player on the speech translation scene. It has in particular taken a leading role in developing shared tasks,

organizing associated data collection exercises, and in general facilitating communication between different groups involved in speech translation. CMU has also been involved to some extent in many other major speech translation projects, including VerbMobil, NESPOLE!, Tides, Babylon, and TC-Star.

Although CMU has been the most important single center, primarily due to the fact that it has been continuously involved in speech translation throughout the whole period that the field has existed, there have been several other groups which have played important roles in its development. Verbmobil [30] was a very large speech translation effort funded by the German government throughout the greater part of the 1990s and Spoken Language Translator [31–33], a substantial project funded by Telia Research and involving SRI International, was active during roughly the same period. Other large speech translation projects from the 1990s were those from NEC [34] and ATR [35] in Japan, the NESPOLE! Consortium [36], and the AT&T speech translation system [37]. More recently, LC-Star and TC-Star are two European efforts which have played a leading role in gathering the data and the industrial requirements to enable pervasive speech-to-speech translation [38]. In the United States, the most substantial program during the first few years of the 21st century has been DARPA's TransTac, in which the key participants have been BBN, CMU, Fluential, IBM, SRI International, and USC.

Over the last 15–20 years, the architecture of speech translation systems has evolved through several stages. Reflecting popular thinking in the early 1990s, several important early speech translation systems featured linguistically motivated approaches, in which a substantial part of the processing was based on grammars structured in accordance with accepted linguistic theories. The starting point for systems of this kind was a pair of substantial grammars for the source and target languages, together with compilers that could be used to turn them into parsers and generators. These could be combined into a translation system by adding a transfer component, which mapped semantic representations on the source side into corresponding semantic representations on the target side. Further addition of recognition and speech synthesis modules yielded a speech translation system. The best-known examples of these early linguistically motivated systems are Verbmobil and the Spoken Language Translator. SLT, which we choose as a representative example of this style of speech translation system, is described in the first half of Section 10.3.1.

As we will soon see, the pure linguistically motivated approach turned out to be less promising than it first appeared, and at the same time the advantages of data-driven approaches, although more expensive, became widely apparent. This pushed development in several different directions. Systems which still used linguistically motivated approaches now combined these with data-driven methods to yield various kinds of hybrid architectures, as described in the second half of Section 10.3.1. Another reaction was to retreat toward simple fixed-phrase systems, concentrating on making these as robust and practically useful as possible. The best-known system of this kind is Phraselator, described in Section 10.3.2. A related idea, which was initially explored by CMU researchers, was to develop architectures suitable for systems which could be rapidly deployed to new language pairs. Two influential

systems in this area were Diplomat and Tongues, which are described in Section 10.3.3.

As already indicated, however, the mainstream moved toward data-driven architectures, which rapidly became dominant. The biggest problem with building statistical systems is the need for large parallel corpora. For a text-based translation system, corpus data can often be acquired by web mining; unfortunately, this is not generally a feasible way to collect parallel conversational corpora. However, as the EU's C-STAR data became publicly available for the travel domain, and as the DARPA Babylon and TransTac programs started, there have been more parallel corpora available for building statistical and hybrid systems. Section 10.3.4 describes a state-of-the-art mainstream system developed by Flunential, whose top-level architecture is a statistical-/rule-based hybrid.

10.3.1 SLT and MedSLT

This section describes a series of related systems developed over the last 15 years, starting with the first version of the Spoken Language Translator (SLT-1) in the early 1990s and continuing up to the MedSLT system, which is still very active in 2008. SLT-1 [39] was constructed between 1992 and 1993 by a team of researchers from SRI International, Telia Research, Stockholm, and the Swedish Institute of Computer Science; it translated questions in the Air Travel Information Services (ATIS) domain from English to Swedish, using a vocabulary of about 1500 words. An English-to-French version was built in 1994 in collaboration with ISSCO, Geneva. The main processing-level components of the system were SRI's DECIPHER speech recognition and Core Language Engine (CLE) language processing platforms, and Telia's Prophon speech synthesis system. DECIPHER [40], an HMM-based recognition platform, was later developed into the commercial Nuance system.

The Core Language Engine [41] was a unification-based processing platform that supported grammar-based parsing and generation of language. Transfer was performed at the level of Quasi Logical Form (QLF) [42], an elaborate higher-order logical-based surface semantic form. For example, the QLF representation of "Where will the plane stop?" was

```
[whq,
 form(verb(no,no,no,will1,yes),E,
  [[where1,E,term(q(wh,X,[place,X])],
   [stop1,E,term(ref(def,the,sing),
    Y,[plane1,Y])]])])]
```

The basic idea is that `terms` represent NPs, while forms represent VPs and some other relations. Here, the representation of "the plane" is

```
term(ref(def,the,sing),Y,[plane1,Y])
```

The first field, $\text{ref}(\text{def}, \text{the}, \text{sing})$, represents the (definite, singular) determiner “the,” the second field, Y , is the variable associated with the noun phrase, and the third field, $[\text{plane1}, Y]$, is the noun “plane.” The piece of representation immediately enclosing the term, $[\text{stop1}, E, \dots]$, represents the fact that the subject of the verb “stop” is “the plane.” The tense and aspect information associated with “stop” is, however, moved out to be the first argument of the form :. The five arguments indicate that there is no explicit tense inflection (the first occurrence of “no”), the aspect is neither perfect nor continuous (the second and third occurrences of “no”), there is a modal auxiliary (the “will”), and the voice is active (the final “no”).

The motivation for this complex formalism was that it offered a deep level of structure in which utterances with widely differing surface syntax but equivalent meaning could often receive a uniform representation. For example, a possible French translation for “Where does the plane stop?” is “*Où l’avion s’arrêtera-t-il?*” (“Where the-plane itself-stop-FUTURE-it?”). At first glance, this translation appears quite different from the English; however, the QLF representation,

```
[whq,
  form(verb(fut,no,no,no,yes),E,
    [[où1,E, term(q(wh,X,[place,X])],
      [se_arrêter1,E, term(ref(def,le,sing),
        Y,[avion1,Y])]])])]
```

has exactly the same structure. This example shows both the strengths and the weaknesses of the QLF idea; although “*Où l’avion s’arrêtera-t-il?*” is indeed a correct French translation with a semantic structure isomorphic to the original English, it is considerably more idiomatic to say “*Où l’avion fait-il escale?*” (“Where the plane makes-it stop?”). In general, aiming toward translations that preserve structural correspondence tends to foster this type of problem, something that became increasingly apparent as the project progressed and higher expectations were placed on translation quality.

In SLT-1, all processing except speech recognition was rule-based, and all components were connected together in a pipeline. DECIPHER delivered its 1-best speech hypothesis to the source (English-grammar) version of the CLE, which converted it to a QLF representation. This was transferred into a target (French or Swedish) QLF, and a target representation was then generated using a second copy of the CLE equipped with a target-language grammar. In retrospect, it is unsurprising that performance was weak. Pure rule-based processing is usually brittle, and SLT was no exception. Almost immediately, the project began introducing data-driven methods. The first of these was a trainable method for inducing parse preferences [43]. Shortly afterward, the general English source-language grammar was replaced by a specialized version that was created using the corpus-driven Explanation Based Learning (EBL) method [44]; this also necessitated replacing the original CLE left-corner parser with a new one, which used an adapted version of the LR algorithm [45].

The two methods described above considerably improved performance with respect to both speed and accuracy, but the system still suffered from a generic problem that was more deeply embedded in the architecture. Since all language processing was grammar-based, recognized utterances that failed to be inside grammar coverage could never be processed at all. Although grammar coverage was carefully adapted to the ATIS domain, the recognizer was not directly aware of the grammar's constraints, and random misrecognitions, typically triggered by various kinds of dysfluencies, often resulted in ungrammatical recognition results being produced. Performance on unseen ATIS data, as reported in [31], resulted in only about 42% of all utterances receiving a correct translation.

In SLT-2, the second (1995–1996) phase of the project, the system's designers decided that it would be impossible to achieve the desired improvements in performance without a substantial reorganization of the architecture. They consequently introduced a version of the “multi-engine translation” idea originally proposed in [46], and added shallower levels of fall-back processing. The new version of the system permitted translation at three separate levels of representation. The parsing algorithm was reorganized to be a bottom-up version of the one from [45]. This meant that partial constituents could be produced and added to a chart structure, and that linguistic analysis could be viewed as an “anytime” algorithm, which could be given a fixed timeout parameter and still always be able to return meaningful results. When linguistic analysis stopped, the system first looked for a constituent that spanned the whole input, and attempted to translate it. If this failed, it then backed off to finding as good a sequence as possible of the incomplete parsed constituents, translating each one individually. Any remaining gaps, consisting of unparsed individual words, were processed using surface phrase translation rules.

With further improvements, as described in detail in [33], the final version of the system, SLT-3, was able to produce adequate translations for about 75% of all utterances on unseen speech data, and was generally acknowledged to be an impressive demo system. Unfortunately, this apparently good result concealed less welcome news. Although it was in the end possible to overcome the lack of robustness in the original prototype, the cost was an over-elaborate architecture that was hard to extend and maintain. One key issue was the complex linguistically motivated QLF representations, which were as much a hindrance as an asset. Though elegant, they were also challenging to learn, and non-trivial translation mismatches tended to result in convoluted transfer rules. The worst problem, however, was the incompatibility between the recognition and language processing architectures. The statistical language models used by the recognizer were fundamentally different from the rule-based grammars on which the language processing was based, and a large part of the system's complexity resulted from the need to smooth over these differences.

It became clear that the system would have been much easier to construct if the designers had thought more about two fundamental design issues at the start. First, it would have been better to use compatible architectures for speech and language processing; both could have been fundamentally data driven, or both could have been fundamentally rule-based, but the combination of a statistical recognizer with a rule-based language processing engine was not a good choice. Second, the semantic

representation language was too fine-grained and made all the linguistic rule-sets extremely complex, with correspondingly longer times required for development, maintenance, and porting.

These lessons enabled the designers to make much better choices in the later MedSLT system. MedSLT, based at Geneva University in Switzerland, is an Open Source speech translation project that has been active since 2003. (A pre-study, resulting in a sketchy initial prototype, is described in [47].) At the name suggests, the domain is medical. Specifically, the project focuses on doctor/patient examination dialogues. At the time of writing, in late 2007, there are versions of the system for about 20 different language pairs and four subdomains. Languages handled include English, French, Japanese, Spanish, Catalan, and Arabic; the subdomains are headaches, chest pain, abdominal pain, and sore throat. Performance has been tuned most intensively for English \leftrightarrow French and Japanese \leftrightarrow English, and the headache domain, but is respectable for several other language pairs and subdomains as well. Vocabulary size varies depending on the input language, being determined mainly by the number of inflected forms of verbs and adjectives. It ranges from about 1,100 for French (highly inflected) to about 400 for Japanese (highly uninflected).

Despite the similarity in names and a non-trivial overlap in key project personnel, MedSLT differs in many respects from SLT. Learning from their experience with the earlier system, the developers made several critical changes. Although, as before, they chose an architecture that combined both rule-based and data-driven aspects, they were much more careful to ensure that these two methodologies were used in compatible ways. As in SLT, language processing is based on general, linguistically motivated grammars. However, instead of performing recognition using statistical language models, which frequently results in production of out-of-coverage recognition results and associated complications, MedSLT also uses a grammar-based approach to language modeling. In fact, the *same* grammar is used for both recognition and source-language analysis, guaranteeing that all recognition results are within grammar coverage. The process of compiling linguistically motivated grammars into language models is handled by the Open Source Regulus platform [48]. Data-driven methods are still used, but only within the bounds of the encompassing grammatical framework. In particular, just as in SLT, domain-specific grammars are derived from the original general ones using the corpus-based EBL method; the training corpus is also used a second time, to perform probabilistic tuning of the generated CFG language model. This probabilistic training has a large impact on recognition performance, and for some languages can reduce the semantic error rate by as much as 50%, even with quite modest quantities of training data [48].

With MedSLT, N-gram-based language modeling is not abandoned altogether, but used to add robustness, as part of an intelligent help module. After each utterance has been processed by the primary, grammar-based recognizer, it is then passed to a secondary recognizer equipped with a conventional N-gram model. Although the N-gram-based recognizer performs considerably worse than the grammar-based one on in-coverage data, its WER is, unsurprisingly, better on out-of-coverage utterances. (Similar results were obtained for the Clarissa system; cf. Section 12.4). The

secondary recognition result is matched against a precompiled library of in-domain utterances that are already known to produce correct translations, and the most similar examples are displayed to the user. In practice, we have found that addition of the help module makes it considerably easier for new users to learn the grammar's coverage [49], with users typically becoming confident after only one to two hours of practice [50].

A third difference between MedSLT and SLT is a greatly simplified representation language. All semantic representations are unordered lists of attribute-value pairs, with at most one level of nesting permitted. A representation of the source-language analysis of an English sentence with a subordinate clause is shown in Fig. 10.1; the nested list starting with "clause" represents the subordinate clause. Subordinate clauses are important in MedSLT, since the doctor often wants to ask the patient when symptoms occur.

Fig. 10.1 Source representation for "Do you have headaches when you drink red wine?"

```
[ [utterance_type, ynq],
  [pronoun, you], [state, have_symptom],
  [symptom, headache],
  [tense, present], [voice, active],
  [sc, when],
  [clause, [ [utterance_type, dcl],
             [pronoun, you], [action, drink],
             [cause, red_wine]
             [tense, present]]]]
```

In a large, open domain, this minimalist style of representation would clearly lose too much information. Since the semantic representations are unordered lists, there would, for example, be no distinction between "John acquired the company" and "The company acquired John." For the fairly restricted domains used in MedSLT, however, the formalism appears adequately expressive. Even though the original resource grammars are quite general, the *actual* grammars used for analysis and generation are heavily specialized to the domain, both by the corpus-based grammar specialization method and by sortal (semantic-type) constraints inherited from a domain-specific lexicon. For example, although both "You have a headache" and "A headache has you" could in principle be grammatical, sortal constraints on the lexical entries for "have," "you" and "headache" ensure that only the first of these can actually be generated.

Since one of the key goals of the project is to enable rapid porting to new language pairs and subdomains, translation is interlingua-based. It consists of four main stages: (1) mapping from the source representation to interlingua; (2) ellipsis processing (when it is necessary); (3) mapping from interlingua to the target representation; and (4) generation, using a suitably compiled Regulus grammar for the target language. These stages are shown in Fig. 10.2, which illustrates most of the key representational issues in the MedSLT system. Interlingual structures

```

Source = [[utterance_type,ynq],
          [modal,can], [cause,bright_light],
          [action,give], [voice,active],
          [pronoun,you],
          [symptom,headache]]

Interlingua = [[utterance_type,ynq],
               [pronoun,you],
               [state,have_symptom],
               [symptom,headache],
               [tense,present]

               [sc,when],
               [clause,
                [[utterance_type,dcl],
                 [pronoun,you],
                 [cause,bright_light],
                 [state,experience],
                 [tense,present],
                 [voice,active]]]]

Target = [[utterance_type,sentence],
          [cause,lumière_forte], [event,causer],
          [tense,present], [voice,passive]
          [symptom,mal_de_tête]]

```

Fig. 10.2 Translation flow in English → French MedSLT for “Does bright light give you headaches?” → “Vos maux de tête sont-ils causés par une lumière forte?”

are essentially canonical versions of English representations. Here, the English source-language expression “Does bright light give you headaches?” gives rise to an interlingual expression that could be paraphrased as “Do you have headaches when you experience bright light?” Similarly, the main systematic transformation carried out when moving from source to interlingua is concerned with the temporal and causal concepts central in the medical domains. In the same way, “Are the headaches accompanied by nausea?” would give rise to a representation that could be paraphrased as “Do you experience nausea when you have headaches?” Similarly, all the different temporal and causal constructions are mapped to one of the following three interlingua schemas: (1) Clause1 WHEN Clause2; (2) Clause1 BEFORE Clause2; and (3) Clause1 AFTER Clause2. The details are presented in [49].

Apart from the small number of rules dealing with the temporal/causal transformations described above, most of the translation rules are simple; they merely map short lists of feature–value pairs to short lists of feature–value pairs. It is clear that the solutions used would not scale to an arbitrarily complex domain. The range of language covered is, however, enough to cover a large range of medical

questions; during post-experimental debriefing, nearly all the medical student subjects used in [50] stated that, although they were often not able to use the exact phrase they would have preferred, it was usually easy to find an alternative phrasing that was within system coverage. These and other experiments suggest that the representation language of MedSLT is appropriate to the level of difficulty of the domain.

In general, MedSLT shows that linguistically motivated architectures do not have to be baroquely over-complex. By making sensible design choices at the beginning of the project, in particular incorporating data-driven methods when appropriate, the developers were able to create a speech translation application that featured a domain of practical significance, and was at the same time linguistically principled and easy to port to new domains and language pairs.

10.3.2 *Phraselator*

Spoken language translation research in the United States received a major boost in the mid-1990s, when the US Defense Department's Defense Advanced Research Projects Agency (DARPA) was searching for technologies that might have an immediate impact on the actual operations of the Defense Department, especially in its intervention in the conflicts between the states of the former Yugoslavia. In addition to funding research in spoken language translation, DARPA funded a "one-way translation" effort led by contractor Ace Sarich, a former Navy SEAL, beginning in 1997. Sarich's work was intended to be immediately deployable in a very carefully limited application, using currently working technology. Based on an original text-to-speech system by Lee Morin, a US Navy physician, which had been further developed into a speech-to-speech system by the Naval Operational Medical Institute (NOMI) and Dragon Systems, the resulting "Phraselator" incorporates speech technology from SRI International [51].

The Phraselator is specifically intended to be used *only* in the English-to-other-language direction ("one-way"), by a trained English-speaking user. This permits the further assumption that the trained user will compose sentences from a relatively small set of English phrases (typically 500–1,000) designed to cover some particular application. This strong constraint allows clever engineering of all the component technologies to be based entirely on phrases. Thus the speech recognizer in the device recognizes whole phrases as indivisible units; this makes recognition easier and more effective, since the individual units to be recognized are few and acoustically distinct. The translation component is trivial: a table of the English phrases and their manual translations into the target language. Similarly, there is no actual synthesizer; rather, the device simply plays a selection from a set of recordings of all the target-language phrases.

As a result of this great simplicity, a working system can be assembled for a new target language in literally a few days. All that is necessary is to translate the set of phrases for the desired domain into the new language, and then have a native speaker

of the new language record the target phrases. Since this process is so inexpensive, Phraselator language modules are now available in a number of domains, in more than 60 target languages.

In keeping with the theme of immediate military usefulness, special rugged, handheld hardware platforms were quickly developed to run the Phraselator software. These platforms have since been used for other spoken language translation projects. The Phraselator software includes well-engineered GUIs, capabilities for field updates, and other support for practical use.

Since its original trial in Bosnia, the Phraselator has been used in numerous locations, including the Arabian Gulf, Afghanistan, and Iraq. Some anecdotal evidence suggests that it has been useful to personnel in the field, but at the time of writing, there has never been a published technical evaluation of the Phraselator in field use, to the best of the authors' knowledge. In any event, the strong limits on its application to situations where the speaker knows which phrases to use prevent it from being a real competitor with the other systems described here.

10.3.3 Diplomat/Tongues

The Diplomat project [52] at Carnegie Mellon University began as part of the same US DARPA search for immediately usable technology that spawned Phraselator (Section 10.3.2). Researchers at Carnegie Mellon presented a spoken language translation demo to DARPA using already existing technology, and convinced DARPA to fund a “two-way translation” research effort at Carnegie Mellon, with the understanding that this was much more difficult than the “one-way” Phraselator, but might produce usable technology in the relatively near future. This effort was distinguished from the preceding and concurrent Carnegie Mellon Interactive Systems Laboratory (ISL) work (Section 10.3) in that it was intended to be truly portable (indeed, “wearable,” without wireless links to off-board servers) using hardware available at the time, and rapidly retargetable to new domains and languages.

Beginning in 1996, the Diplomat project constructed speech-to-speech translation systems between English and a variety of other languages: Croatian, Haitian Creole, and Spanish. (Spanish was added to allow demonstrations that would be interesting to civilian audiences.) Preliminary work was also done on Korean and Arabic. The system was designed to run on “wearable” hardware, which was produced by another research group at Carnegie Mellon, as well as on the laptops of the day. This hardware consisted of multiple modules that would be worn, e.g., on a belt or strapped to a forearm, to allow field use by a soldier on foot. The wearable platform had two major impacts on the system design: the available computational resources were quite limited and the human factors design was a serious issue.

The requirement to run self-contained on a small platform largely dictated the selection of the speech and translation components. The Sphinx II speech recognizer

[53] was able to perform in real-time on the platform. The speech synthesizer [54] was locally developed for the Diplomat project, and was again designed to work in real-time on the small platform required. To the authors' knowledge, this was the first speech synthesis system developed at Carnegie Mellon.

Similarly, the Carnegie Mellon EBMT system that had been developed during the Pangloss project [11, 55] was able to translate in real-time using the relatively small disks and RAM, and relatively slow processor speeds, which were then available. This EBMT system took a very "shallow" approach to EBMT, similar to early SMT work, but in contrast to some other EBMT systems that attempt to work with linguistic structure. A variety of heuristics are used to produce a lattice of translation hypotheses from a parallel corpus. Then a typical stack decoder – similar to ones used in ASR or SMT systems – is utilized to search this lattice. This was further developed into Generalized EBMT (G-EBMT) [18], which incorporates an ability to generalize matching phrases by using non-terminals representing easily replaceable words such as <CITY>, and other approaches similarly adapted from grammar-based transfer MT systems. In contrast to this EBMT system, the SMT systems of the time would not run in real-time on such a slow platform or work at all with such small RAM. The Diplomat system was also required to *not* be strongly limited-domain, so a KBMT solution was not feasible.

Since the Diplomat system was intended for real field trials, serious attention was again paid to its human factors aspects [56]. The system was designed for use by a well-trained but computer-naïve English speaker, and a totally unprepared non-English-speaker who might not know how to use a keyboard or mouse/joystick. Wizard-of-Oz experiments were carried out early in the development of the system with local computer-illiterates: a project member in a neighboring cubicle would simulate speech recognition and translation output, sending it to the device that the test user was operating, to test the usability/naturalness of the graphical user interface. These experiments led the developers to remove any complexity from the non-English side of the interaction. For example, to confirm the recognition of their utterance, the non-English users would be presented with a display showing only a question in their language about whether this is roughly what they said, their speech recognizer output, and two large buttons saying "Yes" and "No" in their language. This illustrates another aspect to the human factors design: in order to interact with totally naïve users in unpredictable field environments, a portable touch-screen tablet display was included in the system. Similarly, a noise-canceling microphone in the shape of a telephone handset was used, rather than a more typical headset that would have had to have been placed on the user's head *before* talking with them.

The Diplomat system was also intended to be rapidly portable to new language pairs, since the US government saw that its translation needs were changing rapidly in unpredictable directions. The languages of interest were generally not major world languages, and thus had very little online text available. For example, even when Croatian is combined with closely related languages such as Serbian, there are roughly 21 million native speakers; Haitian Creole is even smaller, being

spoken natively by about 7 million people. The small amount of data available for these languages was a significant issue in developing these systems. The Diplomat EBMT system was able to function with much less data than SMT, but significant project effort went into translating newswire texts into the languages of interest, to create broad-coverage parallel text. The system made use of online dictionaries and glossaries as well, which generally required significant manual clean-up to be usable. The speech recognition research in the project was also largely geared toward problems of very small data resources [57].

Due to these computational resource and training data limitations, the Diplomat system was never expected to perform at impressive levels of quality compared to systems running on large machines working with large-data languages. Rather, the interesting evaluation question in the context of the Diplomat system was whether the output quality was sufficient for communication between the intended users; similar concerns have led to the task-based evaluations [58] carried out in other spoken language translation projects. In the case of the Diplomat system, there was an expectation that the system would eventually be field-tested with a US peace-keeping force, either in Bosnia or in Haiti.

Unfortunately, despite repeated efforts on the part of the researchers, a field test did not occur during the Diplomat project proper. However, such a field test was part of a follow-on project called Tongues, sponsored by the US Army Chaplains School. The Tongues project [52, 59] was led by Lockheed Martin, but was essentially an extension of the spoken language translation technology developed in the Diplomat project, with Carnegie Mellon continuing the development of the same software base as a sub-contractor. The US Army program under which Tongues was funded consisted of one year of development followed by a field test. Systems that passed the field test would be acquired and deployed for use by the Army.

There were only a few significant technical differences between Diplomat and Tongues. The platform was required to fit into the large “cargo pocket” of military fatigues, rather than being wearable. This allowed the use of the smallest commercial laptop available at the time (a 200 MHz Toshiba Libretto with 192 MB RAM), providing significantly more processing power. There were two significant software differences. First, Lockheed Martin developed a new, robust GUI addressing some issues that had been observed during Diplomat. For example, the GUI included buttons for playing pre-recorded information, both for introductory explanations and for emergency situations (“Put down the gun!”). The second significant software change was the replacement of the locally developed speech synthesizer with a lightweight version of the system it had been modeled on, FestVox [60]. The foreign language chosen was, fortuitously, Croatian, and the intended application domain was essentially a sub-domain of the Diplomat one, refugee relief. English speech and vocabulary data was acquired from chaplains by recording role-playing exercises in which they re-enacted actual encounters they had had in the field. This data was translated into Croatian and then recorded by native Croatian speakers to provide Croatian speech data and additional, relevant parallel texts for further MT development.

After one calendar year of prototype development, a field test was carried out in Zagreb, Croatia, in April 2001. It was important for this test to be very realistic, due to the possibility of real adoption of the technology if the test was considered successful. The English speakers were actual US Army chaplains; the 21 Croatian speakers were naïve native Croatian speakers, with no advance knowledge of the technology or test situation. Since the Croatians were not actual refugees, they were given a variety of refugee scenarios (printed in Croatian) to enact. The test was hosted by the University of Zagreb.

Qualitatively, approximately one half of the dialogues using the system seemed to be successful. The MT was found to be the weak link in the test; the speech technology worked surprisingly well, considering the small amount of somewhat unnatural training data used. The largest single source of difficulties was the need to provide feedback to the monolingual English user concerning the accuracy of the output translation. Since spoken language translation systems often make errors, most of these systems try to provide some feedback about what translation was produced. An important general issue regarding MT technology is that, like SMT, EBMT does not produce any intermediate internal representation of the input; it maps fairly directly from input strings to output strings. In contrast, the much more labor-intensive KBMT approach produces an internal representation. One major side-benefit of creating this internal representation is that a reliable paraphrase can be generated *back* into the source language, allowing the speaker to know the meaning that the system will attempt to synthesize in the other language. Systems without an internal representation often resort, as Tongues did, to creating a second, independent back-translation of their output from the target language back into the source language. The problem this creates is that the translation error rate is essentially doubled for the back-translation. In fact, during the Tongues test, a team member fluent in Croatian indicated that many translations that were abandoned by the users due to incomprehensible back-translations were actually reasonably correct in the forward direction.

Another serious issue that has also been noted with other systems was the overall slowness of the communication. Part of this slowness was simply due to the necessary repetition of each utterance in both languages; part was due to the time required by the recognition/translation/synthesis processing; and another part was due to the time taken up by having to repeat canceled translations. Future tests would do well to compare the system speed to a baseline of communication time using a human interpreter, since even expert human interpreters will add significant time to the communication process. On the other hand, the slowness of the interaction style may actually have improved the quality of the speech recognition, by making the speech less dysfluent than in normal spontaneous speech.

The formal report by the US Army Chaplains concluded that this technology was worth further development, but was not yet ready for field deployment. While this conclusion was correct, the level of performance achieved in the Tongues system was quite good, given its time and resource constraints, and the realism of the field test. The Tongues system set a new standard for realism in testing spoken language translation, influencing subsequent tests of such technology.

10.3.4 S-MINDS

Since 2002, Fluential has been developing a speech translation system called S-MINDS¹ with a hybrid architecture (Fig. 10.3) that combines multiple ASR engines and multiple translation engines. This approach only slightly increases the development cost of new translation applications, but greatly improves the accuracy and the coverage of the system by leveraging the strengths of both statistical and grammar-/rule-based systems. Further, this hybrid approach enables rapid integration of new speech recognition and translation engines as they become available. S-MINDS has a modular architecture with the components described below.

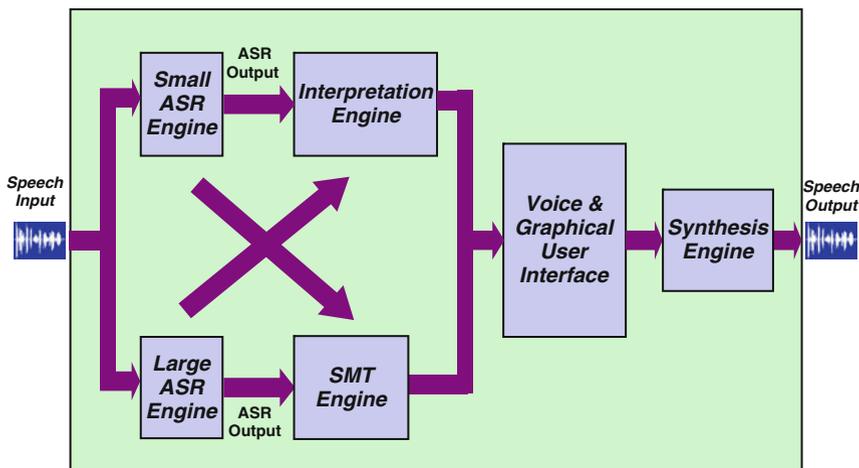


Fig. 10.3 Component architecture of the S-MINDS system

10.3.4.1 ASR Component

S-MINDS employs multiple recognition components within each language, with two separate language models active at the same time. A smaller, more directed language model with higher accuracy is used to capture important and frequently used concepts, or responses in situations which heavily constrain the range of possible utterances (e.g., the answer to a question – “how are you”). For less frequently used concepts, a larger language model is used, that has broader coverage but somewhat lower accuracy. The combination of these two models provides high accuracy for responses that can readily be anticipated and slightly lower accuracy but broader coverage for everything else. This method also allows development of new domains with very little data, since only the domain-specific small language model needs to be built for a given new domain. The S-MINDS system by default performs

¹Speaking Multilingual Interactive Natural Dialog System

N-best/lattice processing in the ASR as well as the MT components, but also can operate in 1-best mode.

10.3.4.2 Translation Component

S-MINDS has three different translation engines. Depending on the language, the domain, and the availability of data and linguistic expertise, one or more of these translation engines is used in either direction.

The first engine is a very simple rule-based architecture. Its rules are built by first having a non-expert bilingual speaker semantically tag words and concepts in a parallel corpus. These rules are used to create a robust surface parser, where each rule is tagged with the corresponding aligned translation in the target language. When multiple rule-sets are available, S-MINDS allows them to be applied at runtime either in series or in parallel. The series method applies each rule-set to each recognition result sequentially, in order of highest to lowest precision. The parallel approach applies all rule-sets simultaneously; if any apply, the system selects the best processing result based on a voting scheme. This engine has the advantage of requiring very little data; the simple structure of the rules, compared for example to those in the SLT or MedSLT systems described in Section 10.3.1, also means that they are highly accurate, since there are few ways in which rules can combine unexpectedly. The downside is that the rule-sets can be sensitive to certain types of speech recognition errors such as word insertion errors. These can cause multiple rules to be activated, after which the under-constrained translation algorithm often has no way to prefer the correct rule.

The second engine is a paraphrase translation engine that is primarily used in the English to second-language direction. This engine works by extracting an abstract semantic representation from the speech recognition, and then performing a paraphrase translation. This process is similar to what human interpreters do when they convey the essential meaning without providing a literal translation. The engine we have implemented performs information extraction using a Support Vector Machine (SVM) algorithm [61, 62]; other classification approach, such as K-means or Neural Net would also have been feasible. In the current system, the SVM-based classification effort has been focussed on extraction of robust linguistic features (lexical classes, part-of-speech tags), which is combined with statistical features (n-gram-based language models, and bag-of-words). Experiments show that the hybrid feature set, containing a mixture of surface-form and extracted-lexical classes, is a good tradeoff in terms of robustness and accuracy.

The advantage of an interpretation/paraphrase engine of this kind is that new domains can be added more quickly and with perhaps an order of magnitude less data than is possible with an SMT engine. For high-volume, routine interactions, a paraphrase interpretation engine can be extremely fast and accurate, with only a few examples of each concept needed to perform adequate training. The downside is that coverage is constrained by a knowledge engineering bottleneck related to the number of implemented concepts (currently only in thousands). Another problem is that the paraphrase-style translation may lose important nuances.

The third engine is a statistical translation engine of the kind described in Section 10.2.2. The S-MINDS SMT engine capitalizes on the intuition that language is broadly divided into two levels: structure and vocabulary. Traditional statistical approaches force the system to learn both types of information simultaneously. However, if the acquisition of structural information is decoupled from the acquisition of vocabulary, the resulting system is able to learn both levels more efficiently. In addition, by modifying the existing corpus to separate structure and vocabulary, the developer is able to take full advantage of all the information in the bilingual corpus, again producing higher-quality MT without requiring large bodies of training data.

This idea has been implemented using a syntactic chunk-based, two-level Machine Translation algorithm, which learns vocabulary translations within syntactically and semantically independent phrase units, and separately acquires global structural relationships at the phrase chunk level. Syntactic chunking is a relatively simple process, as compared to deep parsing. It only segments a sentence into syntactic phrases such as noun phrases, prepositional phrases, and verbal clusters without hierarchical relationships between phrases. These chunks are aligned using a parallel corpus, and augment the existing statistical alignment table using different weighing criteria. A phrasal re-ordering component that re-orders the chunks before they are translated enables the system to make the phrasal structure for the input language very similar to the output language.

This sub-system is still under active development. Recent improvements include the use of distance-based ordering [63] and lexicalized ordering [64] to allow for multiple language models, including non-word models such as part-of-speech improved search algorithm, in order to improve its speed and efficiency.

10.3.4.3 N-Best Merging of Results

The merge module's task is to extract a single optimal translation from the multiple N-best recognition-translation pairs produced when the different translation engines process the multiple recognition hypotheses arising from speech recognition. Note that the SLT-2 and SLT-3 systems described in Section 10.3.1 contained a similar module. Each pair includes the associated recognition confidence and translation confidence scores. Based on these scores, the merge algorithm ranks all pairs and produces an ordered list of pairs. The component scores are of several different kinds. ASR confidence values reflect the likelihood that the given acoustic input could have produced a specified word sequence, and combine probabilities derived both from acoustic models and language models. The robust parser scores are binary quantities (match or no-match). The interpretation engine score is another composite measure, which combines the classifier mapping score as well as other values which are correlated with the expected reliability of a template application, including an algorithm's a priori likelihood of producing correct output and the number of words/classes of the input sentence that the template covers. The SMT engine score is as usual based on the Bayesian translation probability

$$\operatorname{argmax} P(e|f) = \operatorname{argmax} P(e)P(f|e)$$

which is composed from the language model score of the target language $P(e)$ and the translation model score $P(f|e)$. All these different types of scores (ASR and the various MT engine scores) are combined using the trainable Powell search algorithm [65].

10.3.4.4 User Interface Component

S-MINDS has a flexible user interface that can be configured to use VUI only or a combination of VUI and GUI for either the English speaker or the second-language speaker. (It is possible to configure the system so that the English speaker and the second-language speaker use different types of interface). Speech recognition can be confirmed by VUI, GUI, or both, and it can be configured to verify all utterances, no utterances, or just utterances that fall below a certain confidence level. The system can be deployed on multiple types of hardware platforms, and can use multiple types of microphones, including open microphones, headsets, and telephone headsets.

10.3.4.5 Speech Synthesis Component

By default, S-MINDS uses Cepstral's text-to-speech engine [27] throughout to produce spoken output. For some domains, however, it is possible to attain better performance by using a more elaborate generation strategy which combines TTS for translations produced by the SMT-based system and spliced-together chunk-based recordings for translations produced by the interpretation/paraphrase engine. This strategy becomes increasingly attractive as the interpretation/paraphrase engine's use increases in frequency and the chunks it produces become larger.

10.3.4.6 Evaluations

Slightly different versions of this system have been evaluated in various settings. Many of the design decisions that we have described above have been based on the results of these evaluations. The following presents a rough summary of a typical evaluation of this kind, in this case for a bidirectional English-to-Arabic system. The evaluation was conducted over the course of one week in July 2007 at NIST and involved around 15 English and 15 Arabic speakers, who produced a total of about 20 live scenarios of 10 min length each. Our Arabic vocabulary size was about 77,000 words and the English vocabulary size was 28,000 words. The topics covered were various military and medical interactions, assumed to be between an American service member and an Iraqi native speaker [66].

A bilingual Arabic speaker closely examined and analyzed 109 utterances of the total set, representing about 15% of the total evaluation data. The goal was to determine where the system made the most mistakes, rather than to measure the exact translation accuracy; as such, the priority was to analyze what kinds of errors the system made and where future effort could be most productively focused. Of the

utterances examined, about 65% were correctly recognized and translated. On the remaining set, the most frequent cause of error – about 49% of the total – was ASR homonyms in Arabic. This high frequency was due to the fact that Arabic has a rich morphology, with many words only one phoneme apart. The second largest group, roughly 17%, was caused by incorrect word sense assignment. The third largest, about 8%, consisted of problems due to software issues, including user errors and user misunderstandings of what had been said. The remaining 26% belonged to many different categories, with no dominant trends. Note that in this analysis, the error is always assigned to the first component; thus for example, if both an ASR error and a machine translation error occurred, the fault was only assigned to the ASR engine and not the machine translation engine.

10.4 Further Directions

In this chapter, we have reviewed the component technologies used by speech translation systems, the history of the subject, and several representative systems. With speech translation technology now about 20 years old, a measure of agreement seems to be developing concerning methods and architectures. Nearly all groups are now using HMM-based recognition and concatenative speech synthesis, and there is a strong consensus about the importance of being able to deploy systems on appropriate platforms: the advantages of highly portable/wearable systems are obvious, and Moore's Law has now reached the point where very substantial applications can readily be loaded onto machines weighing less than a kilogram. Agreement is not as strong on choice of translation engine, but the trend appears to be toward hybrid methods which combine both data-driven and rule-based aspects. An important question that needs to be resolved here, though, is how best to reconcile the two approaches to translation which so far have been most popular: knowledge-based interlingua on the one hand, and statistical machine translation on the other.

Another key question is the choice of domains on which to concentrate development work. In the United States, political, financial, and resources considerations have forced research groups to focus largely on military systems. Even leaving aside the ethical aspects, it is, however, far from clear that military domains are the most promising ones. Certainly, military funding has been a very important seed source for technological development for everything from the Internet to speech recognition engines; however, in almost every case, military R&D funding has been exceeded and often dwarfed by spending in the commercial market.

In particular, the authors of this chapter believe that medical applications are going to be one of the key commercial markets that will enable mass adaptation of this technology. Above and beyond commercial rewards, there are many advantages to building systems in the medical domain. For example, one of the main practical problems associated with development of military systems is that there is no clearly defined "military" or "force protection" domain. (According to one source, it is "everything that a member of the armed forces has to do"). In contrast, many

medical domains, in particular doctor–patient examination dialogues, are tightly specified, and can be covered well with vocabularies of only a few hundred to at most a couple of thousand words. There is also a large demand for automatic medical speech translation services; in the United States, studies like [67] suggest that as many as half of all Limited English Proficiency patients presenting to emergency departments are not provided with a medical interpreter. It would not be surprising if the combination of these facts acted over the next few years to push speech translation technology away from military domains, and toward medical ones.

References

1. Adams, D. (1979). *The Hitchhiker's Guide to the Galaxy*, London: Pan Books.
2. Levinson, S., Liberman, M. (1981). Speech recognition by computer. *Sci. Am.*, 64–76.
3. Weinstein, C., McCandless, S., Mondshein, L., Zue, V. (1975). A system for acoustic-phonetic analysis of continuous speech. *IEEE Trans. Acoust. Speech Signal Process.*, 54–67.
4. Bernstein, J., Franco, H. (1996). Speech recognition by computer. In: *Principles of Experimental Phonetics*, St. Louis: Mosby, 408–434.
5. Young, S. (1996). A review of large-vocabulary continuous-speech recognition. *IEEE Signal Process. Mag.*, 45–57.
6. Ehsani, F., Knodt, E. (1998). Speech technology in computer-aided language learning: strengths and limitations of a new CALL paradigm. *Language Learning and Technology*, 2, 45–60. Available online, February 2010: <http://lt.msu.edu/vol2num1/article3/index.html>.
7. Deng, L., Huang, X. (2004). Challenges in adopting speech recognition. *Commun. ACM*, (47-1), 69–75.
8. Nyberg, E., Mitamura, T. (1992). The KANT system: fast, accurate, high-quality translation in practical domains. In: *Proc. 14th Conf. on Computational Linguistics*, Nantes, France.
9. Cavalli-Sforza, V., Czuba, K., Mitamura, T., Nyberg, E. (2000). Challenges in adapting an interlingua for bidirectional english-italian translation. In: *Proc. 4th Conf. Assoc. Machine Translation in the Americas on Envisioning Machine Translation in the Information Future*, 169–178.
10. Somers, H. (1999). Review article: Example-based machine translation. *Machine Translation*, 113–157.
11. Brown, R. (1996). Example-based machine translation in the pangloss system. In: *Proc. 16th Int. Conf. on Computational Linguistics (COLING-96)*, Copenhagen, Denmark.
12. Trujillo, A. (1999). *Translation engines: Techniques for machine translation*. London: Springer.
13. Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., Roossin, P. (1990). A statistical approach to machine translation, *Comput. Linguistics*, 16(2), 79–85.
14. Berger, A., Della Pietra, V., Della Pietra, S. (1996). A maximum entropy approach to natural language processing. *Comput. Linguistics*, 22(1), 39–71.
15. Brown, R., Frederking, R. (1995). Applying statistical English language modeling to symbolic machine translation. In: *Proc. 6th Int. Conf. on Theoretical and Methodological Issues in Machine Translation (TMI-95)*: Leuven, Belgium, 221–239.
16. Knight, K. (1999). A statistical MT tutorial workbook. Unpublished. Available online, May 2010: <http://www.isi.edu/natural-language/mt/wkbk.rtf>.
17. Koehn, P., Knight, K. (2001). Knowledge sources for word-level translation models. In: *Proc. EMNLP 2001 Conf. on Empirical Methods in Natural Language Processing*, Pittsburgh, PA, 27–35.
18. Brown, R. (1999). Adding linguistic knowledge to a lexical example-based translation system. In: *Proc. TMI-99*, Chester, England.

19. Yamada, K., Knight, K. (2001). A syntax-based statistical translation model. In: Proc. 39th Annual Meeting on Association for Computational Linguistics, Toulouse, France, 523–530.
20. Alshawi, H., Douglas, S., Bangalore, S. (2000). Learning dependency translation models as collections of finite-state head transducers. *Comput. Linguistics*, 26(1), 45–60.
21. Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguistics*, 23(3), 377–403.
22. Wang, Y. (1998). Grammar inference and statistical machine translation. Ph.D. thesis, Carnegie Mellon University.
23. Och, F., Tillmann, J., Ney, H. (1999). Improved alignment models for statistical machine translation. In: Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora: University of Maryland, College Park, MD, 20–28.
24. Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., Thayer, I. (2006). Scalable inference and training of context-rich syntactic translation models. In: Proc. 21st Int. Conf. on Computational Linguistics, Sydney, 961–968.
25. Venugopal, A. (2007). Hierarchical and Syntax Structured Models, MT Marathon, Edinburgh, Scotland.
26. Chiang, D. (2007). Hierarchical phrase-based translation. *Assoc. Comput. Linguistics*, 33(2), 201–228.
27. Schultz, T., Black, A. (2006). Challenges with rapid adaptation of speech translation systems to new language pairs. In: Proc. ICASSP2006, Toulouse, France.
28. Waibel, A. (1996). Interactive translation of conversational speech. *Computer*, 29(7), 41–48.
29. Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita, T., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., Ward, W. (1993). Recent advances in JANUS: A speech translation system. In: Proc. Workshop on Human Language Technology, Princeton, NJ.
30. Wahlster, W. (2002). *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, Berlin.
31. Rayner, M. H., Alshawi, I., Bretan, D., Carter, V., Digalakis, B., Gambck, J., Kaja, J., Karlgren, B., Lyberg, P., Price, S., Pulman, S., Samuelsson, C. (1993). A speech to speech translation system built from standard components. In: Proc. 1993 ARPA workshop on Human Language Technology, Princeton, NJ.
32. Rayner, M., Carter, D. (1997). Hybrid language processing in the spoken language translator. In: Proc. ICASSP'97, Munich, Germany.
33. Rayner, M., Carter, D., Bouillon, P., Wiren, M., Digalakis, V. (2000). *The Spoken Language Translator*, Cambridge University Press, Cambridge.
34. Isotani, R., Yamabana, K., Ando, S., Hanazawa, K., Ishikawa, S., Iso, K. (2003). Speech-to-Speech Translation Software on PDAs for Travel Conversation. NEC Research and Development.
35. Yasuda, K., Sugaya, F., Toshiyuki, T., Seichi, Y., Masuzo, Y. (2003). An automatic evaluation method of translation quality using translation answer candidates queried from a parallel corpus. In: Proc. Machine Translation Summit VIII, 373–378. Santiago de Compostela, Spain.
36. Metz, F., McDonough, J., Soltau, H., Waibel, A., Lavie, A., Burger, S., Langley, C., Laskowski, K., Levin, L., Schultz, T., Pianesi, F., Cattoni, R., Lazzari, G., Mana, N., Pianta, E., Besacier, L., Blanchon, H., Vaufraydaz, D., Taddei, L. (2002). The NESPOLE! Speech-to-speech translation system. In: Proc. HLT 2002, San Diego, CA.
37. Bangalore, S., Riccardi, G. (2000). Stochastic finite-state models for spoken language machine translation. In: NAACL-ANLP 2000 Workshop on Embedded Machine Translation Systems, Seattle, WA, 52–59.
38. Zhang, Y. (2003). Survey of Current Speech Translation Research. Unpublished. Available online, May 2010: <http://projectile.sv.cmu.edu/research/public/talks/speechTranslation/sst-survey-joy.pdf>

39. Agnas, M. S., Alshawi, H., Bretan, I., Carter, D. M., Ceder, K., Collins, M., Crouch, R., Digalakis, V., Ekholm, B., Gambäck, B., Kaja, J., Karlgren, J., Lyberg, B., Price, P., Pulman, S., Rayner, M., Samuelsson, C., Svensson, T. (1994). Spoken language translator: first year report. SRI Technical Report CRC-043.
40. Digalakis, V., Monaco, P. (1996). Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognizers. *IEEE Trans. Speech Audio Process.*, 4(4), 281–289.
41. Alshawi, H. (1992). *The Core Language Engine*. MIT Press, Cambridge, MA.
42. Alshawi, H., van Eijck, J. (1989). Logical forms in the core language engine. In: *Proc. 27th Annual Meeting on Association for Computational Linguistics*, Vancouver, British Columbia, Canada, 25–32.
43. Alshawi, H., Carter, D. (1994). Training and scaling preference functions for disambiguation. *Comput. Linguistics*, 20(4), 635–648.
44. Rayner, M., Samuelsson, C. (1994). Grammar Specialisation. In: [39], 39–52.
45. Samuelsson, C. (1994). *Fast natural-language parsing using explanation-based learning*. PhD thesis, Royal College of Technology, Stockholm, Sweden.
46. Frederking, R., Nirenburg, S. (1994). Three heads are better than one. In: *Proc. 4th Conf. on Applied Natural Language Processing*, Stuttgart, Germany.
47. Rayner, M., Bouillon, P. (2002). A flexible speech to speech phrasebook translator. In: *Proc. ACL Workshop on Speech-to-Speech Translation*, Philadelphia, PA.
48. Rayner, M., Hockey, B. A., Bouillon, P. (2006). Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler. CSLI Press, Stanford, CA.
49. Rayner, M., Bouillon, P., Santaholma, M., Nakao, Y. (2005). Representational and architectural issues in a limited-domain medical speech translator. In: *Proc. TALN 2005*, Dourdan, France.
50. Chatzichrisafis, N., Bouillon, P., Rayner, M., Santaholma, M., Starlander, M., Hockey, B. A. (2006). Evaluating task performance for a unidirectional controlled language medical speech translation system. In: *Proc. 1st Int. Workshop on Medical Speech Translation, HLT-NAACL*, New York, NY.
51. Sarich, A. (2004). Development and fielding of the phraselator phrase translation system. In: *Proc. 26th Conf. on Translating and the Computer*, London.
52. Frederking, R., Rudnicky, A., Hogan, C., Lenzo, K. (2000). Interactive speech translation in the Diplomat project. *Machine Translation J., Special Issue on Spoken Language Translation*, 15(1–2), 27–42.
53. Huang X., Alleva F., Hon H. W., Hwang K. F., Lee M. Y., Rosenfeld R. (1993). The SPHINX-II Speech Recognition System: An overview. *Comput. Speech Lang.*, 2, 137–148.
54. Lenzo, K., Hogan, C., Allen, J. (1998). Rapid-deployment text-to-speech in the DIPLOMAT system. In: *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP-98)*, Sydney, Australia.
55. Frederking, R., Brown, R. (1996). The Pangloss-Lite machine translation system. In: *Proc. Conf. Assoc. for Machine Translation in the Americas (AMTA)*.
56. Nielsen, J. (1993). *Usability Engineering*. AP Professional, Boston, MA.
57. Rudnicky, A. (1995). Language modeling with limited domain data. In: *Proc. ARPA Workshop on Spoken Language Technology*, Morgan Kaufmann, San Francisco, CA, 66–69.
58. Gates, D., Lavie, A., Levin, L., Waibel, A., Gavaldà, M., Mayfield, L., Woszczyna, M., Zhan, P. (1996). End-to-end evaluation in JANUS: A speech-to-speech translation system. In: *Workshop on Dialogue Processing in Spoken Language Systems. Lecture Notes in Computer Science*, Springer, Berlin.
59. Black, A., Brown, R., Frederking, R., Lenzo, K., Moody, J., Rudnicky, A., Singh, R., Steinbrecher, E. (2002). Rapid development of speech-to-speech translation systems. In: *Proc. ICSLP-2002*, Denver.
60. Black, A., Lenzo, K. (2000). Building voices in the festival speech synthesis system. Unpublished. Available online, May 2010: <http://www.festvox.org/festvox/index.html>.

61. Joachims, T. (2002). Learning to Classify Text Using Support Vector Machines. Dissertation, Kluwer.
62. Klinkenberg, R., Joachims, T. (2000). Detecting concept drift with support vector machines. In: Proc. 17th Int. Conf. on Machine Learning (ICML), Morgan Kaufmann, San Francisco, CA.
63. Zens, R., Ney, H. (2004). Improvements in phrase-based statistical machine translation. In: Proc. Human Language Technology Conf. (HLT-NAACL), Boston, MA, 257–264.
64. Tillmann, C., Zhang, T. (2005). A localized prediction model for statistical machine translation. In: Proc. 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, MI, 557–564.
65. Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (2007). Numerical Recipes 3rd Edition: The Art of Scientific Computing. Cambridge University Press, Cambridge.
66. Schlenoff et al. (2007). Transtac July 2007 Evaluation Report, NIST Internal Document. Published in September 2007.
67. Baker, D. W., Parker, R. M., Williams, M. V., Coates, W. C., Pitkin, K. (1996). Use and effectiveness of interpreters in an emergency department. *JAMA*, 275, 783–788.

Chapter 11

Application of Speech Technology in Vehicles

Fang Chen, Ing-Marie Jonsson, Jessica Villing, and Staffan Larsson

11.1 Introduction

Speech technology has been regarded as one of the most interesting technologies for operating in-vehicle information systems. Cameron [1] has pointed out that under at least one of the four criteria that people are using speech system more likely. These four criteria are the following: (1) They are offered no choice; (2) it corresponds to the privacy of their surroundings; (3) their hands or eyes are busy on another task; and (4) it is quicker than any other alternatives. For driver, driving is a typical “hands and eyes are busy” task. In most of the situations, the driver is the only person inside the car, or with some passengers who know each other well, so the “privacy of surroundings” criterion is also met. There are long histories of interests of applying speech technology into controlling in-vehicle information system. Up to now, some of the commercial cars have already equipped with imbedded speech technology. In 1996, however, the S-Class car of Mercedes-Benz introduced Linguatronic, the first generation of in-car speech system for anybody who drives a car [2]. Since then, the number of in-vehicle applications using speech technology is increasing [3].

Normally, the speech systems used in vehicle are imbedded solutions with either separate hardware box or integrated into infotainment systems running under a variety of real-time operating systems. The output of the system is using the loudspeakers of the car. Considering the noise environment (include driver’s listening to music or conversation with passengers, making phone calls, etc.), a push-to-talk button or lever to start the system is needed.

The concept of modern car is very different from before. Many new features have been introduced in cars, beginning from the 1980s and rapidly increasing in recent years. This means that a growing number of features need to be controlled, which leads to a growing number of buttons to be pushed and menus to be looked at. This takes the driver’s attention off the road. Speech interface applied to the driver is

F. Chen (✉)

Interaction Design, Department of Computer Science and Engineering, Chalmers University of Technology, Sweden
e-mail: fanch@chalmers.se

considered to increase the safety on the roads. Speech technology would enable for the driver to keep the hands on the steering wheel and the eyes – and hopefully the attentions – on the road. It would also solve the space problem, as one only needs to have a good noise cancellation microphone and loudspeakers which are already installed inside vehicles. These advantages are the main reasons why the vehicle industry is experiencing a fast-growing interest in speech technology. Safety has been an important argument for all brands; it is no longer an issue only for a few exclusive manufacturers. However, many engineers and researchers do not consider speech technology to be mature enough for the application in drive environment. Factors, such as the intrinsically noisy vehicle environment and the often distracted and stressed users, with passengers' speech distraction, make this into a challenging area. Market forces, both on the demand and on the supply side, are however busy making sure there are deployed systems such as Pioneers In-car Navigation system, SoundClear from Acoustic Technologies, and Linguatronic.

In recent years, continuous speech recognition technology for the vehicle environment has progressively been improved in terms of robustness and flexibility. The increasing number of microprocessors and decreasing costs are factors that enable new technologies to be introduced to the car [3]. The first microprocessors designed for engine control appeared in the early 1980s. Since then the number of microprocessors has increased to about 15–25 in an ordinary car and about 50–70 in a premium car. This development has enabled advanced in-vehicle techniques. There are also emerging demands for hands-free, safe, reliable, and easy-to-use systems that allow the driver to control devices such as the onboard computer, RDS-tuner or air-conditioner, global positioning and wireless communication systems, remote information services access, and Web browsing by voice [4]. To remove some buttons and instead control of the device using speech would be cheaper for the manufacturers, and hence speech technology is beginning to be a business case for the car industry.

Speech recognition technology can be used to control different functions inside the vehicle [3]. Spoken dialogue systems enable voice communication between driver and vehicle. Synthesis speech can be used to read long text messages to the driver. Emotional speech can be used to manipulate the driver's stress level, etc. and enhance the safety drive [5, 6]. In each of these in-vehicle application areas, researchers and designers are facing different challenges, such as to meet the differences of driver's expectation and driver experiences and the usability requirements of speech interaction design.

In this chapter we provide a brief review of ongoing research and development and indicate potential trends for future researches.

11.2 Complicated Vehicle Information Systems

There are huge differences between today's and future transportation system (the so-called intelligent transportation systems, ITS). The transportation infrastructure will change to provide more traffic-related information to the driver, such as road

condition, traffic situations, and possible services. The ITS network will most likely provide connections and communication between vehicles and the roadside infrastructure. Different intelligent systems, so-called Advanced Drive Assistance System (ADAS) [7], will be installed in the vehicles. Some of them are already available in the market, such as adaptive cruise control, lane departure warning, forward distance alert, blind spot information, diver alert system, and night vision. These systems are designed to help the driver in various ways: prevent from potential accidents by providing traffic information, evaluate driver performance, warn the driver of potentially dangerous situations, and in some cases also cease driver's control of the vehicle or part of the vehicle. This will make the relationship between driver and vehicle very different from what we have today. Drivers may in addition to operating the vehicle also communicate with these additional information systems. This could come in handy to warn other drivers about upcoming road hazards or perceived dangerous drivers, or select the best route [8].

The infotainment systems in vehicle will become more and more complex. In addition to providing access to information in the vehicle, the systems will offer internet connections and other functions that personal devices may offer [9]. This opens up a new plethora of functions and potential information access from the vehicle. Basically, whatever one can do in his home or office can be done in vehicle in the near future.

The number of portable devices that are used by the driver for supporting, assisting, communicating, or entertaining while driving is increasing. The most common used portable device so far is the mobile phone, with or without hands-free kit. Navigation systems and especially PDA-based navigation systems provide a rapidly growing market due to falling prices. These systems that used to be offered in speciality stores are now often sold in discount supermarkets or bundled with new vehicles. Last but not least, music players are becoming more common in vehicles, where drivers carry their MP3 players or Apple iPod's between home, car, and working or study place.

With the increased complexity of integrating ADAS, IVIS, and portable devices, and even the Internet access to the vehicle, important issues such as driver safety, human-machine interaction (HMI), and integration of devices have to be considered. To keep driver's eyes and minds on the road and their hands on the steering wheel are the essential requirements for driving safely. Today, most of the IVIS and portable devices that are used inside vehicles have the potential of competing for the driver's attention, re-focusing attention from the driving task to information interaction.

The primary task for a driver is driving, any other activity performed by the driver while driving is, and must be, regarded as a secondary task. Driver distraction is generally defined as a driver performing a secondary task. The course of a distraction can be due to (1) performing a secondary task with hands, moving the hands from the steering wheel, (2) shifting the driver's focus from the road to the information device, and (3) the secondary task is more compelling than driving, causing full secondary task focus. These secondary tasks can be physical, visual, and cognitive [10], and range from reading, eating, putting on makeup, or interacting with unsuitable or poorly located information devices. The automotive industry focuses attention

on speech technology in cars. Special focus is on in-vehicle systems using ASR (automatic speech recognition) instead of manual input and TTS (text-to-speech) technologies instead of visual display. From a safety perspective this would reduce the necessity of physical manipulation and re-focus of visual attention for interaction. It is, however, less clear that speech interaction can reduce the cognitive distraction over physical/visual interactions. Speech is a transient media, once the word has been spoken they are gone, which makes heavy demands on the system output being easy to understand and remembered. Without proper visual display, speech navigation into a complicated manual may add heavy working memory load to the driver. Today, most high-end vehicles come with speech-based systems for climate control, radio control, voice dialing, telematics services, traffic information, and navigation. Drivers who purchase high-end vehicles generally expect high performance, and speech interfaces that are built into the car are still exception. The ultimate goal is to provide voice interfaces that approach human-to-human interaction. Talk to your car the way you want, but focus on driving.

The challenges of applying speech technology into vehicle can be summarized as:

1. Fluid situation for driver: The environment outside the car and inside the vehicle is constantly changing. Traffic patterns and traffic events are dynamic and vary over time. The environment inside the car changes with passengers and use of IVIS and other information systems or devices. These may affect the driver's emotional state and hence influences both speech patterns and language. It requires the speech technology to be robustness to handle noisy environments (AC, fan, traffic, passengers), driver's emotional state, varying speech rate, amplitude, and multiple speakers.
2. Cognitive load: Driving safety is always the driver's first priority; therefore, the task performed using speech technology is always going to be a secondary task. With inferior design, it may increase the mental workload of the driver and hence compete with cognitive processes needed for driving [11].
3. The complexity of the task: to control/navigate among multiple devices and information systems using an adaptive integrated driver-vehicle interface requires the speech system to be highly natural, robust, seamless, and highly intuitive and easy-to-use. Natural speech requires a system that is robust, both regarding ASR and NLU (Natural Language Understanding), concerning different mode of expression.
4. Matching the emotional state between the driver and the voice of the car [5].

11.3 Driver Distraction Due to Speech Interaction

Many studies have evaluated driver distraction due to secondary task performance. Baron [12] reviewed 15 paper and summarized the human factors literature on the use of speech interfaces for tasks such as music selection, email processing, dialing, and destination entry while driving. Most papers they reviewed focused on

identifying differences between the speech and manual input modality from the viewpoint of safety and driver distraction. They concluded that people generally drove at least as well, if not better (less lane variation, speed was steadier), when using speech interfaces compared to manual interfaces, but using a speech interface was often worse than just driving. Speech interfaces led to less workload than manual interfaces and reduced eyes-off-the-road time, all pro-safety findings. Task completion time was less with speech interfaces, but not always (as in the case of manual phone dialing). Missing from the literature were firm conclusions about how the speech/manual recommendation varies with driving workload, recognizer accuracy, and driver age [12].

Lee et al. [11] studied the effect of using an in-vehicle email device with 100% simulated speech recognition accuracy on driver braking performance in a driving simulator. Self-paced use of the speech recognition system was found to affect braking response time with 30% increase in the time it took for drivers to react to an intermittently braking lead vehicle. This demonstrated that speech-based interaction with an in-vehicle device increases the cognitive load on the driver. Similar results are found when compared using hand-held phone or hands-free phone [13]. Treffner's study [14] in real driving (not in simulator) confirmed that conversing on a mobile phone, regardless of conversation type, will detract from a driver's ability to control a vehicle compared to when driving in silence. It did not matter if the conversation is simple or complex, as speaking on a hands-free mobile phone while driving may significantly degrade critical components of the perception-action cycle. On the contrary, a study carried out by [15] revealed that drivers in fact adapt their dialogue to the traffic situation. They signal in different ways when they need to make a pause in the dialogue in order to concentrate on the traffic, and when it is safe to start speaking again.

Keeping in mind that speech-based interactions with a secondary task while driving always lead to driver distraction and decrease driving performance, new speech-based IVIS interface must be thoroughly tested before application. These tests should take place in a driving environment, initially in a driving simulator, but ultimately in real-time real-world driving to ensure ecological validity of how the interaction affects the driving task.

When evaluating interactions with infotainment systems in simulators, the fidelity level of the simulator as well as performance and attitudinal measures should be considered carefully [16]. There are yet no standard methods that indicate how to measure the usability of an interactive speech-based in-vehicle system. It is an urgent requirement to develop methods that take into consideration of evaluating (1) the driver's mental workload, (2) the distraction caused by the interactions with the system, (3) how traffic interacts with the use of the system (for example, in busy traffic with many road users, shall we remind the user of some kind of potential danger of using the system [17]), and (4) how passengers interact with the use of the system. This would include guidelines as to which performance and attitudinal measures to use when testing interactions with speech-based IVIS. The evaluation can take place either on a real road drive or in a driving simulator. It is, however, important to be aware of the fact that the driving behavior differ between a real road drive and a driving simulator, which should be taken under consideration (see, e.g.,

[12, 18, 19]). To be able to compare results and measures used in different studies, certain testing conditions should be standardized, such as the participant screening and description, the fidelity level of the simulator, the traffic scenarios, and the driving task.

Common methods for driving performance measure the longitudinal acceleration or velocity, steering wheel behavior or lane keeping displacement [12]. The driver's visual behavior during a driving session is normally measured using an eye-tracking system to measure the eye glance pattern [12, 20]. The driver's mental workload is normally evaluated by using the NASA-TLX method [21], or a modified version that specially developed for driving situation called DALI can also be used [22]. The problem with driving performance measurement is that different drivers may use different behavior strategies to cope with the distraction. Some of them may reduce speed, others may position the car close to the right side of the road for a larger safety margin, and some may combine both behaviors. This can make data analysis difficult, and the results may not reflect the true situation [23]. The mental workload can also change due to traffic or road conditions when drivers are engaged with a speech-based IVIS. During light traffic and easy road conditions, the driver may be able to use more resources to cope with the speech task than during high traffic situations. On the other hand, the speech task can potentially also keep the driver alert resulting in improved driving performance, even though the speech task increases the workload. It is important to note that drivers need to concentrate on the driving task if, e.g., the traffic is heavy, the road is demanding, or the weather is bad. During these conditions, even simple speech tasks may significantly increase the mental workload and result in decreased driving performance. Different sub tasks of speech interactions may also impact driving performance differently. It is therefore important to investigate how different speech tasks impact driving performance and overall driving experience. In addition to the NASA-TLX or DALI, it might also be necessary to develop special methods, tailor-made to continually measure workload for drivers [24, 25].

The trend is to equip new vehicles with more and more ADAS systems. The information from ADAS is either presented by visual or by auditory displays. As these systems assume to improve driving safety, the drivers' behavior and delegation of control of driving task to the vehicle may change. The challenge is hence to keep up with a moving target on how to evaluate the usability of and distraction from an interactive speech system in a dynamically changed in-vehicle environment. There are yet no research have (1) results from studies on combinations of ADAS and speech interactions, and (2) no proposed methodology for this type of studies.

11.4 Speech as Input/Output Device

11.4.1 Noise Inside Vehicles

The use of speech recognition and speech synthesis to operate in-vehicle systems has the potential to provide head-up and hands-off interactions while driving.

Speech recognition errors, however, are greatly increased by the noisy in-vehicle environment, where sounds originate both from inside and outside the vehicle. Changes in speech patterns and inflections, due to the driver's workload, stress and emotional state further reduces the speech recognition accuracy. Correcting speech recognition errors is both irritating and requires mental resources. This is also complicated by the fact that comprehension of a synthetic message requires more mental effort than comprehension of a spoken message [26].

Noise from the engine, air conditioner, wind, music, echoes, etc. makes the signal-to-noise ratio (SNR) of the speech signal relatively low. Separating the driver's speech from background noise is further complicated in the presence of other voices such as passengers talking, baby's crying, children screaming, and by sounds from passenger activities such as movie and mobile games. It is hard to find reliable patterns that indicate a particular speaker in these types of environments, and placing the microphone close to the driver's mouth (headset) is not an option. At present time, this renders the speech recognition performance to be unacceptable in most cases.

There is much ongoing work to remedy the situation, and techniques to improve the performance of speech recognition in vehicles can be divided into three main categories, listed as follows:

Robust speech feature extraction: work in this field is focused on the task to find a set of parameters or similarity measures that are robust to variations in the speech signal caused by noise. The techniques developed in the area can be found as following:

- RASTA filtering [27]
- cepstral mean normalization (CMN) [28]
- use of dynamic spectral features [29]
- short-time modified coherence (SMC) [30]
- one-sided autocorrelation LPC (OSALPC) [31]
- differential power spectrum (DPS) [32]
- relative auto-correlation sequence (RAS) [33, 34]

Speech enhancement: for solutions in this field, information about speech and noise is needed. Researchers are working on the following techniques:

- Spectral subtraction (SS) [35]
- Wiener filtering [36]

Model-based compensation for noise: work in this field is focused on different speech models such as hidden Markov models (HMMs). The aim is to use compensation techniques to remove the mismatch between the trained models and the noisy speech to improve the recognition rate. Some of the methods are:

- Parallel model combination (PMC) [37, 38]
- Vector Taylor series (VTS) [39, 40]
- weighted projection measure (WPM) [30]

Recently, solutions using microphone arrays (multiple-microphones) configurations for speech processing combined with speech enhancement scheme have been presented. These technologies integrate spatial and temporal signal-processing methods for robust speech recognition in noisy environments and are going to play an ever-increasing role in multimedia systems and possibly in in-vehicle systems [41]. Generally, push-to-talk buttons are used which has two purposes; (1) forcing the driver to push the button for every utterance makes the driver aware of when the system is listening and when it is not (2) the system is only listening when the button is pushed which decreases the “listening time” and by that also the risk of perceiving additive noise.

Although all the aforementioned efforts were used in speech recognition tasks with certain levels of success, it is still necessary to investigate new algorithms to further improve the performance of ASR for practical applications. One way to improve the performance of ASR systems is to develop a new feature set for in-vehicle speech, since all processes in ASR systems are highly dependent on the quality of the extracted features [42]. At present, however, most of the speech enhancement systems are developed to solve certain kind of in-vehicle noise, and lack a broad coverage of different sources of noise. One common way to adapt the ASR system and compensate for the effects of unknown additive noise during the front-end processing is to assume that the first frames of the speech signal are noise only, and compute the average noise spectrum from that [43]. However, the speech recognition rate in vehicles is still too low due to noise inside the vehicle, and this presents an obstacle for the introduction of ASR applications in the vehicle.

It is important to understand that there will never be 100% accuracy for every user utterance, not even humans can achieve this. One thing that humans have, though, is the capability of fast error recognition and error recovery. The challenge is to make dialogue systems that can detect and handle misrecognition in a way that feels natural and minimally distracting for the driver. Adding linguistic knowledge to improve dialogue management will increase naturalness of system behavior. Ecologically, valid user tests of human dialogue can give clues to how humans handle misrecognition without repeatedly asking distracting clarification questions. So far, the studies have concentrated on other environments than the in-vehicle environment. The studies show that humans do not always have to hear every word correctly to be able to interpret their dialogue partner's utterances. When humans face speech recognition problems, a common strategy is to ask task-related questions that confirm their hypothesis about the situation instead of signaling non-understanding. Compared to other strategies, such as asking for a repetition, this strategy leads to better understanding of subsequent utterances, whereas signaling non-understanding leads to decreased experience of task success [44, 45]. The in-vehicle environment is crucially different from most other environments where a dialogue system is used. Here, interaction with the dialogue system is a secondary task – while driving always must be the primary task – therefore there is a need to carry out similar user studies also in this environment.

11.4.2 Identify Suitable Functions

Even though speech interactions can provide head-up and hands-off operation of in-vehicle systems, it does not mean that speech is always the optimal modality to use. Results from studies where speech was compared with manual task performance indicate that speech did not always give the best results [46]. Greenberg et al. [47] compared hands-free and hand-held voice mail retrieval during a driving task where the driver had to detect and avoid swerving vehicles. The hands-free voice mail retrieval condition showed that drivers significantly reduced driving speed during mail retrieval, but otherwise showed the same driving behaviors as the baseline condition (no mail retrieval). Whereas drivers in the hand-held voice mail retrieval condition showed significant reductions in detection of wavering vehicles. This result indicates that although a hands-free solution lets the drivers keep their eyes on the road, and might reduce the cognitive workload, it is not a 100% solution as the drivers were clearly compromised reducing driving speed while retrieving emails.

It is clear that the evaluation of speech systems and speech interactions in vehicles is still a field under development. Some important properties of driving tasks and speech systems that should be considered when conducting research in the field include the characteristics of

- the driving task
- the driving environment
- the selected group of drivers
- the ADAS integrated in the vehicle
- the speech-based driver–vehicle information tasks
- the performance parameters of the speech recognition system
- the driver–vehicle information system interface [48]

11.5 Dialogue System Design

The automotive industry is in the process of taking interactive voice technology to the next level. Speech technology will be embedded in vehicles and will be expected to integrate and work with speech-enabled telematics services located outside the vehicle. Limits will be pushed and dialogue flexibility improved, allowing drivers to speak more naturally. It is therefore of utmost importance that the feedback from the system evolves accordingly.

More importantly, what will studies of new dialogue-based systems show? Today, feedback is kept short and concise to prompt drivers to only speak what is necessary. This is mostly due to difficulties with speech technology, and in particular speech recognition; it is easier for the ASR system to discriminate between the words if the vocabulary is small. This has traditionally made vehicle dialogues very directed. Complex dialogue designs have waited for the technology to mature. Using

complex, conversational dialogue systems is a potential for in-vehicle systems that are able to handle first time users who have not taken the time to read the manual.

11.6 In-Vehicle Research Projects

When discussing dialogue system design for in-vehicle applications, we need to take a look at the European project VICO (Virtual Intelligent CO-driver) (<http://www.vico-project.org/>). The goal of the project is the development of an advanced in-car dialogue system. VICO supports natural language speech interaction with an agent that provides services such as navigation, route planning, hotel and restaurants reservation, tourist information, and car manual consultation. VICO includes a robust speech recognizer, connected with a natural language understanding module and a dialogue manager that can adapt itself to a wide range of dialogues, allowing the driver to address any task and sub-task in any order using any appropriate linguistic form of expression [49]. Two prototypes were built in the project: the first with text-only output and the second displaying road maps and/or route icons. Both prototypes can handle spoken street names and numbers, cities, parts of cities, parts of countries, gas stations, and hotels. They can also answer questions about the VICO system itself. The second prototype can also handle hotel reservations over the web, plan routes that might be of interests for tourists. The driver can get car manual information and control devices in the car by voice. This prototype also has information awareness, meaning that in stressed traffic situations, e.g., when the driver brakes, the system will stop the dialogue so that the driver can concentrate on the traffic. A Wizard-of-Oz study has been carried out to evaluate user expectations of natural language input [50]. The subjects' reactions turned out to be very positive. Speech, when allowed to use natural language, was considered easy and comfortably to use.

The EU-funded TALK project focused on the development of new technologies for adaptive multimodal and multilingual human-computer dialogue systems. The long-term vision is one of users interacting naturally with devices and services using speech, graphics, or a combination of the two. There are two task domains: the in-home and the in-car domain. The main objectives of the TALK project are to (1) develop dialogue systems using strategies learned from human multimodal interactions and modeled within the ISU (Information State Update) approach, thus making dialogue systems a more efficient tool for communicating with information technology devices and services; (2) separate application-specific information from dialogue systems development, resulting in reconfigurable systems and re-usable developer tools, and (3) separate modality- and language-specific information from the dialogue system core, allowing cost-effective and easy portability of a single system to multiple languages and modalities. The in-car system SAMMIE controls an MP3 player and supports natural, mixed-initiative interaction, with particular emphasis on multimodal turn-planning and natural language

generation. Input can be given through any of the modalities speech or graphics and is not restricted to answers to system queries. The user can provide new tasks as well as any information relevant to the current task at any time.

DICO [51] is a project, financed by Swedish VINNOVA foundation, focusing on in-vehicle dialogue systems. The dialogue system used in the project is GoDiS (Gothenburg Dialogue System) [52]. The philosophy of the GoDiS system is to provide general solutions to general dialogue management problems. Problems that arise in every dialogue, such as the need for clarification questions and the possibility to switch to another task in the middle of a conversation, are taken care of generally. This means that irrespective of what application within the dialogue system the user wants to use – e.g., the mobile phone, the navigation system, or any telematics system – the dialogue system behavior is the same, giving the user a feeling of habitability and decreasing the risk of cognitive overload. The focus of the DICO project is to develop dialogue management techniques to handle user distraction, integrated multimodality (see Section 11.7), and noisy speech signals.

CU-move is a DARPA (Defense Advanced Research Projects Agency) project with the goal of developing algorithms and technology for robust access to information via spoken dialog systems in mobile, hands-free environments. The natural conversational system includes intelligent microphone arrays, auditory and speech enhancement methods, environmental noise characterization, and speech recognizer model adaptation methods for changing acoustic conditions in the car. The initial prototype system allows users to get driving directions for the Boulder area via a hands-free cell phone, while driving a car.

The trend is pointing in the direction of developing conversational dialogue systems, i.e., dialogue systems that allow the user to speak in a more natural way (compared to commands), and that is able to negotiate with the user to find out what action to perform. The main difference between dialogue system designs for in-vehicle use, compared to other environments, is the fact that the user has to pay full attention to something else but the dialogue system, namely, the traffic. The fact that the user does not pay full attention to the dialogue system puts heavy demands on the dialogue manager, allowing the system to handle a flexible interaction. The user might, at any time in the dialogue, need to pause to be able to concentrate on the traffic. When the traffic situation allows it, the user should be able to resume the dialogue in a convenient way. This means that the dialogue system must be able to decide how to continue, e.g., move on with the dialogue as if nothing has happened, repeat the last utterances or perhaps restart the whole dialogue. The challenge is to develop techniques to detect the level of user distraction and stress. Most premium brands (i.e., high feature cars) have systems to detect user distraction, e.g., by measuring speed maintenance and lane keeping, or detecting when the driver, for example, uses the blinker. When these systems indicate that the driver is distracted, information that is not of great importance (such as, e.g., warnings of lack of screen washer fluid) is not shown to the driver until the distraction is considered to be over. However, what is distracting for one driver might not at all be distracting for another, so techniques need to be developed to measure each driver's distraction level so that the dialogue system can adapt to each user.

To decrease the risk of cognitive overload, more work has to be done to allow each user to express herself in a way that feels natural, so that the user does not have to remember exactly the right command or words to perform a certain task, and also not having to remember the menu structure. This is a question for dialogue management as well as speech recognition.

In-vehicle dialogue system development needs to find techniques to:

- handle a flexible dialogue where the user can speak in a natural way and not use special command words
- adapt to users with different experience; both the novice user and the expert user must be able to use the dialogue system in a convenient way
- negotiate with the user if there is ambiguity in the user utterance, or if the user asks for a service that the dialogue system is not able to help with
- detect user distraction and adapt to each user's level of distraction and cognitive load
- resume a paused dialogue in a natural way

11.7 Commercial In-Vehicle Dialogue Systems

Blue&Me[®] is a cooperation between Fiat Auto and Microsoft. It is a speaker independent in-car infotainment system based on Windows Mobile for Automotive, a gateway based on Windows CE 5.0 platform. There are two versions; the basic version for connecting mobile phones via Bluetooth and MP3 players via a USB port, and the standard version that also includes GPS (Global Positioning System) connectivity and a build-in GSM phone for delivering off-board services such as real-time traffic information, navigation, vehicle safety and security features, and remote diagnostics. Other services are under development, such as an interactive navigation system, possibilities to check addresses, weather and traffic forecasts, and satellite localization in case of theft and SOS. To start a dialogue the user presses a push-to-talk button placed on the steering wheel. The input is unimodal; the user gives voice commands to, e.g., make a phone call or listen to a song. The output is multimodal; the system gives visual feedback on a dashboard display and auditory feedback via the car speakers. It is used to integrate in-car communication, entertainment, and information.

BMW Voice Recognition enables the driver to control the radio, the phone, and the navigation system by voice. It is a speaker-independent system with a vocabulary of 140 words. It is possible to store phone numbers and names and to recall a number. The driver can also take notes through a memo function. In some models it is also possible to control the so-called iDrive functions such as climate control and trip computer, and the system can use text-to-speech to read SMS and emails aloud. The user gives commands from a fixed vocabulary and the dialogue follows a predefined structure. The system gives feedback on everything that is recognized. The user presses a button to start the dialogue, the dialogue then continues until the

user presses the button again, speaks the word “cancel” or is silent for 10 seconds. An incoming phone call will interrupt the dialogue. There is also a help function that lists the available commands in the requested domain.

Honda uses IBM’s voice recognition technology ViaVoice to control the navigation system Touch by Voice. The driver can ask for directions from the US navigation system and hear responses over the existing car audio system using a “talk” button on the steering wheel. The system can recognize commands in English such as “Find the nearest gas station,” or “Find the nearest Italian restaurant,” the user can also specify an address or location. It is also possible to control the vehicle’s climate control systems, the audio system, and the DVD entertainment system.

There are no public publications regarding usability evaluation of the commercial systems, making it hard to get a view of how efficient the systems are. As can be seen, so far commercial in-vehicle dialogue systems have been mainly command-based. However, since it has become clearer that this approach in many cases increases the risk of cognitive load and user distraction things are starting to change. The trend now is to move toward advanced, conversational dialogue systems.

11.8 Multimodal Interaction

Traditionally, dialogue systems have been either text-based or speech-based. The last few years, however, multimodal systems (i.e., a combination of at least two modalities) have become more common. To enable the driver to keep the hands on the steering wheel and the eyes on the road, speech seems to be the most appropriate modality. Sometimes, however, visual and/or haptic media might be more convenient to use. For example, a single push at the volume switch is perhaps the easiest way to mute the radio. A quick glance at a song list might also be less distracting than having to hear and remember a long list of spoken titles.

There are two main approaches to combining several modalities; the parallel approach and the integrated approach. The parallel approach presupposes that all modalities convey the full information alone, without support from the other [53]. The user can choose either of the possible input modalities to interact with the computer. The major advantage of this approach is that the user does not have to rely on one single modality. If, for example, speech is unsuitable at the moment, the user can choose another modality and still have access to all the functionality. The output modalities, too, might be parallel, but this raises issues regarding dialogue management since it might be inconvenient to let all modalities give all information. Sometimes it may be more informative and less tedious for the user to let the verbal output be brief, and to let the visual output be comprehensive (using maps or tables, for example).

The integrated approach to multimodality means that the modalities in combination convey the full information. In a navigation system the input could, for example, be a combination of speech and mouse clicks or pointings at a touch screen (*I want to go from here [click] to here [click]*). This way of combining modalities is useful

in multimodal systems where one modality is more suited to a specific task than the other. In the above example, speech can be used for the whole task, but sometimes it might be more convenient and easy-to-use mouse clicks to mark the stops instead of trying to remember their names.

The VICO prototypes both have speech as input modality and the interaction is activated by pushing a button. The button may not need a physical push; the dialog system can be activated and deactivated in different ways such as time interval between speech segments or the use of certain commands, and even non-speech sounds. The output modalities are speech and graphics; the screen shows a red or green light to indicate whether the system is listening or not, and also additional information. The first prototype shows text output and the second displays road maps and/or route icons.

The trend is pointing toward multimodal user interfaces, but does this hold for in-vehicle dialogue systems as well? Research needs to be done to find forms for multimodal interaction in vehicles. A study [54] using a map-based multimodal interface shows that people tend to use more than one modality when the cognitive load increases due to task difficulty or communicative complexity. Another study [55] indicates that there is less need for visual displays when driving a car, since the driver is busy keeping her eyes on the road. The challenge is to find out what kind of multimodality is best suited for the in-vehicle environment, and which modalities should be used. Another study [56] points out the need for ecologically valid user tests, e.g., tests where the participants drive a real car in real traffic (as opposed to a driving simulator) and have a well-defined task to perform. People do not always behave in the same way when driving a simulator where a possible accident does not have any consequences, compared to what happens if you drive a real car. Clearly, there is a need to further investigate how multimodality should be designed in vehicles both regarding which modalities to use and how to use them.

11.9 Drivers States and Traits, In-Vehicle Voice and Driving Behavior

11.9.1 Driver States and Traits

People are affected by traits and states and by the environment. Traits and states are concepts that people use to both describe and understand themselves and others. These concepts are prototype-based categories that have fuzzy boundaries. Prototypical traits are stable, long-lasting, and internally caused, such as personality and cognitive abilities. Prototypical states are temporary, brief, and caused by external circumstances, such as emotions and moods. Trait concepts permit people to predict the present from the past; state concepts identify those behaviors that can be controlled by manipulating the present situation. To be able to predict an individual's behavior at any given moment in time requires attention to state as well

as traits. State is defined to be current feelings influenced by the physical situation of the individual, characteristics, and traits are defined to be age, gender, and personality. In all our activities, states and traits will influence everything we do and experience from answering the phone to driving down the highway. Recent studies have shown that driver states and traits interact with characteristics of the voice of in-vehicle information systems and affects both attitude and driving performance [5, 16, 57–60]. The results from these studies indicate that vehicles should know their drivers, and that one system does not fit all. As an example of how states and traits affect drivers and driving performance, we will take an in-depth look at emotions, and then briefly touch upon age and personality.

11.9.2 Emotions and Driving

Emotion is a fundamental component of being human and motivates actions that add meaning to our experiences. Recently, there has been an explosion of research on the psychology of emotion [61] in human–computer interaction. Emotion in human–computer interaction is no longer limited to excitement when a hard task is resolved or frustration when reading an incomprehensible error message. New results show that emotions play a critical role in all goal-directed activities. Off the shelf technologies can be used to assess, detect, and identify emotions or emotional states in real time [62]. Individuals communicate most of their emotions by a combination of words, sounds, facial expressions, and gestures. Anger, for example, causes many people to frown and yell. People also learn ways of showing their emotions from social interactions, though some emotional behavior might be innate. Paralinguistic cues such as tone of voice and prosodic style are among the most powerful of these social signals even though people are usually unaware of them.

Emotions can direct our attention to objects and situations that are important to our needs and goals. This is done through emotion-relevant thoughts that dominate our conscious processing. The focus increases with the importance of the situation [63]. We can also divert attention through emotion regulation [64]. If we, on some cognitive level, determine that a particular emotion is undesirable, we can re-focus our attention to some other object or task. As an example, becoming angry with an in-vehicle information system is not seen as productive by most drivers, so instead they either focus attention on something else in the car or they simply just turn the system off. Emotion regulation can also be used for positive emotions; there are times when we re-focus our attention to prevent, for instance, inappropriate laughter. Emotion regulation works in most cases, there is, however, the danger of failure if a situation or information system is too arousing. High arousal levels makes it harder for us to re-focus attention, and in the case of an in-vehicle information system, the driver will be unable to ignore the system.

Differentiating between emotion and mood is important since they influence interactions differently. The difference between emotion and mood is based on time. Mood is a longer term affective states that bias people's responses, whereas

emotion is a more immediate and short duration affective state [65]. Mood has a different impact on attention and people tend to pay more attention to events and visuals that are congruent with their current mood [66]. It has also been shown that people that are in a good mood often regulate mood by performing tasks that sustain their mood or counteract undesired moods.

Driving, in particular, presents a context in which an individual's emotional state can have enormous impact. Emotions influence attention, performance, and judgment, and their properties are extremely important behind the steering wheel of a car. In complex and dense traffic situations even the smallest disturbance has the potential to be disastrous. Considering the effects of emotion, and, in particular, that positive affect leads to better performance and less risk-taking, it is not surprising that research and experience demonstrate that happy drivers are better drivers [67].

Interactive information systems are moving into the car. Most car manufacturers offer in-vehicle navigation systems, and these systems, regardless of whether they use screen-based interactions, speech-based interactions or a mix thereof, will affect the driver's attitude and driving performance. Screen-based interaction requires the driver's eyes and focus to move from the road to the screen [68, 69], and attention theory suggests that speech-based interactions would work better to keep the driver focused on the road. Results, however, show that even with speech-based interactions, drivers tended to take risks during interactions and often failed to compensate for slower reaction times [70]. Speech-based interactions in the car also share some of the characteristics of mobile phone conversation and may show the same effect on driving performance [71]. Using a mobile phone, part of the driver's attention transfers from the road to the ongoing communication. This, together with the communication partner's lack of knowledge of the driving condition and the driver's current situation, increases the risk of unintentionally creating a hazardous driving situation. It is important to note that for all these systems, and in particular speech-based systems, emotions can be used to focus and direct attention.

There are, however, fundamental differences between conversation with in-vehicle computers, conversations using mobile phones, and conversations with passengers. There are studies that show the difference in impact on driving performance between conversation with a passenger and conversation using a mobile phone [72]. There is, however, very little data published on the impact of a conversational interface on a driver's attitude and driving performance. Even so, car manufacturers are increasingly turning to voice as a promising strategy for safe and engaging interactions with in-vehicle systems. This has made it critical to know how linguistic features of the in-vehicle voice such as emotion interact with a driver's emotion in affecting attention, performance, and judgment.

The characteristics of an in-vehicle voice have the potential to impact the drivers focus, attention, performance, and judgment. Furthermore, it becomes important to learn what happens when the characteristics of the in-vehicle voice and the emotional state of the driver are matched and mismatched. In a driving-simulator-based study we investigated the question to see what happened when an upset driver encountered a happy car voice, and what happened when the upset driver encountered an upset car voice [73]. To assess the effects of the link between driver emotion

and voice emotion on drivers' performance, the number of accidents that drivers were involved in during their time in the simulator was recorded. In addition to this, the driver's attention to the driving task was measured as a reaction time to honk the horn in response to randomly occurring honks. The driver's engagement with the system was measured by the amount of time drivers spent talking back to the voice in the car while driving down the simulated road.

Matching the voice of the car to the drivers' emotions had enormous consequences. Drivers who interacted with voices matched to their own emotional state (energetic voice for happy drivers and subdued voice for upset drivers) had less than half as many accidents on average as drivers who interacted with mismatched voices [5]. This reduction is greater than the effects of happy versus sad on driving performance, providing cost effective means to safer driving. Using the wrong or mismatched in-vehicle voice seemed to distract drivers. They divided their attention and cognitive power between driving and unconsciously trying to figure out why someone interacting with them had such a different emotional state. Possibly, the drivers also pondered why the in-vehicle voice did not follow the social protocol and adapt to their (the driver's) emotional state. The result was that drivers with mismatched in-vehicle voice had more accidents and paid less attention to the driving task.

Drivers that were paired with a similar car voice, on the other hand, communicated much more with the car voice than drivers paired with a mismatched voice. It is interesting to note that the in-vehicle said exactly the same thing in all conditions. This suggests that drivers who interacted with in-vehicle voices similar to themselves or their own emotional state were able to speak more while avoiding accidents.

The effects of matching an in-vehicle voice to the driver emotion were more powerful than driver emotion alone. Although there was a slight tendency for happy drivers to be better drivers, even this effect was unimportant compared to the effects of matching. In other words, finding the appropriate in-vehicle voice for the driver's emotion stood out as the most critical factor in enabling a safe and engaging driving experience.

This suggests that detecting and responding to a driver's emotional state can have a dramatic influence on driver safety. Simply changing the para-linguistic characteristics of a voice was sufficient to impact on driving performance. Using the same words, spoken at the same times, by the same voice, under the same road conditions, driver performance can be altered by changing the emotional coloring of the in-vehicle voice.

11.9.3 Age of Voice, Personality, and Driving

Further studies on how voice characteristics affect and interact with driver state and driver characteristics and traits show similar results, one voice does not work for all.

The driving task places significant perceptual and cognitive demands on the driver, and the normal aging process negatively affects many of the perceptual,

cognitive, and motor skills necessary for safe driving [74, 75]. Voice prompts and speech messages provide reminders concerning previous interaction for those with poor memories in the domain of computing [76]. Voice messages can play an important part in helping people to execute everyday tasks or to help them execute them more efficiently. The messages can provide useful information about the environment and current events, which especially older people may not readily absorb for themselves. Speech support is also used to help with strategizing, making contextually relevant suggestions, i.e., what to do next in computer interactions, and to provide warnings in safety critical situations. This type of speech support could easily be visualized as part of an in-vehicle information system to help older adult with the driving task.

It is important to consider both the selection of voice and the format of the informational content when designing a speech-based in-vehicle system for older adults. Most older adults are found to be less able to absorb long instructions than younger people [77]. This supports the usefulness of an in-vehicle system with timely speech-based traffic-related information, since the information then appears at the point it is required and useful, instead of at the beginning of the driving task. Varying the linguistic and para-linguistic properties of the voices used by in-vehicle system can also influence driving performance, as characteristics of voices can influence people's attention and affect performance, judgment, and risk-taking [6, 57]. Introducing a speech-based in-vehicle system for older adults could potentially prove beneficial and improve driving performance and driver satisfaction, or it could prove disastrous and distract their attention from the driving task.

Characteristics of the voice affect listeners perception of liking and credibility of what is said either by another human or by a speech-based system [6]. The psychological literature suggests that consistency is important. People expect and prefer consistency. When inconsistency is encountered, people enter a state where they are motivated to adapt their perceptions in order to resolve inconsistency [78], and this creates a mental workload and hence has the potential to affect driving performance. It has also been shown [79] that most successful human communication will occur between a source and a receiver who are alike, i.e., homophilous, and have a common frame of reference. Communication is more effective [80] when source and receiver share common meanings, belief, and mutual understanding. Individuals enjoy the comfort of interacting with others who are similar. Talking with those who are markedly different from a person requires more effort to make communication effective, once again with a potential negative impact on driving performance.

An in-vehicle hazard and warning system was designed to address the question of how voice selection and information content of an in-vehicle information system affects driving performance and attitude for drivers 55 years of age and older. The age group was selected based on a publication by the American Automobile Associations Foundation for Traffic Safety [58]. To minimize the cognitive load induced by the system, the hazard and warning system was designed as a pure information system. The system provided information on upcoming road hazards and traffic events, without engaging the driver in a dialogue. Information was designed to focus attention to traffic situations and by providing older adults with relevant information, extra time, and distance to evaluate the situation, the hope was to

improve their ability to react confidently. The in-vehicle hazard and warning system was realized as two versions with the same information content. One version had all warning and hazard messages read by a young adult (20 years of age) and the other version had all the messages read by an older adult (76 years of age).

Assessment of the older adult voice and the young voice in a lab setting shows that the older adult voice is more credible, more trustworthy, and more intelligible than the young voice. This assessment of trust and credibility changes dramatically when the older adult voice is used in an in-vehicle hazard and warning system. The hypothesis based on similarity attraction and homophiles that the older adult voice would be better liked and trusted by older adults driving with the in-vehicle hazard and warning system was contradicted by the results from the driving simulator experiment. The data from the driving simulator experiment instead show that the driving performance of older adult drivers was significantly better when driving with the young voice than the older adult voice. These findings highlight the importance of context and emphasize that the car is different from a lab setting that is not associated with a driving task. Hence, in-vehicle systems or products targeted for the car should, for reliability of results, be tested in a driving simulator and ultimately in a real car [58]. This experiment demonstrates that there is significant potential for increasing the safety older adult drivers (over 55 years of age) and also for young drivers [57] by providing timely information concerning road hazard. Results also show that the information is well-received by the drivers if the voice of the system is carefully selected. It is clear that the choice of voice for an in-vehicle information system is very important, and especially so for older adult drivers. The judgments of older participants in this study were far more affected by a change in context than those of younger people. Dulude [81] found more flexibility in younger people when investigating performance with interactive voice response systems.

The results indicate that the car represents a truly different setting for information systems. Voices must be tested and selected in the relevant context: properties in a voice-only setting might be perceived differently in a driving simulator or a car than in other settings, resulting in unexpected influences on driving performance. It would therefore be advisable to test the voice intended for use in the car in a driving simulator or car, before proceeding to introduce to productize the voice for an in-vehicle information system.

Another study was setup to investigate the effect of matching the personality of the voice and language used by an in-vehicle navigation system with the personality of the driver. The dimensions of dominant/extrovert and submissive/introvert were selected for the study based on Wiggins 8 dimensional Interpersonal Adjective Scale. Two versions of a navigation system with the same information content were realized. One version used recorded navigation information and suggestions read by a dominant voice using matching words and sentence structures, "It is faster to turn right." The other version was made up of the same navigation information recorded by the same voice, however read in a submissive voice with matching words and sentence structures, "It might be faster to turn right." The dominant/voice was characterized by a faster speech rate, higher amplitude, higher pitch, and also more pitch variation than the submissive voice.

In the study, 40 drivers, 20 dominant/extrovert and 20 submissive/introvert were selected based on their personality assessment. The personality of the 40 drivers were then either matched or mismatched with the personality of the in-vehicle voice during a 30-min navigation task in a driving simulator. The results show that matching the personality of the voice of the in-vehicle navigation system to the personality of the driver had an enormous effect. Matched cases show better driving performance, drivers have a higher opinion of the quality of the system and the car, and most fundamentally, they also follow suggestions more often than in non-matched conditions.

The key finding here is that there is not one effective voice for all drivers. For both attitude and driving performance, drivers clearly benefit from a car, or an in-vehicle information system that knows its driver. This suggests that voices used in in-vehicle systems must adapt to their drivers. This presents two important questions: (1) How can an interface detect driver state and traits (emotion, age, gender, personality)? (2) How can that information be used most effectively to ensure safe driving and a pleasurable driving experience?

11.10 Usability and Acceptance

One of the most critical issues evaluating the user interfaces is the demand for users' limited attention as they navigate the external world. This issue becomes crucial in the domain of driving. The drivers' primary task is clearly safe driving, and to safely navigate the vehicle in traffic using the steering wheel and controlling speed. It is, however, equally clear that drivers engage in a wide variety of secondary tasks through manual or speech interactions.

When giving drivers access to in-vehicle information system via speech technology, driver acceptance and usability are very important, in addition to the consideration of safety. The in-vehicle interface will strongly influence how a user views and understands the functionality of the system. Driver acceptance of both interfaces and functionality will determine if and how the in-vehicle systems are used, and will hence play a critical role in what intelligent vehicles look like and how they perform.

Acceptance of a system is not just a measure of how the functionality of the system, it is also a measure of how the interactions with the system are designed. How do drivers input information to the system and how the results are presented?

As described above in how driver states and traits interact with how and what information is presented, the audio output from the system matters. Male and female voices are both available as recorded speech, text-to-speech, or both. Some in-vehicle systems allow the driver to select the voice for the system, offering a range of unknown or known female and male voices. Integrating different in-vehicle systems in a car, there is also the possibility that these systems use different voices. There might be one voice for the voice dialing application, another for the controlling the radio, and a third for the navigation system. Since the car presents a different

context, and most importantly one where performance is crucial, this is an area that needs to be studied.

There is currently no standard in how to evaluate acceptance of new technology and new in-vehicle systems. Van Der Laan et al. [82] proposed a tool for how to study the acceptance of new technology in vehicle. The authors describe the tool and how it should be used in a report where they compared data from six studies of in-vehicle information systems. Driver experience is basically measured by using questionnaires with nine items: useful/useless; pleasant/unpleasant; bad/good; nice/annoying; effective/superfluous; irritation/likeable; assisting/worthless; undesirable/desirable; and raising alertness/sleep-inducing. There is, however, no correlation between specific functions or interaction of the systems and the questionnaire. The tool can thus be used to rate the overall acceptance of a system, but there is no support to use the tool to diagnose and describe specific design problems. A thorough search of the literature shows no method for how to evaluate the acceptance of speech system in vehicle. This is an area that needs to be explored.

There are, however, different methods for evaluating the usability of an interactive speech system [83–86]. But usability evaluation for interactive speech-based system in-vehicle application is a new topic. As stressed previously, functionality and usability of interactive speech systems should be done simultaneously [87]. Fulfilling the requirements of the functionality for an application does not guarantee good usability and user acceptance. As history has shown us, the best technology is not always the winner, an inferior technology with a superior usability and hence acceptance rating can win the race [88]. A review of the literature did not reveal any criteria to be used when evaluating the usability of an interactive in-vehicle speech system. Besides, traditional requirements for usability have proved not to be enough to rate driver acceptance and driver experience. Enjoyment, emotions, and mood have been shown to play an important role for acceptance [89]. This is an area that is crucial but not fully explored, what is the use of the best and safest speech-based in-vehicle system, if the driver does not like it and turns it off?

References

1. Cameron, H. (2000). Speech at the interface. In: Workshop on “Voice Operated Telecom Services”. Ghent, Belgium, COST 249.
2. Heisterkamp, P. (2001). Linguatronic – Product-level speech system for Mercedes-Benz Cars. In: Proc. HLT, San Diego, CA, USA.
3. Hamerich, S. W. (2007). Towards advanced speech driven navigation systems for cars. In: 3rd IET Int. Conf. on Intelligent Environments, IE07, Sept. 24–25, Ulm, Germany.
4. Goose, S., Djennane, S. (2002). WIRE3: Driving around the information super-highway. Pers. Ubiquitous Comput., 6, 164–175.
5. Nass, C., Jonsson, I.-M., Harris, H., Reaves, B., Endo, J., Brave, S., Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. In: CHI '05 Extended Abstracts on Human factors in Computing Systems. ACM Press, New York, NY.
6. Nass, C., Brave, S. B. (2005). Wired for Speech: How Voice Activates and Enhance the Human Computer Relationship. MIT Press, Cambridge, MA.
7. Bishop, R. (2005). Intelligent Vehicle Technology and Trends. Artech House, Boston.

8. van de Weijer, C. (2008). Keynote 1: Dutch connected traffic in practice and in the future. In: IEEE Intelligent Vehicles Sympos. Eindhoven, The Netherlands, June 4–6.
9. Gardner, M. (2008). Nomadic device integration in Aide. In: Proc. AIDE Final Workshop and Exhibition. April 15–16, Göteborg, Sweden.
10. Johansson, E., Engström, J., Cherri, C., Nodari, E., Toffetti, A., Schindhelm, R., Gelau, C. (2004). Review of existing techniques and metrics for IVIS and ADAS assessment. EU Information Society Technology (IST) program IST-1-507674-IP: Adaptive Integrated Driver-Vehicle Interface (AIDE).
11. Lee, J. D., Caven, B., Haake, S., Brown, T. L. (2001). Speech-based interaction with in-vehicle computer: The effect of speech-based e-mail on driver's attention to the roadway. *Hum. Factors*, 43, 631–640.
12. Barón, A., Green, P. (2006). Safety and Usability of Speech Interfaces for In-Vehicle Tasks while Driving: A Brief Literature Review. Transportation Research Institute (UMTRI), The University of Michigan.
13. Saad, F., Hjälm Dahl, M., Cañas, J., Alonso, M., Garayo, P., Macchi, L., Nathan, F., Ojeda, L., Papakostopoulos, V., Panou, M., Bekiaris, E. (2004). Literature review of behavioural effects. EU Information Society Technology (IST) program: IST-1-507674-IP, Adaptive Integrated Driver-Vehicle Interface (AIDE).
14. Treffner, P. J., Barrett, R. (2004). Hands-free mobile phone speech while driving degrades coordination and control. *Transport. Res. F*, 7, 229–246.
15. Esbjörnsson, M., Juhlin, O., Weilenmann, A. (2007). Drivers using mobile phones in traffic: An ethnographic study of interactional adaptation. *Int. J. Hum. Comput. Inter., Special Issue on: In-Use, In-Situ: Extending Field Research Methods*, 22 (1), 39–60.
16. Jonsson, I.-M., Chen, F. (2006). How big is the step for driving simulators to driving a real car? In: IEA 2006 Congress, Maastricht, The Netherlands, July 10–14.
17. Chen, F., Jordan, P. (2008). Zonal adaptive workload management system: Limiting secondary task while driving. In: IEEE Intelligent Transportation System, IVs' 08, Eindhoven, The Netherlands, June 2–6.
18. Esbjörnsson, M., Brown, B., Juhlin, O., Normark, D., Östergren, M., Laurier, E. (2006). Watching the cars go round and round: designing for active spectating. In: Proc. SIGCHI Conf. on Human Factors in computing systems, Montréal, Québec, Canada, 2006.
19. Recarte, M. A., Nunes, L. M. (2003). Mental workload while driving: Effects on visual search, discrimination, and decision making. *J. Exp. Psychol.: Appl.*, 9 (2), 119–137.
20. Victor, T. W., Harbluk, J. L., Engström, J. A. (2005). Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transport. Res. Part F*, 8 (2), 167–190.
21. Hart, S. G., Staveland, L. E. (1988). Development of NASA-TLX (task Load Index): Results of empirical and theoretical research. In: Meshkati (ed) *Human Mental Workload*, P. A. H. a. N. Elsevier Science Publishers B.V., North-Holland, 139–183.
22. Pausie, A., Sparpedon, A., Saulnier, G. (2007). Ergonomic evaluation of a prototype guidance system in an urban area. Discussion about methodologies and data collection tools, in Vehicle Navigation and Information Systems Conference. In: Proc. in conjunction with the Pacific Rim TransTech Conf. 6th Int. VNIS. "A Ride into the Future", Seattle, WA, USA.
23. Wang, E., Chen, F. (2008). A new measurement for simulator driving performance in situation without interfere from other vehicles, *International Journal of Transportation Systems F. AEI* 2008. In: Applied Human Factors and Ergonomics 2008, 2nd Int. Conf., Las Vegas, USA, July 14–17.
24. Wilson, G. F., Lambert, J. D., Russell, C. A. (2002). Performance enhancement with real-time physiologically controlled adaptive aiding. In: HFA Workshop: Psychophysiological Application to Human Factors, March 11–12, 2002. Swedish Center for Human Factors in Aviation.
25. Wilson, G. F. (2002). Psychophysiological test methods and procedures. In: HFA Workshop: Psychophysiological Application to Human Factors, March 11–12, 2002. Swedish Center for Human Factors in Aviation.

26. Lai, J., Cheng, K., Green, P., Tsimhoni, O. (2001). On the road and on the web? Comprehension of synthetic and human speech while driving. In: Conf. on Human Factors and Computing Systems, CHI 2001, 31 March–5 April 2001. Seattle, Washington, USA.
27. Hermansky, H., Morgan, N. (1994). RASTA processing of speech. *IEEE Trans. Speech Audio Process.*, 2 (4), 578–589.
28. Kermorvant, C. (1999). A comparison of noise reduction techniques for robust speech recognition. IDIAP research report, IDIAP-RR-99-10, Dalle Molle Institute for perceptual Artificial Intelligence, Valais, Switzerland.
29. Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoustics, Speech Signal Process.*, 34 (1), 52–59.
30. Mansour, D., Juang, B.-H. (1989). The short-time modified coherence representation and noisy speech recognition. *IEEE Trans. Acoustics Speech Signal Process.*, 37 (6), 795–804.
31. Hernando, J., Nadeu, C. (1997). Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition. *IEEE Trans. Speech Audio Process.*, 5 (1), 80–84.
32. Chen, J., Paliwal, K. K., Nakamura, S. (2003). Cepstrum derived from differentiated power spectrum for robust speech recognition. *Speech Commun.*, 41 (2–3), 469–484.
33. Yuo, K.-H., Wang, H.-C. (1998). Robust features derived from temporal trajectory filtering for speech recognition under the corruption of additive and convolutional noises. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, April 21–24, 1997, Munich, Bavaria, Germany.
34. Yuo, K.-H., Wang, H.-C. (1999). Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences. *Speech Commun.*, 28, 13–24.
35. Lebart, K., Boucher, J. M. (2001). A new method based on spectral subtraction for speech dereverberation. *Acta Acoustic ACUSTICA*, 87, 359–366.
36. Lee, C.-H., Soong, F. K., Paliwal, K. K. (1996). *Automatic Speech and Speaker Recognition*. Kluwer, Norwell.
37. Gales, M. J. F., Young, S. J. (1995). Robust speech recognition in additive and convolutional noise using parallel model combination. *Comput. Speech Lang.*, 9, 289–307.
38. Gales, M. J. F., Young, S. J. (1996). Robust continuous speech recognition using parallel model combination. *IEEE Trans. Speech Audio Process.*, 4 (5), 352–359.
39. Acero, A., Deng, L., Kristjansson, T., Zhang, J. (2000). HMM adaptation using vector Taylor series for noisy speech recognition. In: Proc. ICASSP, June 05-09, 2000, Istanbul, Turkey.
40. Kim, D. Y., Un, C. K., Kim, N. S. (1998). Speech recognition in noisy environments using first-order vector Taylor series. *Speech Commun.*, 24 (1), 39–49.
41. Visser, E., Otsuka, M., Lee, T.-W. (2003). A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments. *Speech Commun.*, 41, 393–407.
42. Farahani, G., Ahadi, S. M., Homayounpour, M. M. (2007). Features based on filtering and spectral peaks in autocorrelation domain for robust speech recognition. *Comput. Speech Lang.*, 21, 187–205.
43. Choi, E. H. C. (2004). Noise robust front-end for ASR using spectral subtraction, spectral flooring and cumulative distribution mapping. In: Proc. 10th Australian Int. Conf. on Speech Science & Technology. Macquarie University, Sydney, December 8–10.
44. Fernandez, R., Corradini, A., Schlangen, D., Stede, M. (2007). Towards reducing and managing uncertainty in spoken dialogue systems. In: The Seventh International Workshop on Computational Semantics (IWCS-7). Tilburg, The Netherlands, Jan 10–12.
45. Skantze, G. (2005). Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Commun.*, 45 (3), 325–341.
46. Gellatly, A. W. a. D., T. A. (1998). Speech recognition and automotive applications: using speech to perform in-vehicle tasks. In: Proc. Human Factors and Ergonomics Society 42nd Annual Meeting, October 5-9, 1998, Hyatt Regency Chicago, Chicago, Illinois.
47. Greenberg, J., Tijenna, L. Curn, R., Artz, B., Cathey, L., Grant P, Kochhar, D., Koxak, K., Blommer, M. (2003). Evaluation of driver distraction using an event detection

- paradigm. In: Proc. Transportation Research Board Annual Meetings, January 12-16, 2003, Washington, DC.
48. McCallum, M. C., Campbell, J. L., Richman, J. B., Brown, J. (2004). Speech recognition and in-vehicle telematics devices; Potential reductions in driver distraction. *Int. J. Speech Technol.*, 7, 25–33.
 49. Bernsen, N. O., Dybkjaer, L. (2002). A multimodal virtual co-driver's problems with the driver. In: ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments Proceedings. Kloster Irsee, Germany, June 17–19.
 50. Geutner, P., Steffens, F. Manstetten, D. (2002). Design of the VICO Spoken Dialogue System: Evaluation of User Expectations by Wizard-of-Oz Experiments. In: Proc. 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002). Las Palmas, Spain, May.
 51. Villing, J.a.L., S. (2006). Dico: A multimodal menu-based in-vehicle dialogue system. In: The 10th Workshop on the Semantics and Pragmatics of Dialogue, brandial'06 (Sem-Dial 10). Potsdam, Germany, Sept 11–13.
 52. Larsson, S. (2002). Issue-based dialogue management. PhD Thesis, Goteborg University.
 53. Bringert, B., Ljunglöf, P., Raanta, A. and Cooper, R. (2005). Multimodal dialogue systems grammars. In: The DIALOR'05, 9th Workshop on the Semantics and Pragmatics of Dialogue. Nancy (France), June 9–11, 2005.
 54. Oviatt, S. (2004). When do we interact multimodally? Cognitive load and multimodal communication patterns. In: Proc. 6th Int. Conf. on Multimodal Interfaces. Pennsylvania, Oct 14–15.
 55. Bernsen, O., Dybkjaer, L. (2001). Exploring natural interaction in the car. In: Proc. CLASS Workshop on Natural Interactivity and Intelligent Interactive Information Representation, Verona, Italy, Dec 2001.
 56. Esbjörnsson, M., Juhlin, O., Weilenmann, A. (2007). Drivers using mobile phones in traffic: An ethnographic study of interactional adaption. *Int. J. Hum Comput Interact.*, Special Issue on In-Use, In-Situ: Extending Field Research Meth., 22 (1), 39–60.
 57. Jonsson, I.-M., Nass, C., Endo, J., Reaves, B., Harris, H., Ta, J. L., Chan, N., Knapp, S. (2004). Don't blame me I am only the driver: Impact of blame attribution on attitudes and attention to driving task. In: CHI '04 extended Abstracts on Human Factors in Computing Systems, Vienna, Austria.
 58. Jonsson, I.-M., Zajicek, M. (2005). Selecting the voice for an in-car information system for older adults. In: Human Computer Interaction Int. Las Vegas, Nevada, USA.
 59. Jonsson, I.-M., Zajicek, M., Harris, H., Nass, C. I. (2005). Thank you I did not see that: In-car speech-based information systems for older adults. In: Conf. on Human Factors in Computing Systems. ACM Press, Portland, OR.
 60. Jonsson, I. M., Nass, C. I., Harris, H., Takayama, L. (2005). Got Info? Examining the consequences of inaccurate information systems. In: Int. Driving Symp. on Human Factors in Driver Assessment, Training, and Vehicle Design. Rockport, Maine.
 61. Gross, J. J. (1999). Emotion and emotion regulation. In: John, L. A. P. O. P. (ed) *Handbook of Personality: Theory and Research*. New York: Guilford, 525–552.
 62. Picard, R. W. (1997). *Affective Computing*. MIT Press, Cambridge, MA.
 63. Clore, G. C., Gasper, K. (2000). Feeling is believing: Some affective influences on belief. In: Frijda, A. S. R. M. N. H., Bem, S. (eds) *Emotions and Beliefs: How Feelings Influence Thoughts*, Editions de la Maison des Sciences de l'Homme and Cambridge University Press (jointly published), Paris/Cambridge, 10–44.
 64. Gross, J. J. (1998). Antecedent- and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology. *J. Personality Social Psychol.*, 74, 224–237.
 65. Davidson, R. J. (1994). On emotion, mood, and related affective constructs. In: Davidson, P. E. R. J. (ed) *The Nature of Emotion*, Oxford University Press, New York, 51–55.
 66. Bower, G. H., Forgas, J. P. (2000). Affect, memory, and social cognition. In: Eich, J. F. K. E., Bower, G. H., Forgas, J. P., Niedenthal, P. M. (eds) *Cognition and Emotion*. Oxford University Press, Oxford, 87–168.
 67. Groeger, J. A. (2000). *Understanding Driving: Applying Cognitive Psychology to a Complex Everyday Task*. Psychology Press, Philadelphia, PA.

68. Lunenfeld, H. (1989). Human factor considerations of motorist navigation and information systems. In: Proc. Vehicle Navigation and Information Systems, September 11-13, Toronto, Canada.
69. Srinivasan, R., Jovanis, P. (1997). Effect of in-vehicle route guidance systems on driver workload and choice of vehicle speed: Findings from a driving simulator experiment. In: Ian Noy, Y. (ed) Ergonomics and Safety of Intelligent Driver Interfaces, Lawrence Erlbaum Associates Inc., Publishers, Mahwah, New Jersey, 97–114.
70. Horswill, M., McKenna, F. (1999). The effect of interference on dynamic risk-taking judgments. *Br. J. Psychol.*, 90, 189–199.
71. Strayer, D., Drews, F., Johnston, W. (2003). Cell phone induced failures of visual attention during simulated driving. *J. Exp. Psychol.: Appl.*, 9 (1), 23–32.
72. Merat, N., Jamson, A. H. (2005). Shut up I'm driving! Is talking to an inconsiderate passenger the same as talking on a mobile telephone. In: 3rd Int. Driving Symp.on Human Factors in Driver Assessment, Training, and Vehicle Design. Rockport, Maine.
73. Nass, C. et al. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. In: CHI '05 Extended Abstracts on Human Factors in Computing Systems. ACM Press, New York, NY.
74. Brouwer, W. H. (1993). Older drivers and attentional demands: consequences for human factors research. In: Proc. Human Factors and Ergonomics Society-Europe, Chapter on Aging and Human Factors. Soesterberg, Netherlands, 93–106.
75. Ponds, R. W., Brouwer, W. H., Wolffelaar, P. C. (1988). Age differences in divided attention in a simulated driving task. *J. Gerontol.*, 43 (6), 151–156.
76. Zajicek, M., Hall, S. (1999). Solutions for elderly visually impaired people using the Internet. In: The 'Technology Push' and The User Tailored Information Environment, 5th Eur. Research Consortium for Informatics and Mathematics – ERCIM. 2000. Dagstuhl, Germany, November 28–December 1.
77. Zajicek, M.a.M., W. (2001). Speech output for older visually impaired adults. In: Blandford, A., Vanderdonckt, J., Gray, P. (eds) People and Computers XV - Interacting without Frontiers, Spring Verlag, 503–513.
78. Fiske, S., Taylor, S. (1991). *Social Cognition*. McGraw-Hill, New York, NY.
79. Lazarsfeld, P., Merton, R. (1948). Mass communication-popular taste and organized social action. In: Bryson, L. (ed) Institute for Religious and Social Studies, Nueva York.
80. Rogers, E., and Bhowmik, D. (1970). Homophily-Heterophily: Relational concepts for communication research. *Public Opinion Q.*, 34, 523.
81. Dulude, L. (2002). Automated telephone answering systems and aging. *Behav. Inform. Technol.*, 21, 171–184.
82. Van Der Laan, J., Heino, A., De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. *Transport Res. C*, 5 (1), 1–10.
83. Dybkjær, L., Bernsen, N. O., Minker, W. (2004). Evaluation and usability of multimodal spoken language dialogue systems. *Speech Commun.*, 43, 33–54.
84. Graham, R., Aldridge, L., Carter, C., Lansdown, T. C. (1999). The design of in-car speech recognition interfaces for usability and user acceptance. In: Harris, D. (ed) *Engineering Psychology and Cognitive Ergonomics*, Ashgate, Aldershot, 313–320.
85. Larsen, L. B. (2003). Assessment of spoken dialogue system usability – what are we really measuring? In: 8th Eur. Conf. on Speech Communication and Technology – Eurospeech 2003. September 1–4, Geneva, Switzerland.
86. Zajicek, M., Jonsson, I. M. (2005). Evaluation and context for in-car speech systems for older adults. In: The 2nd Latin American Conf. on Human–Computer Interaction, CLIHC, Cuernavaca, México, October 23–26, 2005.
87. Chen, F. (2004). Speech interaction system – how to increase its usability. In: The 8th Int. Conf. on Spoken Language Processing, Interspeech. ICSL, Jeju Island, Korea, Oct 4–8, 2004.
88. Norman, D. (2007). *The Design of Future Things*. Basic Books, New York.
89. Jordan, P. W. (2000). *Designing Pleasurable Products*. Taylor & Francis, London and New York.

Chapter 12

Spoken Dialogue Application in Space: The Clarissa Procedure Browser

Manny Rayner, Beth Ann Hockey, Jean-Michel Renders,
Nikos Chatzichrisafis, and Kim Farrell

12.1 Introduction

Anyone who has seen more than three science-fiction films will probably be able to suggest potential uses for spoken dialogue systems in space. Until recently, however, NASA and other space agencies have shown a surprising lack of interest in attempting to make this dream a reality and it is only in the last few years that any serious work has been carried out. The present chapter describes Clarissa, an experimental voice-enabled system developed at NASA Ames Research Center during a 3-year project starting in early 2002, which enables astronauts to navigate complex procedures using only spoken input and output. Clarissa was successfully tested on the International Space Station (ISS) on June 27, 2005, and is, to the best of our knowledge, the first spoken dialogue application in space.

The comparative success of the Clarissa project is perhaps more than anything due to its organisation, which from the beginning has been structured as a close cooperation between the Ames spoken dialogue systems group and the NASA Astronaut Corps. The original conception for the system came from astronauts,¹ in response to a request for suggestions about possible uses for speech technology in manned spaceflight environments. The astronauts explained that a large proportion of their working days on orbit was spent executing complex procedures, typically containing dozens of steps, and covering tasks as diverse as maintaining life support systems, checking out space suits, conducting science experiments and performing medical exams. There are over 12,000 such procedures in the current inventory and a large number of specialists at Mission Control and elsewhere are continually updating, testing and improving them. It is completely impossible for astronauts to memorise more than a small fraction of this huge set.

Today, when carrying out procedures, an astronaut most often reads from a PDF viewer on a laptop computer; when working in an enclosed space, they may have

M. Rayner (✉)
ISSCO/TIM, University of Geneva, CH-1211, Geneva 4, Switzerland
e-mail: emmanuel.rayner@unige.ch

¹In this connection, we would particularly like to mention T.J. Creamer and Mike Fincke.

to print out a paper copy, or in extreme cases ask a colleague to read the procedure aloud for them. Given that procedure tasks are frequently hands-busy and eyes-busy, and that astronaut time is extremely expensive, all these solutions have obvious drawbacks. The astronauts suggested that it should be possible to use speech technology to implement a voice-enabled system, which would read out steps to the user in response to spoken commands. The initial prototype [1] had a vocabulary of less than 50 words and implemented a handful of commands. Successive rounds of development and testing added a large amount of extra functionality. The final version, as used in the 2005 test, had a vocabulary of about 260 words and supported about 75 different commands, including reading steps, scrolling forwards or backwards in the procedure, moving to an arbitrary new step, reviewing non-current steps, adding or removing voice notes, displaying pictures and setting alarms or timers. Section 12.2 gives an overview of the system's functionality and architecture.

To build Clarissa, we have had to address several interesting research problems. We have also had to do a good deal of mainstream software engineering, for example to build the GUI display component and the process on the ISS server machine which holds the database of procedures, and in general to make sure that the system works well enough to pass the stringent NASA flight requirements. In this chapter, we focus on the research aspects. First, consider the relationship between the original PDF version of the procedure and the interactive spoken version presented by Clarissa. These two versions will typically differ in a variety of ways. There are several written constructions, like tables, which cannot be directly transposed into the spoken modality; it is also frequently the case that material which is left implicit in the written version needs to be made explicit in the spoken one. These issues are discussed in Section 12.3.

The next point concerns the user language supported by the system. In general, any type of spoken dialogue system has to steer a course between two basic strategies. At one end, the system can try to adapt to the user. The implementers collect data representing what users would ideally like to be able to say; a common way to do this is to use Wizard of Oz methods. The system is then configured to perform as well as possible on the sample data, usually using statistical techniques. At the other end of the scale, the burden of adaptation is put on the user: the system is designed to offer a predefined range of coverage, which the user is required to learn. In practice, of course, some compromise between these two positions is normal.

For many applications, the robustness inherent in the data-driven approach makes it the preferable choice. Astronauts, however, are very far from being typical users. In most cases, they come from an aeronautics background where use of controlled language is the norm; they have moreover been carefully selected from a large pool of candidates, among other things for their ability and willingness to learn complex new skills in a short time. For this kind of user, a designed coverage approach has clear attractions. Performance for expert users is better, and it is easier to reconfigure the system in response to changes in the specification. Our approach to recognition is discussed in Section 12.4.

Robustness with respect to variation in user language is not particularly important for an application like Clarissa. However, two other kinds of robustness are critical. Since the whole point of the system is to support hands- and eyes-free operation, recognition has to be performed in an “open-mic” mode. This implies that the system needs to be able to reject spoken input (“cross-talk”) not intended for it. Our approach to handling cross-talk is described in Section 12.5. Robustness is also crucial at the level of dialogue management. For the usual reasons, recognition can never be completely accurate; if the system misrecognises, it must be easy for the user to undo or correct the misrecognition. Our approach to dialogue management, described in Section 12.6, supports an elegant and robust handling of correction moves. Finally, Section 12.7 presents the results from the initial on-orbit test and the appendix summarises quantitative performance results referred to in the main body of the chapter.

12.2 System Overview

The Clarissa system runs on a standard A31p IBM laptop under Windows 2000. It consists of a set of software modules, written in several different languages, which communicate with each other through the SRI Open Agent Architecture [2]. Commands can be issued either using voice or through the GUI. This section gives an overview of the system as a whole. We start by listing the different types of supported functionalities and then describe the main modules.

12.2.1 Supported Functionality

The system supports about 75 individual commands, which can be accessed using a vocabulary of about 260 words. Many commands can also be carried out through the GUI. The main system functionality is as follows. In each case, we briefly describe the type of functionality, and give examples of typical commands.

- Navigation: moving to the following step (“next”), the preceding step (“previous”), or a named step (“go to step three”, “go to step ten point two”).
- Visiting non-current steps, either to preview future steps or recall past ones (“read step four”, “read note before step nine”).
- Opening and closing procedures (“open E M U checkout procedure”, “close procedure”).
- Recording, playing and deleting voice notes (“record voice note”, “play voice note on step three point one”, “delete voice note on substep two”).
- Setting and cancelling alarms (“set alarm for five minutes from now”, “cancel alarm at ten twenty one”).
- Showing or hiding pictures (“show figure two”, “hide the picture”).
- Changing volume for TTS and prerecorded audio (“increase volume”, “quieter”).

- Temporarily switching off speech recognition, to put the system in a mode where it will only react to a key-phrase used to restore normal function (“suspend”, “resume”).
- Querying status (“where are we”, “list voice notes”, “list alarms”).
- Commands associated with “challenge verify mode”. This is a special mode suitable for particularly critical parts of a procedure, which aggressively asks for confirmation on each step. The user can directly enter or exit this mode (“enter challenge verify mode”, “exit challenge verify mode”) or else set challenge verify mode on a specified step or range of steps (“set challenge verify mode on steps three through twelve”).
- Responding to system questions. Most of the dialogue is user-initiative, but the system can enter short information-seeking subdialogues in certain situations. The most important types of responses are yes/no words (“yes”, “affirmative”, “no”, “negative”) and numerical values (“zero”, “eleven”, “eight thousand two hundred four”, “sixty one point five”, “no value”).
- Undoing and correcting commands. Any command can be undone (“undo”, “go back”) or corrected (“no I said increase volume”, “I meant exit review mode”). In some cases, the command can be expressed elliptically (“no I said three point one”, “I meant step four”).

12.2.2 Modules

The main software modules of the system are the following:

Speech Processor. The Speech Processor module is built on top of the commercial Nuance Toolkit platform [3] and is implemented in C++ using the RCEngine API. The top-level functionalities it provides are speech recognition and spoken output. Speech recognition can be carried out using either a grammar-based or a statistical language model; speech output can be either through playing recorded audio files, or by using a TTS engine. The Speech Processor’s output from speech recognition always includes a list of words, each tagged with a confidence score. When using a grammar-based language model, it also includes a logical form representation defined by the grammar.²

Semantic Analyser. The Semantic Analyser is implemented in SICStus Prolog. It receives the output of the Speech Processor (a string of words, possibly combined with a logical form), and converts it into a dialogue move. The methods used to do this combine hand-coded patterns and corpus-derived statistical information, and are described in Sections 12.1 and 12.2.

Response Filter. Since recognition is carried out in open-microphone mode, at least some of the speech input needs to be rejected. This function is performed by the Response Filter. The Response Filter receives the surface input from the recogniser (a list of words tagged with confidence scores) and produces a binary judgement,

² Note that the grammar’s “logical forms” and the dialogue manager’s “dialogue moves” are *not* the same.

either to accept or to reject. It is implemented in C using Support Vector Machine techniques described in Section 12.5.

Dialogue Manager. The Dialogue Manager is implemented in SICStus Prolog. It accepts dialogue moves from the Semantic Analyser and the Output Manager, and produces a list of abstract dialogue actions as output. It also maintains a dialogue state object, which encodes both the discourse state and the task state. The Dialogue Manager is entirely side-effect free; its operation is specified by a declarative *update function*. This is described further in Section 12.6.

Output Manager. The Output Manager accepts abstract dialogue actions from the Dialogue Manager, and converts them into lists of procedure calls. Executing these calls results in concrete system responses, which can include production of spoken output, sending display requests to the GUI and sending dialogue moves back to the Dialogue Manager. The Dialogue Manager is implemented in SICStus Prolog.

GUI. The GUI is written in Java Swing, and mediates normal keyboard and screen-based interaction with the user. It accepts input from the Output Manager in the form of display requests. It can also convert keyboard input from the user into dialogue moves, which are sent to the Dialogue Manager.

12.3 Writing Voice-Navigable Documents

The problem of representing written documents for spoken use can be surprisingly complex. In the case of Clarissa, several important and potentially conflicting constraints had to be satisfied by the design of the procedure representation. First, since ISS procedures are critical formal documents that typically reflect hundreds or even thousands of person-hours of effort, including a lengthy approval process, it is not feasible to replace them with a new structure. Second, although the associated visual display is required to faithfully reproduce the official procedures, reading out the procedures verbatim is unworkable. Third, the NASA procedure writing community must accept the spoken version as equivalent in content to the original procedure, or it cannot be formally approved as a valid substitute in the safety-critical environment of the ISS. And finally, all procedure-specific information must be incorporated into the procedure representation, rather than the browser code, to enable use of the same procedure browser for many procedures.

Our approach to satisfying these constraints represents procedures in an XML format that contains all the text and layout information present in the original written procedure, together with additional information which specifies how the text is to be read out in the context of procedure execution. For each procedure, an XML file is compiled into an HTML display document which will exactly mimic the appearance of the original paper document, and also an annotated structure that can be followed by the dialogue manager and which will permit the text to be augmented and paraphrased where appropriate to enable it to be read aloud in a natural manner.

The browser treats this compiled XML procedures as data. This makes it possible to drop in an updated procedure without re-compiling the entire Clarissa system.

Clarissa currently handles five ISS procedures. These procedures are fairly elaborate; they average approximately 53 steps each and require an average of 980 lines of XML to represent them.

Given that the spoken version can, and in a number of cases must, differ from the original written version of the procedure, we are compelled to address the question of how much and in what ways the versions can differ. A basic design principle in Clarissa is that the spoken version models the way a human would read the written version aloud while using it to instruct another person in performing the task. In some parts of the procedures the written and spoken versions are the same and in others they diverge. The divergences, which we describe in more detail in the next few paragraphs, are the major source of complexity in the XML representation. These divergences arise from basic differences between the text and speech modalities and perhaps even more crucially from the fact that the spoken version must be adequate for possible values of the document's dynamic content as well as the document's use.

In some cases, the differences are minor: wording for fluent speech often differs from highly abbreviated and/or acronym-filled text. For example, "H2O vlv↔MSB" would read better as "water valve, disconnect from micro-sample bag". In other cases, visual and spoken structures are so different that, even if one wanted to read that part of the document verbatim, it would not be clear how to do it. Tables are an obvious example. Visually, a table provides information in the formatting. One can scan top and side headers to understand what the values in a table cell mean or what kind of material should be entered in the table cell. Lines typically individuate the cells. What should a spoken version of a table be like? How does one "read" lines separating cells, or "read" the spatial layout of the table? A human reading aloud would probably not give details of the formatting or read all the headers, but would present the table information in a way motivated by how the table needed to be used. Was the table to be read out, or filled in? In what order should cells be read?

Figure 12.1 illustrates the relationship between the written and spoken versions in the case of a table used to elicit values from the user and record them. In this procedure, maintenance is done on between one and three space suits (EMUs). At the beginning of the procedure the spoken system adds queries about which EMUs will be used so that it can restrict its later activity to only the relevant columns of the table. To fill in the table, a human reader would likely elicit the values by incorporating the column and row header information into a query or directive for each cell. The system follows this strategy, associating a query to the user with each cell. Once the user answers, the information is displayed in the relevant cell of the table. Note that language must also be added for reading out symbols and some of the abbreviations and acronyms.

12.3.1 Representing Procedure-Related Discourse Context

As we have already seen, procedures will frequently need to encode requests to acquire information from the user. This information will then become part of the

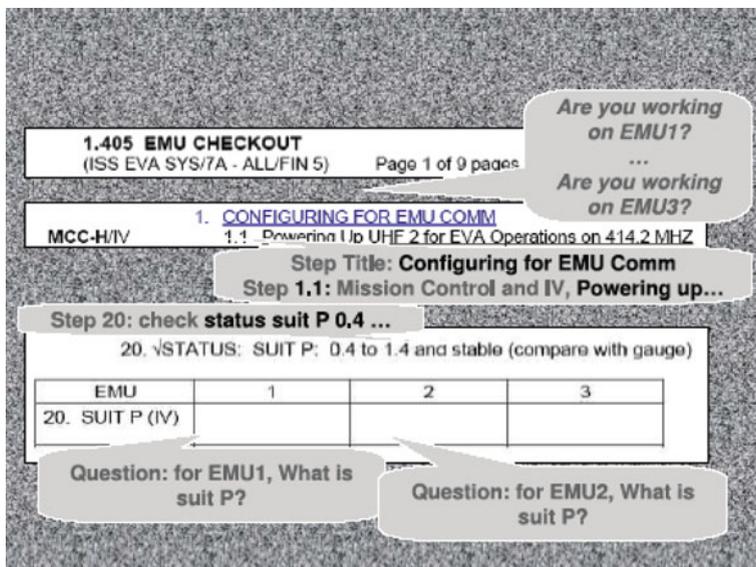


Fig. 12.1 Adding voice annotations to a table

task-related discourse context, and is stored as the bindings of variables defined in the procedure. Later steps may access the variable bindings so as to make decisions about conditional steps, using suitable XML constructs.

The use of context variables creates problems when combined with the requirement that users should be able to move around freely in the procedure. (It is worth pointing out that these problems are very general in nature and would apply to any interactive document which is intended to support both conditional content and free navigation.) Suppose that a value for variable *V* is acquired at step 5, and is then used as part of a condition occurring in step 10. If the user skips straight from step 4 to step 10, *V* will be unbound and the condition will not be evaluable. In situations like these, we would like the dialogue manager to be able to alert the user that they need to execute the earlier step (in this case, step 5) before moving to step 10.

In the general case, this problem appears to be quite intractable; use of a sufficiently expressive variable-setting language can require arbitrarily complex reasoning to determine which steps need to be executed in order to give a binding to a specific variable. A workable solution must involve restricting the way in which variables may be set. The solution we have implemented is fairly drastic, but was sufficient for the purposes of the initial Clarissa project: for each variable *V*, we only permit *V* to be assigned a value in at most one procedure step, so that it is possible to associate *V* with the step in which it can potentially receive a value. If a jump results in attempted evaluation of an unbound variable, the dialogue manager can then tell the user that they first need to execute the relevant earlier step.

12.4 Grammar-Based Recognition

As described in Section 12.2, Clarissa uses a grammar-based recognition architecture. At the start of the project, we had several reasons for choosing this approach over the more popular statistical one. First, we had no available training data. Second, the system was to be designed for experts who would have time to learn its coverage; although there is not a great deal to be found in the literature, an earlier study in which we had been involved [4] suggested that grammar-based systems outperformed statistical ones for this kind of user. A third consideration that affected our choice was the importance of “open-mic” recognition; there was again little published research, but folklore results suggested that grammar-based recognition methods had an edge over statistical ones here too.

Obviously, none of the above arguments are very strong. We thus wanted to implement a framework which would allow us to compare grammar-based methods with statistical ones in a methodologically sound way, and also retain the option of switching from a grammar-based framework to a statistical one if that later appeared justified. The Regulus and Alterf platforms, which we have developed under Clarissa and other earlier projects, are designed to meet these requirements.

The basic idea behind the Regulus platform is to construct grammar-based language models using example-based methods driven by small corpora. Since grammar construction is now a corpus-driven process, the same corpora can also be used to build normal statistical language models, facilitating a direct comparison between the two methodologies. On its own, however, Regulus only permits comparison at the level of recognition strings. Alterf extends the paradigm to the semantic level, by providing a uniform trainable semantic interpretation framework which can work on either surface strings or logical forms.

12.4.1 *Regulus and Alterf*

Regulus [5, 6] is an Open Source toolkit, which can be used to derive a domain-specific Nuance recogniser from a training corpus in a series of steps:

1. The starting point is a general unification grammar [5], Chapter 8 loosely based on the Core Language Engine grammar for English [7]. For a given domain, such as Clarissa, this grammar is extended with a set of domain-specific lexicon entries.
2. The training corpus is converted into a “treebank” of parsed representations. This is done using a left-corner parser representation of the grammar.
3. The treebank is used to produce a specialised grammar in Regulus format, using the Explanation Based Learning (EBL) algorithm [8, 9].
4. The specialised grammar is compiled into a Nuance GSL grammar, using the methods described in [5], Chapter 8.
5. The Nuance GSL grammar is converted into runnable speech recognisers by invoking the Nuance Toolkit compiler utility.

Alterf [10] is another Open Source toolkit, whose purpose is to allow a clean combination of rule-based and data-driven processing in the semantic interpretation phase. Alterf characterises semantic analysis as a task slightly extending the “decision-list” classification algorithm [11, 12]. We start with a set of *semantic atoms*, each representing a primitive domain concept, and define a semantic representation to be a non-empty set of semantic atoms. For example, in Clarissa we represent the utterances.

“please speak up”
 “set alarm for five minutes from now”
 “no i said next”

respectively as

```
{increase_volume}
{set_alarm, 5, minutes}
{correction, next_step}
```

where *increase_volume*, *set_alarm*, *5*, *minutes*, *correction* and *next_step* are semantic atoms. As well as specifying the permitted semantic atoms themselves, we also define a *target model* which for each atom specifies the other atoms with which it may legitimately combine. Thus here, for example, *correction* may legitimately combine with any atom, but *minutes* may only combine with *correction*, *set_alarm* or a number.

Alterf works by training an association between utterances and semantic representations. Training data consists of a set of utterances, in either text or speech form, each tagged with its intended semantic representation. We define a set of *feature extraction rules*, each of which associates an utterance with zero or more features. Feature extraction rules can carry out any type of processing. In particular, they may involve performing speech recognition on speech data, parsing on text data, application of hand-coded rules to the results of parsing or some combination of these. Statistics are then compiled to estimate the probability $p(a|f)$ of each semantic atom a given each separate feature f .

At runtime, these probability estimates are used to associate a set of semantic atoms with an utterance. The decoding algorithm is very simple: we just walk down the list of available semantic atoms, starting with the most probable ones, and add them to the semantic representation we are building up when this does not conflict with the consistency rules in the target model. We stop when the atoms suggested are too improbable, that is, have probabilities below a specified threshold.

12.4.2 Using Regulus and Alterf in Clarissa

The Clarissa Regulus grammar is composed of the general Regulus grammar and the general function-word lexicon, together with a Clarissa-specific domain

Table 12.1 Summary information for Clarissa lexicon

POS	#Entries		Example	Example context
	Lemmas	Words		
Verb	129	645	go	“ go to step three”
Noun	99	127	caution	“read caution before step eleven”
Number	25	25	zero fiver	“set alarm for ten zero fiver ”
Interjection	20	20	copy	“ copy go to step three”
Preposition	15	15	on	“give me help on navigation”
Adjective	15	15	skipped	“list skipped steps”
Adverb	10	10	louder	“speak louder ”
Total	313	857		

lexicon containing 313 lemmas, which realise 857 surface lexical rules. Table 12.1 summarises the data for the domain-specific lexicon.

The training corpus used for grammar specialisation contains 3297 examples; of these, 1349 have been hand-constructed to represent specific words and constructions required for the application, while the remaining 1948 are transcribed examples of utterances recorded during system development. The parameters guiding the grammar specialisation process have been chosen to produce a fairly “flat” grammar, in which many noun-phrases become lexical items. This reflects the generally stereotypical nature of language in the Clarissa domain. The specialised unification grammar contains 491 lexical and 162 non-lexical rules; Table 12.2 shows examples of specialised grammar rules, together with associated frequencies of occurrence in the training corpus. The specialised grammar is compiled into a CFG language model containing 427 non-terminal symbols and 1999 context-free productions. Finally, the training corpus is used a second time to perform probabilistic training of the CFG language model using the Nuance `compute-grammar-probs` utility and the resulting probabilistic version of the language model is compiled into a recognition package using the `nuance-compile` utility.

Table 12.2 Examples of rules in specialised version of Clarissa grammar

Rule	Freq	Example
$S \rightarrow V \text{ NP}$	606	“[[delete] [the voice note]]”
$\text{NP} \rightarrow \text{step NUMBER}$	481	“go to [[step] [four]]”
$\text{SIGMA} \rightarrow \text{INTERJECTION NP}$	344	“[[no i meant] [four point one]]”
$S \rightarrow V \text{ NP POST_MODS}$	295	“[[set] [timer] [for two minutes]]”
$\text{POST_MODS} \rightarrow \text{P NP}$	228	“set alarm [[at] [three zero six]]”
$V \rightarrow \text{go back}$	108	“[go back]”
$\text{NP} \rightarrow \text{the voice note}$	40	“cancel [the voice note]”
$S \rightarrow V \text{ P NP POST_MODS}$	28	“[[go] [to] [the note] [before step one]]”

“Rule” = context-free skeleton of rule; “Freq” = Frequency of occurrence in the training corpus

Semantic representations produced by the Clarissa grammar are general domain-independent logical forms. By construction, the same representations are produced by the specialised grammar and the derived recogniser. The Alterf package is used to convert these general representations into unordered lists of semantic atoms; a final post-processing stage transforms Alterf output into the “dialogue moves” used as input by the dialogue manager. Figure 12.2 shows examples of these different levels of representation.

```

Surface
    “no i said go to step five point three”

Logical form
[[interjection, correction],
 [imp,
  form(imperative,
    [[go,
      term(pro, you, []),
      [to, term(null, step, [[number, [decimal,5,3]]]]]]]]]]

Alterf output
[correction, go_to, [decimal,5,3]]

Dialogue move
correction(go_to([decimal,5,3]))

```

Fig. 12.2 Examples showing different levels of representation for a Clarissa utterance. We show the surface words, the general logical form produced by the Regulus grammar and derived recogniser, the list of semantic atoms produced by Alterf and the dialogue move

Recall that the Alterf algorithm requires definition of feature extraction rules, so that it can then be trained to acquire associations between extracted features and co-occurring semantic atoms. We have experimented with three different kinds of feature extraction rules: surface N-grams, hand-coded surface patterns and hand-coded logical form patterns. Unsurprisingly, we discovered that surface N-gram features were not particularly reliable. We then implemented two more sets of feature extraction rules, which defined different types of hand-coded patterns. The first set consists of conventional phrasal patterns over the tokenised recognition string, written in a simple string-matching language; the second set encodes structural patterns in the logical form. Examples of string-based and logical-form-based patterns are shown in Fig. 12.3. The version of Clarissa described here has 216 string-based patterns and 305 logical-form-based patterns. The patterns have been developed and debugged using the 3297 utterance training corpus: on this corpus, each set of patterns has a classification error rate of about 0.5%.

12.4.3 Evaluating Speech Understanding Performance

There are a large number of types of experiments which we could potentially have carried out to evaluate the speech understanding architecture. Given limited resources, we decided to focus on two main questions:

String based patterns

```
% "decrease" followed by "volume" → decrease_volume
surface_pattern([decrease, '...', volume], decrease_volume).
% "back" not following "go" and at the end → previous_line
surface_pattern([not_word(go),back,'*end*'], previous_line).
% "put" followed by "voice note" → record_voice_note
Surface_pattern([put,'...',voice,note], record_voice_note).
```

Logical-form-based patterns

```
% "decrease" with object "volume" → decrease_volume
lf_pattern([decrease, _, term(_, volume,_)], decrease_volume).
% "back" used as an interjection → previous_line
lf_pattern([interjection,back], previous_line, back).
% "put" with "voice_note" → record_voice_note
lf_pattern([put, _, term(_, voice_note,_),_], record_voice_note).
```

Fig. 12.3 Examples of string-based and logical-form-based patterns used in Clarissa

- How does the Regulus grammar-based framework compare against a more conventional framework using a class N-gram language model and a set of phrase-spotting rules?
- How do different types of features compare against each other? In particular, are logical-form-based patterns more effective than string-based or N-gram patterns, and is it useful to combine several types of pattern?

The next issue to resolve is the choice of appropriate performance metrics and test data. Given that we are essentially interested in speech understanding performance, our primary metric is semantic error rate. The choice of appropriate test data was unfortunately not straightforward. Ideally, we would have liked to test on astronaut subjects, but the virtual impossibility of getting significant quantities of astronaut time forced us to adopt a compromise. We compared the small amount of astronaut data we were able to obtain against the results of a pilot study using naïve subjects with no previous exposure to the system, but this revealed a serious mismatch. The astronauts were very familiar both with the procedure-following task and with use of controlled language, and moreover had a strong motivation to learn to use the system; the naïve subjects had neither the relevant background, nor any particular reason to want to acquire the relevant skills. The performance figures reflected this imbalance, with the astronauts scoring enormously better than nearly all of the naïve subjects. We obtained a much closer fit against the data recorded by system developers during the course of the project. Although the developers know the system a little better than the astronaut users, our intuitive observation was that the difference was not large, and that the astronauts would probably catch up after only a relatively short period of use.

The experiments we describe are thus based on a sample of 8158 in-domain utterances (23,369 words) collected and transcribed during the course of the project.

By “in-domain”, we mean here that the utterances expressed commands meaningful in the context of the Clarissa task and that the system should ideally have responded to them. Behaviour on out-of-domain data, where the best response is to ignore the utterance, is considered in Section 12.5. The data had not previously been used for development purposes and can be considered as unseen.

In order to compare the Regulus-based recogniser with a conventional architecture, we used the Nuance SayAnything© tool and the same 3297 utterance training set to build a standard class N-gram model. A summary of performance results for the two recognisers is shown in Table 12.3. In this table, WER and SER represent Word Error Rate and Sentence Error Rate, and SemER represents the semantic error rate for the best Alterf configuration tested which involved the relevant recogniser; we tested configurations involving logical-form-based features, string-based features and combinations of the two. Detailed results are presented in Tables 12.7 and 12.8 in the appendix.

Table 12.3 WER, SER and SemER for GLM and SLM recognisers trained on the same data. SemER refers to semantic understanding performance with the best Alterf configuration for that recogniser

Recogniser	WER (%)	SER (%)	SemER (%)
GLM	6.3	9.8	6.0
SLM	7.4	12.4	9.6

As can be seen, the GLM performs considerably better than the SLM. The best SLM version has a semantic error rate of 9.6%, while the best GLM version has an error rate of 6.0%, a relative improvement of 37%. Part of this is clearly due to the fact that the GLM has better WER and SER than the SLM. However, the relative improvement in WER is only 15% (7.4% versus 6.3%), and that in SER is 21% (12.4% versus 9.8%).

The larger improvement by the GLM version at the level of semantic understanding is most likely accounted for by the fact that it is able to use logical-form-based features, which are not accessible to the SLM version. Although logical-form-based features do not appear to be intrinsically more accurate than string-based features, the fact that they allow tighter integration between semantic understanding and language modelling is intuitively advantageous.

12.5 Rejecting User Speech

Section 12.4 evaluates the performance of the speech understanding component of the system, assuming that the task can be described as that of taking an input utterance constituting a valid system command and assigning a correct semantic interpretation to it. This characterisation, however, omits an important dimension. At least some of the time, we know that input speech will not consist of valid system commands. The most common reason for this will be *cross-talk*: the user may break

off addressing the system to converse with another person, but their remarks will still be picked up by the microphone and subjected to speech recognition.³ There is also the possibility that the user will say something that is outside the system's sphere of competence, either through carelessness or because of uncertainty about its capabilities.

We will refer to the choice between accepting and rejecting output from the speech recogniser as the "accept/reject decision". Usually, the speech recogniser produces a confidence score as part of its output, and the accept/reject decision is made simply by rejecting utterances whose confidence score is under a specified threshold. A recent example is [13], which reported an accuracy of 9.1% on cross-talk identification using the confidence threshold method.

In this section, we show that adapted versions of standard kernel-based methods from the document classification literature can substantially improve on the baseline confidence threshold approach.

12.5.1 The Accept/Reject Decision Task

We need to introduce suitable metrics for evaluating performance on the accept/reject task. We can define the following three categories of utterance:

Type A: Utterances directed at the system, for which a good semantic interpretation was produced.

Type B: Utterances directed at the system, and to which the system could in principle respond, but for which the semantic interpretation produced was incorrect. Usually, this is due to faulty recognition.

Type C: Utterances not directed at the system, or directed at the system but to which the system has no adequate way to respond.

We would like to accept utterances in the first category, and reject utterances in the second and third. If we want to measure performance on the accept/reject task, the most straightforward approach is a simple classification error rate. Ultimately, however, what we are most interested in is measuring performance on the top-level speech understanding task, which includes the recognition task, the semantic interpretation task and the accept/reject task described here.

Constructing a sensible metric for the top-level task involves taking account of the fact that some errors are intuitively more serious than others. In a Type A error, the user can most often correct by simply repeating themselves. In a Type B error, the user will typically have to wait for the system response, realise that it is inappropriate, and then undo or correct it, a significantly longer operation. A Type C error has all the drawbacks of a Type B error, and affects not only the user, but also

³For long side-conversations, the user has the option of using the "suspend" command (cf. Section 2.1) to pause recognition.

the person with whom they are having the side conversation. This analysis suggests that errors should not all be counted equally, but rather be weighted to define a loss function. It is not easy to give clear justifications for particular choices of weights. A reasonable candidate, which we use in the rest of this chapter, is to assign weights of 1 to Type A, 2 to Type B and 3 to Type C. We divide the total weighted score over a given corpus by the maximum possible loss value, giving a normalised loss function whose value is always between 0 and 1.

The discrepancy between the local loss function associated with the classifier and the task-level loss function raises the issue of how to align classifier objectives with task-level ones. In this initial work, we simplify the problem by decoupling the speech understanding and accept/reject subtasks, using separate metrics. Since the weighting on the task metric penalises false accepts more heavily than false rejects, we introduce an asymmetric loss function on the classifier score, which weights false accepts twice as heavily as false rejects. We will refer to this as the u_2 metric, and use it as the filtering task metric to compare different parameter settings.

12.5.2 An SVM-Based Approach

There are several information sources which could potentially be used as input to the accept/reject classification problem. So far, we have limited ourselves to the surface result returned by the Nuance recogniser, which consists of a list of words, each tagged by a numerical confidence value.

As already noted, the usual way to make the accept/reject decision is through a simple threshold on the average confidence score; the Nuance confidence scores are of course designed for exactly this purpose. Intuitively, however, it should be possible to improve the decision quality by also taking account of the information in the recognised words. Looking through our data, we could see examples which had high enough average confidence scores to be accepted, but contained phrases that were very implausible in the context of the application. In the other direction, we noticed that the confidence scores for several common short utterances (in particular “yes” and “no”) appeared for some reason to be artificially depressed; these utterances could safely be accepted with a lower confidence threshold than usual. We wanted a method that could identify and exploit patterns like these.

The accept/reject decision is clearly a kind of document classification problem. It is well known [14] that margin-based classifiers, especially Support Vector Machines (SVM), get good results for this kind of task. Due to its stability and ready availability for research purposes, we used Joachim’s SVM-light platform [15]. There were two key issues we had to address in order to apply SVM-light to the new task. First, we had to construct a suitable kernel function, which would define similarity between two recognition results; this kernel would need to reflect the fact that the objects were not text documents, but speech documents which had been passed through a noisy channel. Second, we had to take account of the asymmetric nature of the cost function.

12.5.2.1 Choosing a Kernel Function

As recognition results consist of sequences of words tagged with confidence scores, the kernel function must be based on this information. (As usual with kernel methods, choosing the kernel function essentially corresponds to choosing the representation in a traditional architecture.) The simple bag-of-words representation, as traditionally used to represent written documents, loses the important confidence score information and is unlikely to produce good results. Preliminary experiments not reported here confirmed that this was indeed the case.

We consequently modified the bag-of-words representation, so as to weight each word by its associated confidence score. This means that the recognition result is initially mapped into a vector in a V -dimensional space, where V is the vocabulary size; the component on each dimension is the confidence score on the relevant word, or zero if the word did not occur in the recognition result. If a word occurs multiple times, the confidence scores for each occurrence are summed; it is possible that a “max” or “min” operator would have been more appropriate, but multiple word occurrences only obtained in about 1% of the utterances. We also added an extra component to the vector, which represented the average of all the word confidence scores.

By using different kernel functions on the vector space representation, we were able to take into account various kinds of lexical patterns. In particular, nonlinear polynomial kernels of degree N encode unordered co-occurrences of N words (unordered gappy N -grams). In practice, we found that values of N higher than 2 gave no additional gains. We also experimented with string subsequence kernels [16], but these failed to improve performance either, while being computationally much more expensive. In the experiments reported below, we thus restrict ourselves to linear and quadratic kernels, representing unigram and unordered non-local bigram information.

12.5.2.2 Making the Cost Function Asymmetric

There are at least two ways to introduce asymmetric costs in standard SVM implementations. Recall that SVM is optimising a mixed criterion, which combines classification errors on a training set and a measure of complexity related to the margin concept [17, p. 220]. The simplest method is to penalise the distance to the margin for misclassified examples more highly for false positives than for false negatives. This can be done directly using the j parameter in the SVM-light implementation.

The drawback to the first approach is, however, that the algorithm is not really optimising the utility function, but a more or less related quantity [18]; this prompted us to investigate the use of calibration as well. Calibration [19] aims at transforming SVM scores into posterior probabilities in a way that is independent from the class priors (basically $P(s(x)|Class)$, where $s(x)$ is the score associated with observation x). The optimal Bayesian decision can then be adapted, once the new class priors are known ($P(Class)$), as well as error costs. For a binary problem

(accept/reject) with equal cost of errors for all negative examples, when the class distribution can be assumed to be the same on both training and test sets, it is sufficient to approximate $P(\text{Class} = A|s(x))$, as the optimal Bayes decision is then based on minimising the expected loss function.

In our case, the u_2 function penalises false accepts twice as heavily as false rejects: the optimal decision rule is thus to accept the utterance if

$$2P(\text{Class} = B \text{ or } C|s(x)) < P(\text{Class} = A|s(x))$$

or, equivalently, if $P(\text{Class} = A|s(x)) > 2/3$. We used Isotonic Regression [20] to realise the mapping from SVM scores into approximate posterior probabilities.

12.5.3 Experiments

A corpus of 10,409 recorded and labelled spoken utterances was used in order to investigate the impact of three factors on classification and task performance:

Classifier We used three types of classifiers: a simple threshold on the average confidence score; an SVM with a linear kernel; and an SVM with a quadratic kernel. The SVM classifiers used a set of features consisting of the average confidence score together with the weighted bag of words over the total vocabulary.

Asymmetric Error We used two different techniques to deal with asymmetric error costs: the j intrinsic parameter of SVM-light and the recalibration procedure using Isotonic Regression. Recall that recalibration aims at optimising the u_2 loss function at SVM classification level, and not the task-level loss function. Without recalibration, the decision threshold on SVM scores is 0.

Recognition We contrasted GLM and SLM methods, specifically using the best GLM-based recogniser (G-4) and the best SLM-based recogniser (S-3) from Table 12.7.

For each choice of parameters, we performed 10 random splits (training/test sets) of the initial set of labelled utterances, learned the model on the training sets and evaluated the loss functions on the corresponding test sets. The final scores were obtained by averaging the loss functions over all 10 runs. Summary results, contrasting performance for the baseline and best SVM-based methods, are shown in Table 12.4. Detailed results are presented in Table 12.9 in the appendix.

The SVM-based method is clearly very much better than the baseline confidence threshold method. The average classification error falls from 9.4% for the best baseline configuration to 5.5% for the best SVM-based configuration, a relative improvement of 42%; in particular, the false accept rate for cross-talk and out-of-domain utterances improves from 8.9% (close to the 9.1% cited in [13]) to 4.7%, a 47% relative improvement, while the error rates on the other individual classes also

Table 12.4 Summary comparing best baseline (raw confidence score) and best SVM-based methods for making the accept/reject decision

Method	Classif. (%)	False acc. (%)	Task (%)
Baseline	9.4	8.9	7.0
SVM	5.5	6.1	5.4

“Classif.” = classification error; “False acc.” = false accept rate;
 “Task” = performance on task metric

improve. On the task performance metric, the improvement is from 7.0 to 5.4%, or 25% relative. In the appendix, we discuss the contributions made to this result by the different techniques used.

12.6 Side-Effect Free Dialogue Management

Most spoken dialogue systems have some notion of context, which typically will include the preceding dialogue, the current state of the task or both. For example, consider the reaction of a simulated robot to the command “Put it on the block”. This might include both remembering a recently mentioned object to serve as a referent for “it” (dialogue context) and looking around the current scene to find an object to serve as a referent for “the block” (task context). The Dialogue Manager (DM) will thus both access the current context as an input, and update it as a result of processing utterances.

Contextual information is usually distributed through the DM as part of the current program state. This means that processing of an input utterance involves at least some indirect side-effects, since the program state will normally be changed. If the DM makes procedure calls to the output module, there will also be direct side-effects in the form of exterior actions. As every software engineer knows, side-effects are normally considered a bad thing. They make it harder to design and debug systems, since they render interactions between modules opaque. The problem tends to be particularly acute when performing regression testing and evaluation; if a module’s inputs and outputs depend on side-effects, it is difficult or impossible to test that module in isolation. The upshot for spoken language systems is that it is often difficult to test the DM except in the context of the whole system.

In this section, we describe an architecture which directly addresses the problems outlined above, and which has been implemented in Clarissa. There are two key ideas. First, we split the DM into two pieces: a large piece, comprising nearly the whole of the code, which is completely side-effect free, and a small piece which is responsible for actually performing the actions. Second, we adopt a consistent policy about representing contexts as objects. Both discourse- and task-oriented contextual information, without exception, are treated as part of the context object.

12.6.1 Side-Effect Free Dialogue Management

Clarissa implements a minimalist dialogue management framework, partly based on elements drawn from the TRIPS [21] and TrindiKit [22] architectures. The central concepts are those of *dialogue move*, *information state* and *dialogue action*. At the beginning of each turn, the dialogue manager is in an information state. Inputs to the dialogue manager are by definition dialogue moves, and outputs are dialogue actions. The behaviour of the dialogue manager over a turn is completely specified by an *update function* f of the form.

$$f : State \times Move \rightarrow State \times Actions$$

Thus if a dialogue move is applied in a given information state, the result is a new information state and a set of zero or more dialogue actions.

In the Clarissa system, most of the possible types of dialogue moves represent spoken commands. For example, `increase(volume)` represents a spoken command like “increase volume” or “speak up”. Similarly, `go_to(step(2,3))` represents a spoken command like “go to step two point three”. The dialogue move `undo` represents an utterance like “undo last command” or “go back”. Correction utterances are represented by dialogue moves of the form `correction(X)`; so for example

```
correction(record(voice_note(step(4))))
```

represents an utterance like “no, I said record a voice note on step four”. There are also dialogue moves that represent non-speech events. For example, a mouse-click on the GUI’s “next” button is represented as the dialogue move `gui_request(next)`. Similarly, if an alarm goes off at time T , the message sent is represented as a dialogue move of the form `alarm_triggered(T)`. The most common type of dialogue action is a term of the form `say(U)`, representing a request to speak an utterance abstractly represented by the term U . Other types of dialogue actions include modifying the display, changing the volume and so on.

The information state is a vector, which in the current version of the system contains 26 elements. Some of these elements represent properties of the dialogue itself. In particular, the `last_state` element is a back-pointer to the preceding dialogue state and the `expectations` element encodes information about how the next dialogue move is to be interpreted. For example, if a yes/no question has just been asked, the `expectations` element will contain information determining the intended interpretation of the dialogue moves `yes` and `no`.

The novel aspect of the Clarissa DM is that all *task* information is also uniformly represented as part of the information state. Thus for example the `current_location` element holds the procedure step currently being executed, the `current_alarms` element lists the set of alarms currently set, associating each alarm with a time and a message, and the `current_volume` element represents the output volume, expressed as a percentage of its maximum value. Putting the task information into the information state has the desirable

consequence that actions whose effects can be defined in terms of their effect on the information state need not be specified directly. For example, the update rule for the dialogue move `go_to(Loc)` specifies among other things that the value of `current_location` element in the output dialogue state will be `Loc`. The specific rule does not also need to say that an action needs to be produced to update the GUI by scrolling to the next location; that can be left to a general rule, which relates a change in the `current_location` to a scrolling action.

More formally, what we are doing here is dividing the work performed by the update function f into two functions, g and h . g is of the same form as f , i.e.

$$g : State \times Move \rightarrow State \times Actions$$

As before, this maps the input state and the dialogue move into an output state and a set of actions; the difference is that this set now only includes the *irreversible* actions. The remaining work is done by a second function

$$h : State \times State \rightarrow Actions$$

which maps the input state S and output state S' into the set of reversible actions required to transform S into S' ; the full set of output actions is the union of the reversible and the irreversible actions. The relationship between the functions f , g and h can be expressed as follows. Let S be the input state and M the input dialogue move. Then if $g(S, M) = \langle S', A_1 \rangle$, and $g(S, S') = A_2$, we define $f(S, M)$ to be $\langle S', o(A_1 \cup A_2) \rangle$, where o is a function that maps a set of actions into an ordered sequence.

In Clarissa, h is implemented concretely as the set of all solutions to a Prolog predicate containing one clause for each type of difference between states which can lead to an action. Thus we have for example a clause which says that a difference in the `current_volume` elements between the input state and the output state requires a dialogue action that sets the volume; another clause which says that an alarm time present in the `current_alarms` element of the input state but absent in the `current_alarms` element requires a dialogue action which cancels an alarm, and so on. The ordering function o is defined by a table which associates each type of dialogue action with a priority; actions are ordered by priority, with the function calls arising from the higher priority items being executed first. Thus for example the priority table defines a `load_procedure` action as being of higher priority than a `scroll` action, capturing the requirement that the system needs to load a procedure into the GUI before it can scroll to its first step.

12.6.2 Specific Issues

12.6.2.1 “Undo” and “Correction” Moves

As already noted, one of the key requirements for Clarissa is an ability to handle “undo” and “correction” dialogue moves. The conventional approach, as for

example implemented in the CommandTalk system [23], involves keeping a “trail” of actions, together with a table of inverses which allow each action to be undone. The extended information state approach described above permits a more elegant solution to this problem, in which corrections are implemented using the g and h functions together with the `last_state` element of the information state. Thus if we write u for the “undo” move, and $l(S)$ to denote the state that S 's `last_state` element points to, we can define $g(S, u)$ to be $\langle l(S), \emptyset \rangle$, and hence $f(S, u)$ will be $\langle l(S), o(h(S, l(S))) \rangle$. Similarly, if we write $c(M)$ for the move which consists of a correction followed by M , we can define $f(S, c(M))$ to be $\langle S', o(A \cup h(S, S')) \rangle$, where S' and A are defined by $g(l(S), M) = \langle S', A \rangle$.

In practical terms, there are two main payoffs to this approach. First, code for supporting undos and corrections shrinks to a few lines, and becomes trivial to maintain. Second, corrections are in general faster to execute than they would be in the conventional approach, since the h function directly computes the actions required to move from S to S' , rather than first undoing the actions leading from $l(S)$ to S , and then redoing the actions from $l(S)$ to S' . When actions involve non-trivial redrawing on the visual display, this difference can be quite significant.

12.6.2.2 Confirmations

Confirmations are in a sense complementary to corrections. Rather than making it easy for the user to undo an action they have already carried out, the intent is to repeat back to them the dialogue move they appear to have made, and give them the option of not performing it at all. Confirmations can also be carried out at different levels. The simplest kind of confirmation echoes the exact words the system believed it recognised. It is usually, however, more useful to perform confirmations at a level which involves further processing of the input. This allows the user to base their decision about whether to proceed not merely on the words the system believed it heard, but also on the actions it proposes to take in response.

The information state framework also makes possible a simple approach to confirmations. Here, the key idea is to compare the current state with the state that would arise after responding to the proposed move and repeat back a description of the difference between the two states to the user. To write this symbolically, we start by introducing a new function $d(S, S')$, which denotes a speech action describing the difference between S and S' , and write the dialogue moves representing “yes” and “no” as y and n respectively. We can then define $f_{\text{conf}}(S, M)$, a version of $f(S, M)$ which performs confirmations, as follows. Suppose we have $f(S, M) = \langle S', A \rangle$. We define $f_{\text{conf}}(S, M)$ to be $\langle S_{\text{conf}}, d(S, S') \rangle$, where S_{conf} is constructed so that $f(S_{\text{conf}}, y) = \langle S', A \rangle$ and $f(S_{\text{conf}}, n) = \langle S, \emptyset \rangle$. In other words, S_{conf} is by construction a state where a “yes” will have the same effect as M would have had on S if the DM had proceeded directly without asking for a confirmation, and where a “no” will leave the DM in the same state as it was before receiving M .

There are two points worth noting here. First, it is easy to define the function f_{conf} precisely because f is side-effect free; this lets us derive and reason about the *hypothetical* state S' without performing any external actions. Second, the function $d(S, S')$ will in general be tailored to the requirements of the task, and will describe

relevant differences between D and S' . In Clarissa, where the critical issue is which procedure steps have been completed, $d(S, S')$ describes the difference between S and S' in these terms, for example saying that one more step has been completed, or three steps skipped.

12.6.2.3 Querying the Environment

A general problem for any side-effect free dialogue management approach arises from the issue of querying the environment. If the DM needs to acquire external information to complete a response, it may at first glance seem that the relationship between inputs and output can no longer be specified as a pure function, since the information gathering actions will constitute side-effects.

The framework can, however, be kept declarative by splitting up the DM's response into two turns. Suppose that the DM needs to read a data file in order to respond to the user's query. The first turn responds to the user query by producing an action request to read the file and report back to the DM and an output information state in which the DM is waiting for a dialogue move reporting the contents of the file. The second turn responds to the file-contents reporting action by using the new information to reply to the user. The actual side-effect of reading the file occurs outside the DM, in the space between the end of the first turn and the start of the second. Variants of this scheme can be applied to other cases in which the DM needs to acquire external information.

12.6.2.4 Regression Testing and Evaluation

Any substantial software project needs regular regression testing, where a library of previously working examples is tested on new versions of the system to ensure that changes have not broken old functionality. Regression testing and evaluation on context-dependent dialogue systems is a notoriously messy task. The problem is that it is difficult to assemble a reliable test library, since the response to each individual utterance is in general dependent on the context produced by the preceding utterances. If an utterance early in the sequence produces an unexpected result, it is usually impossible to know whether results for subsequent utterances are meaningful.

In our framework, regression testing of contextually dependent dialogue turns is unproblematic, since the input and output contexts are well-defined objects. We have been able to construct substantial libraries of test examples, where each example consists of a 4-tuple $\langle \text{InState}, \text{DialogueMove}, \text{OutState}, \text{Actions} \rangle$. These libraries remain stable over most system changes, except for occasional non-downward-compatible redesigns of the dialogue context format, and have proved very useful.

12.7 Results of the On-Orbit Test

Clarissa was first used on the ISS by Expedition 11 Science Officer and Flight Engineer John Phillips on June 27, 2005. To the best of our knowledge, this is

the first ever use of a spoken dialogue system in space. During the test, Phillips completed the interactive Clarissa training procedure, which exercises all the main system functionality; this procedure contains 50 steps, and took 25 min to complete. Table 12.5 summarises performance per step.

Table 12.5 Performance per step for on-orbit test. One step had problems both with background speech and misunderstandings about command syntax

Condition	#Steps
No problems	45
Bad recognition due to background speech	4
Bad recognition due to misunderstandings about command syntax	2
All steps	50

Of the 50 procedure steps, 45 were completed without incident. In four steps, Clarissa suffered from speech recognition problems, apparently due to the fact that loud speech from Russian Mission Control was coming from a speaker very close to Phillips. In all but one of these steps, the system simply failed to respond, and Phillips was able to correct the problem by repeating himself. In two steps, the training procedure was insufficiently clear about explaining the correct command syntax, and Phillips attempted to phrase requests in ways not acceptable to the recogniser. In the first of these steps (entry of numbers into a table), Phillips quickly ascertained that the recogniser did not permit negative numbers, and completed the step. In the second (setting an alarm), Phillips was unable to find the correct syntax to define the alarm time. This was the only step that was not completed successfully.

While Phillips was navigating the training procedure, the recogniser recorded 113 separate audio files: most of these contained spoken commands, but some were just background noise. Table 12.6 breaks down performance by files. Ninety-nine of the 113 files produced appropriate responses. In 84 cases, the file contained a command, and all the words were recognised correctly. In another four cases, the file was again a command, at least one word was misrecognised, but the system still understood and responded correctly. In 11 cases, the file contained non-command content

Table 12.6 Performance per audio file for on-orbit test

Condition	#Steps
Recognised exactly and understood	84
At least one word misrecognised, but understood	4
Non-command, correctly ignored	11
Appropriate responses	99
No recognition	9
Incorrect recognition	5
Inappropriate responses	14
All responses	113

(usually background noise), which was correctly ignored. Fourteen files produced inappropriate responses. In nine cases, the system failed to respond at all to a command, and in another five it responded incorrectly. The reasons for these problems are described above.

Both Phillips and the Clarissa team considered that the system performed very creditably during its first test.

12.8 Conclusion

We have presented a detailed description of a non-trivial spoken dialogue system, which carries out a useful task in a very challenging environment. In the course of the project, we have addressed several general problems and developed what we feel are interesting and novel solutions. We conclude by summarising our main results.

12.8.1 Procedures

The problem of converting written procedures into voice-navigable documents is not as simple as it looks. A substantial amount of new information needs to be added, in order to make explicit the instructions which were implicit in the original document and left to the user's intelligence. This involves adding explicit state to the reading process. Jumps in the procedure can leave the state inconsistent, and it is not trivial to ensure that the system will always be able to recover gracefully from these situations.

12.8.2 Recognition

For this kind of task, there is reasonable evidence that grammar-based recognition methods work better than statistical ones. The extra robustness of statistical methods does not appear to outweigh the fact that the grammar-based approach permits tighter integration of recognition and semantic interpretation. Speech understanding performance was substantially better with the grammar-based method. We were able to make a clear comparison between the two methods because we used a carefully constructed methodology which built the grammar-based recogniser from a corpus, but there is no reason to believe that other ways of building the grammar-based recogniser would have led to inferior results.

12.8.3 Response Filtering

The SVM-based method for performing response filtering that we have developed is considerably better than the naïve threshold-based method. It is also completely

domain-independent, and offers several possibilities for further performance gains. We consider this to be one of the most significant research contributions made by the project.

We are currently extending the work described here in two different directions. First, we are implementing better calibration models (cf. Section 5.2.2), which in particular will allow us to relax the assumption that the class distributions are the same in the training and test data; second, we are investigating the use of word-lattice recognition hypotheses and rational kernels [24, 25]. Initial results show that both these ideas have the potential to improve performance, though the gains so far have been fairly modest.

12.8.4 Dialogue Management

The fully declarative dialogue management framework that we have implemented is applicable to any domain, like ours, where the dialogue state can be fully specified. If this condition is met, our method is simple to implement, and offers a clean and robust treatment of correction and confirmation moves. Perhaps even more significantly, it also permits systematic regression testing of the dialogue management component as an isolated module.

12.8.5 General

The challenge of creating a dialogue system to support procedure execution on the ISS posed a number of interesting research problems. It also demanded production of a prototype system robust enough to pass the rigorous software approval process required for deployment in space. We have now reached a point in the project where we think we can claim that the research problems have been solved. The system as a whole performs solidly, and people both inside and outside NASA generally experience it as a mature piece of production-level software. This includes several astronauts with first-hand experience of using other procedure viewers during space missions.

As described earlier, we have been able to carry out detailed testing of the individual system components. We would now like to address the bottom-line question, namely whether Clarissa actually is an ergonomic win compared to a conventional viewer. Obviously, hands- and eyes-free operation is an advantage. This, however, must be balanced against the fact that reading from a screen is faster than reading aloud.

Based on initial studies we have carried out within the development group, our impression is that tight usability experiments are not straightforward to design; an astronaut who is using a speech-enabled procedure reader has to make non-trivial changes in his normal work practices in order to fully exploit the new technology.

In other words, the challenge is now to learn how to work efficiently together with a voice-enabled automatic assistant.

12.8.5.1 A Note on Versions

The lengthy certification and sign-off process required by NASA means that software actually deployed in space typically lags behind the latest development version, and for these reasons the version of Clarissa described in this chapter differs in some respects from the one tested on the ISS. In general, our focus is on presenting what we think are the best solutions to the design problems we have encountered, rather than on strict historical accuracy.

Acknowledgments Work at ICSI, UCSC and RIACS was supported by NASA Ames Research Center internal funding. Work at XRCE was partly supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. Several people not credited here as co-authors also contributed to the implementation of the Clarissa system: among these, we would particularly like to mention John Dowding, Susana Early, Claire Castillo, Amy Fischer and Vladimir Tkachenko. This publication only reflects the authors' views.

Appendix: Detailed Results for System Performance

This appendix provides detailed performance results justifying the claims made in the main body of the chapter. We divide it into two parts: the first concerns the recognition task and the second the accept/reject task.

The Recognition Task

Table 12.7 presents the results of experiments contrasting speech understanding performance for the Regulus-based recogniser and the class N-gram recogniser, using several different sets of Alterf features (cf. Section 4.3). For completeness, we also present results for simulated perfect recognition, i.e. using the reference transcriptions. We used six different sets of Alterf features:

N-grams: N-gram features only.

LF: Logical-form-based patterns only.

String: String-based patterns only.

String + LF: Both string-based and logical-form-based patterns.

String + N-grams: Both string-based and N-gram features.

String + LF + N-grams: All types of features.

The general significance of these results is discussed at the end of Section 12.4. It is interesting to note that the combination of logical-form-based features and string-based features outperforms logical-form-based features alone (rows G-4 and G-2). Although the difference is small (6.0% versus 6.3%), a pairwise comparison shows

Table 12.7 Speech understanding performance for 8158 test sentences recorded during development, on 13 different configurations of the system

Version	Rec	Features	Rejected (%)	Incorrect (%)	All Errors (%)
T-1	Text	N-grams	7.3	5.9	13.2
T-2	Text	LF	3.1	0.5	3.6
T-3	Text	String	2.2	0.8	3.0
T-4	Text	String+LF	0.8	0.8	1.6
T-5	Text	String+LF+N-grams	0.4	0.8	1.2
G-1	GLM	N-grams	7.4	9.7	17.1
G-2	GLM	LF	1.4	4.9	6.3
G-3	GLM	String	2.9	4.8	7.7
G-4	GLM	String+LF	1.0	5.0	6.0
G-5	GLM	String+LF+N-grams	0.7	5.4	6.1
S-1	SLM	N-grams	9.6	11.9	21.5
S-2	SLM	String	2.8	7.4	10.2
S-3	SLM	String+N-grams	1.6	8.0	9.6

“Rec” = type of recognition: either simulated perfect recognition (“Text”), recognition using the Regulus-derived grammar-based language model (“GLM”) or recognition using a class N-gram language model (“SLM”). “Features” = Alerf features used. “Rejected” = Utterances given no semantic interpretation. “Incorrect” = utterances given incorrect semantic interpretation. “All Errors” = Sum of “Rejected” and “Incorrect”

that it is significant at the 1% level according to the McNemar sign test. There is no clear evidence that N-gram features are very useful. This supports the standard folklore result that semantic understanding components for command and control applications are more appropriately implemented using hand-coded phrase-spotting patterns than general associational learning techniques.

Table 12.8 presents a breakdown of speech understanding performance, by utterance length, for the best GLM-based and SLM-based versions of the system. There are two main points to note here. First, speech understanding performance remains

Table 12.8 Speech understanding performance, broken down by utterance length, for the best GLM-based and SLM-based versions of the system (cf. Table 12.7). Results are omitted for the small group of utterances of length 10 or more

Length	#Utts	Best GLM (G-4)			Best SLM (S-3)		
		WER (%)	SER (%)	SemER (%)	WER (%)	SER (%)	SemER (%)
1	3049	5.7	3.5	2.5	6.3	4.2	3.5
2	1589	12.0	12.0	8.7	14.6	18.4	14.6
3	950	7.2	12.8	7.2	10.4	15.2	15.4
4	1046	7.6	14.8	9.9	7.7	15.6	14.7
5	354	5.7	14.4	9.0	6.1	19.8	10.8
6	543	2.8	11.1	7.2	4.1	15.3	9.8
7	231	3.0	16.0	3.5	4.6	19.5	6.5
8	178	4.4	14.6	4.5	3.6	16.3	5.7
9	174	3.9	20.1	9.2	4.0	20.7	10.3

respectable even for the longer utterances; second, the performance of the GLM-based version is consistently better than that of the SLM-based version for all utterance lengths.

12.0.1 The Accept/Reject Task

Table 12.9 presents detailed results for the experiments on response filtering described in Section 12.5. All conclusions were confirmed by hypothesis testing, using the Wilcoxon rank test, at the 5% significance level. In the remainder of this section, we assess the impact made by individual techniques.

Table 12.9 Performance on accept/reject classification and the top-level speech understanding task, on 12 different configurations of the system

ID	Rec	Classifier	j	A (%)	B (%)	C (%)	All (%)	u_2 (%)	Task (%)
ST-1	SLM	Threshold	1.0	5.5	59.1	16.5	11.8	15.1	10.1
SL-1	SLM	Linear	1.0	2.8	37.1	9.0	6.6	8.6	7.4
SL-2	SLM	Linear	0.5	4.9	30.1	6.8	7.0	8.1	7.2
SQ-1	SLM	Quad	1.0	2.6	23.6	8.5	5.5	7.0	6.9
SQ-2	SLM	Quad	0.5	4.1	18.7	7.6	6.0	7.0	7.0
SQ-3	SLM	Quad/r	1.0	4.7	18.7	6.6	6.1	6.8	6.9
GT-1	GLM	Threshold	0.5	7.1	48.7	8.9	9.4	10.7	7.0
GL-1	GLM	Linear	1.0	2.8	48.5	8.7	6.3	8.3	6.2
GL-2	GLM	Linear	0.5	4.7	43.4	6.0	6.7	7.9	6.0
GQ-1	GLM	Quad	1.0	2.7	37.9	6.8	5.3	6.7	5.7
GQ-2	GLM	Quad	0.5	4.0	26.8	6.0	5.5	6.3	5.6
GQ-3	GLM	Quad/r	1.0	4.3	28.1	4.7	5.5	6.1	5.4

“Rec” = Type of recognition: either Regulus-derived grammar-based language model (“GLM”) or class N-gram language model (“SLM”); “Classifier” = type of classifier used: “Threshold” = simple threshold on average confidence; “Linear” = SVM classifier with linear kernel; “Quad” = SVM classifier with quadratic kernel; “Quad/r” = recalibrated version of SVM classifier with quadratic kernel. “ j ” = Value of SVM-light j parameter. “A” = classifier error rate on in-domain utterances with correct semantic interpretation. “B” = Classifier error rate on in-domain utterances with incorrect or no semantic interpretation. “C” = Classifier error rate on out-of-domain or cross-talk utterances. “All” = Classifier error on all data. “ u_2 ” = Weighted average of classifier error rate using u_2 weights. “Task” = Normalised task metric loss

Kernel Types

Quadratic kernels performed better than linear (around 25% relative improvement in classification error); however, this advantage is less marked when considering the task metric (only 3–9% relative increase). Though small, the difference is statistically significant. This suggests that meaningful information for filtering lies, at least partially, in the co-occurrences of groups of words, rather than just in isolated words.

Asymmetric Error Costs

We next consider the effect of methods designed to take account of asymmetric error costs (cf. Section 12.5). Comparing GQ-1 (no treatment of asymmetric error costs) with GQ-2 (intrinsic SVM-optimisation using the j -parameter) and GQ-3 (calibration), we see that both methods produce a significant improvement in performance. On the u_2 loss function that both methods aim to minimise, we attain a 9% relative improvement when using calibration and 6% when using intrinsic SVM optimisation; on the task metric, these gains are reduced to 5% (relative) for calibration, and only 2% for intrinsic SVM-optimisation, though both of these are still statistically significant. Error rates on individual classes show that, as intended, both methods move errors from false accepts (classes B and C) to the less dangerous false rejects (class A). In particular, the calibration method manages to reduce the false accept rate on cross-talk and out-of-domain utterances from 6.8% on GQ-1 to 4.7% on GQ-3 (31% relative), at the cost of an increase from 2.7% to 4.3% in the false reject rate for correctly recognised utterances.

Recognition Methods

Using the confidence threshold method, there was a large difference in performance between the GLM-based GT-1 and the SLM-based ST-1. In particular, the false accept rate for cross-talk and out-of-domain utterances is nearly twice as high (16.5% versus 8.9%) for the SLM-based recogniser. This supports the folklore result that GLM-based recognisers give better performance on the accept/reject task.

When using the SVM-based methods, however, the best GLM-based configuration (GQ-3) performs about as well as the best SLM-based configuration (SQ-1) in terms of average classification error, with both systems scoring about 5.5%. GQ-3 does perform considerably better than SQ-1 in terms of task error (5.4% versus 6.9%, or 21% relative), but this is due to better performance on the speech recognition and semantic interpretation tasks. Our conclusion here is that GLM-based recognisers do not necessarily offer superior performance to SLM-based ones on the accept/reject task, if a more sophisticated method than a simple confidence threshold is used.

References

1. Aist, G., Dowding, J., Hockey, B. A., Hieronymus, J. (2002). An intelligent procedure assistant for astronaut training and support. In: Proc. 40th Annual Meeting of the Association for Computational Linguistics (demo track), Philadelphia, PA, 5–8.
2. Martin, D., Cheyer, A., Moran, D. (1999). The open agent architecture: a framework for building distributed software systems. *Appl. Artif. Intell.*, 13 (1–2), 92–128.
3. Nuance (2006). <http://www.nuance.com>. As of 15 November 2006.

4. Knight, S., Gorrell, G., Rayner, M., Milward, D., Koeling, R., Lewin, I. (2001). Comparing grammar-based and robust approaches to speech understanding: a case study. In: Proc. Eurospeech 2001, Aalborg, Denmark, 1779–1782.
5. Rayner, M., Hockey, B. A., Bouillon, P. (2006). Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler. CSLI, Chicago, IL.
6. Regulus (2006). <http://sourceforge.net/projects/regulus/>. As of 15 November 2006.
7. Pulman, S. G. (1992). Syntactic and semantic processing. In: Alshawi, H. (ed) The Core Language Engine, MIT, Cambridge, MA, 129–148.
8. van Harmelen, T., Bundy, A. (1988). Explanation-based generalization – partial evaluation (research note). *Artif. Intell.*, 36, 401–412.
9. Rayner, M. (1988). Applying explanation-based generalization to natural-language processing. In: Proc. Int. Conf. on Fifth Generation Computer Systems, Tokyo, Japan, 1267–1274.
10. Rayner, M., Hockey, B. A. (2003). Transparent combination of rule-based and data-driven approaches in a speech understanding architecture. In: Proc. 10th Conf. Eur. Chapter of the Association for Computational Linguistics, Budapest, Hungary, 299–306.
11. Yarowsky, D. (1994). Decision lists for lexical ambiguity resolution. In: Proc. 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, 88–95.
12. Carter, D. (2000). Choosing between interpretations. In: Rayner, M., Carter, D., Bouillon, P., Digalakis, V., Wirén, M. (eds) The Spoken Language Translator, Cambridge University Press, Cambridge, MA, 78–97.
13. Dowding, J., Hieronymus, J. (2003). A spoken dialogue interface to a geologist's field assistant. In: Proc. HLT-NAACL 2003: Demo Session, Edmonton, Alberta, 9–10.
14. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In: Proc. 10th Eur. Conf. on Machine Learning, Chemnitz, Germany, 137–142.
15. Joachims, T. (2006). <http://svmlight.joachims.org/>. As of 15 November 2006.
16. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C. (2002). Text classification using string kernels. *J. Machine Learn. Res.*, 2, 419–444.
17. Shawe-Taylor, J., Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.
18. Navia-Vázquez, A., Pérez-Cruz, F., Artés-Rodríguez, A., Figueiras-Vidal, A. R. (2004). Advantages of unbiased support vector classifiers for data mining applications. *J. VLSI Signal Process. Syst.*, 37 (1–2), 1035–1062.
19. Bennett, P. (2003). Using asymmetric distributions to improve text classifier probability estimates. In: Proc. 26th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, Toronto, Ontario, 111–118.
20. Zadrozny, B., Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In: Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Edmonton, Alberta, 694–699.
21. Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., Stent, A. (2000). An architecture for a generic dialogue shell. *Natural Language Engineering, Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering*, 6, 1–16.
22. Larsson, S., Traum, D. (2000). Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering, Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering*, 6, 323–340.
23. Stent, A., Dowding, J., Gawron, J., Bratt, E., Moore, R. (1999). The CommandTalk spoken dialogue system. In: Proc. 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, Maryland, VA, 183–190.
24. Haffner, P., Cortes, C., Mohri, M. (2003). Lattice kernels for spoken-dialog classification. In: Proc. 2003 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '03), Hong Kong, 628–631.
25. Cortes, C., Haffner, P., Mohri, M. (2004). Rational kernels: Theory and algorithms. *J. Machine Learn. Res.*, 5, 1035–1062.

Chapter 13

Military Applications: Human Factors Aspects of Speech-Based Systems

Jan M. Noyes and Ellen Haas

13.1 Introduction

When considering military applications of speech-based interactive systems, there are some which are specific to the military domain, and some which are more general, for example, office-type applications (dictation, directory and information enquiries; see Jokinen [23]) and training. The emphasis in the chapter is on the more specific military applications although some of the general applications are discussed. Two key components of speech-based interactive systems are Automatic Speech Recognition (ASR) and speech synthesis. These are extensively covered in earlier chapters, so are only considered here in terms of characteristics relevant to the military domain. A final comment concerns the definition of military. Traditionally, the military is thought of as comprising the Air Force, Army, Navy and Marine Corps. In addition, there are some peripheral activities relating to the military such as Air Traffic Control (ATC) and defence, for example, the military police and the security agencies. These are also considered in the chapter as part of the section on applications.

13.2 The Military Domain

The military domain, like any application area, has its own unique set of characteristics. When considering the use of speech-based interactive technologies, it is first necessary to look at what characterises military activities in the air, on land and at sea. The following discussion does this according to characteristics of the users, the technology and the environment. Although speech-based systems encompass both speech recognition and generation, the discussion focuses on ASR. There are two primary reasons for this. The development of technology to recognise human speech is inherently a more difficult problem than building a device to generate speech, and

J.M. Noyes (✉)
University of Bristol, Bristol, UK
e-mail: j.noyes@bristol.ac.uk

secondly, there are more issues associated with using ASR than speech synthesis for the users. It is relatively straightforward to listen to speech output whereas speech input often requires some training. This is discussed further in the next section.

13.2.1 Users

Characteristics of the military users have considerable implications for the use of speech-based systems for military applications. Differences in gender and dialect have been found to have a major impact on the performance of speech recognition systems [17, 54, 60]. However, user age is not an important factor; most speaker-independent recognisers are adequate for users in the age range of 15–70 years [67]. The section describes the importance of gender and dialect to military speech recognition systems.

Gender – Most North Atlantic Treaty Organisation (NATO) countries include women in their armed forces. In 2005, the percentage of female soldiers in the 24 countries of NATO ranged from 1% (Italy) to 20% (Latvia), with a mean percentage of 8.87% for all NATO countries [43]. In general, female NATO military personnel are not placed in combat roles, even in countries that make military service mandatory for women – their roles are largely confined to medical, administrative or logistical fields [6]. Several non-NATO countries (including China, Eritrea, Israel, Libya, Malaysia, North Korea, Peru and Taiwan) draft women into their armies. Other countries that do not normally draft women have done so during emergencies. For example, in the Second World War, Britain and the former Soviet Union conscripted women into their armed forces (CBC News). Thus, to be useful for military applications, speech recognition systems should accommodate female as well as male speakers [60].

Dialect – This is defined as a variety of a language used by people from a particular geographic area and includes a different pronunciation for phonemes [27]. Different dialects within the United States have been found to affect speech recogniser performance. Picone [51] found that speaker-independent ASR average word error rate was significantly higher for U.S. Deep Southern speech (19.9%) than for General American speech (5.1%). However, each year from 2003 through 2005, approximately 43.82% of U.S. military recruits were from southern states [25]. Differences in dialect also have been shown to affect performance in Japanese [27] and Arabic ASR systems [26]. Thus, successful military ASR systems must integrate models for dialect differences to be efficient.

User characteristics have implications when considering the implementation and use of speech-based systems. Current ASR technology, for example, is either speaker-dependent, speaker-adaptive or speaker-independent. Speaker-dependent technology “depends” on the user first training the recogniser in order to set up a set of templates relating to their individual speech patterns. This process, known as enrolment, builds a profile, which includes “a vocabulary and a regional language model with the phonetic analysis of the speaker’s voice”, that is, a speaker-specific

speech model [32]. Depending on the size of the vocabulary, enrolment can take approximately 1 h [47]. In contrast, speaker-adaptive and speaker-independent systems do not have this specific training phase, and users engage in their required task with the recogniser immediately. This would be particularly appropriate when the ASR technology needs to be accessed by the general public, for example, telephone-based enquiry systems [34]. The enrolment phase results in an individual's speech with their particular idiosyncrasies being stored; this will lead to better recognition performance. Given the characteristics of the military user groups, speaker-dependent systems requiring training, enrolment and dedicated users should not be a problem as ASR will be used by specifically trained individuals.

With respect to speech synthesis, the main difficulty for the user relates to the ephemeral nature of auditory information. When presented with a stream of spoken or artificially generated information, it is hard for humans to assimilate and process it in real time. As a result, they are likely to forget the content of the information. Single words or short phrases are easier to remember, and thus, more appropriate for listeners.

13.2.2 Technology

When considering the technology, the “speech” content of the interactions is important. Current speech recognition technology works better with constrained task domains where the vocabulary and grammar are limited [31, 66]; the ability of ASR to handle natural language dialogues with vocabularies of 30,000+ words still remains in the realms of science fiction. Williamson et al. demonstrated the benefits of a limited vocabulary and grammar in a UAV (Unmanned Aerial Vehicle) application where participants used ASR for control and data entry tasks. With a grammar of 160 words and short phrases, recognition rates of around 95% were achieved. Likewise, the Clarissa system described in Chapter 12 had a relatively small vocabulary of 260 words.

As a general rule, the military domain tends to lend itself to the use of limited vocabularies and grammar. For example, in ATC communications, pilots use constrained phraseology which would be suitable for ASR [64]. In a combat situation, lack of time coupled with the need to act quickly are paramount, so short exchanges of speech and the issuing of concise commands are needed rather than lengthy dialogues. However, Newman [41] found that ASR users are often unable to remember key words. A common assumption is that when using the technology, people will give their full attention to it, but this is not necessarily the case, especially in dynamic, rapidly changing operational environments [39]. User memory limitations and lack of attention will limit speech-based technology interactions.

A further feature of the military domain is the use of acronyms and abbreviations. Carr [7] generated a database of around 4,000 military acronyms and abbreviations, which he applied to some Commercial Off-The-Shelf (COTS) ASR software. In a command and control military application, speech recognition was being used for

data entry, and speech synthesis for issuing warnings and instructions. Hence, Carr found he needed to improve his COTS software by making it customised to his particular application.

A major problem for speech recognition technology is ASR detection of relevant vocal utterances. Note that in its broadest sense, the use of vocal utterances means that speakers need only generate sounds; as an example, coughs, grunts, sneezes would follow into this category. Since the technology works on pattern-matching principles, it does not matter whether it is speech or non-speech sounds which are being generated for recognition. One solution to overcome the problems of concatenation (i.e. linking together of speech) is isolated word recognition, where the utterances are made in separate units with pauses between them. For many applications, for example, text input of documents, this may pose a problem. However, in the military domain, this may not be an issue because in the cockpit for example, utterances of single command words would not be unusual. Another solution used in military applications is the push-to-talk control, where the user activates a control (such as a pushbutton) to signal that he or she is inputting speech into the ASR system. This approach is useful in a noisy environment or in an environment where the user must talk with others as well as to the ASR system.

In terms of speech synthesis, the issues relate to the quality of the output. Lai et al. [28, p. 321] stated that “for many years now, people have been predicting that the use of computer-generated speech would become widespread”. It is thought that one of the reasons why this has not been realised is that some synthetic speech sounds unnatural and is not particularly pleasant [14]. Synthesised speech is known for its monotonic, robot-sounding characteristics, which listeners tend not to like. The quality of artificial speech is an amalgam of its intelligibility (comprehensiveness), naturalness, presentation rate and emotional content [44]. It has a strong subjective component which relates to acceptability, that is, if people are content with the quality of the synthesised speech, they are more likely to find it acceptable [61]. Interestingly, intelligible speech may not sound natural and some natural sounding speech may not be intelligible. It has been shown that synthetic speech which is less intelligible takes longer to process [52] and is difficult to absorb, review or edit [57]. This may be due to memory limitations and the difficulties of carrying out two higher order cognitive processes (thinking and speaking) simultaneously. Shneiderman suggested this is why it may be easier to use a keyboard when thinking than a speech-based system. In military applications, the emphasis will need to be on intelligibility because of the safety-critical nature of the operating environment.

Speech input has been used in conjunction with other modalities in applications such as teleoperation (the remote operation of a device such as a robot or unmanned vehicle), which is often employed in military reconnaissance and search-and-rescue operations. Speech has been paired with touch, gesture and gaze for the purpose of creating an interface natural to the user that also frees his or her hands for other manual operations.

Touch – Perzanowski [50] performed a study to examine the extent to which speech recognition is used in conjunction with touch input (i.e. a touch screen) to guide a robot in a search task. In examining how often each modality was used to

give commands, Perzanowski [50] found that participants used speech input most often, making an average of five times as many utterances as touch inputs only or combined speech and touch inputs in each search task. Participants making the largest number of speech inputs tended to make the fewest touch inputs and vice versa. Most touch inputs were combined with verbal utterances, and when isolated touch inputs were made, they were intended to be corrections of previous touch inputs.

Gesture – Frigola et al. [16] suggested that human gesture is another means of controlling robots in tasks such as teleoperation. Gesture, which includes human body and arm pose, hand gesture and facial expression, can be static (not moving) or dynamic (where meaning is dependent on gesture motion trajectory). Generally, gesture interfaces are oriented to teleoperation tasks that leave the hands of the operator free [16] and can be provide a full range of interactions that complement and augment speech [69]. Yoshizaki et al. [68] described a service robot that used a vision-based interface to recognise gestures as well as obtain real-world information about objects mentioned in the user's speech. Weimer and Ganapathy [63] developed a teleoperation system using hand gesturing and voice input, using a data glove to track the hand. Hu et al. [21] noted that the gesture interface is advantageous because it is simple for the operator to use, can be used anywhere in the field of view of a camera and allows a wide variety of gestures since it is software based. At present, gesture control requires the presence of a camera system for image acquisition, and images can take a relatively long time for system computers to process, making it impractical for real-time applications.

Gaze – Mitsugami et al. [36] had participants use gaze and speech control to operate multiple robots, first giving a voice command, then using gaze to select a robot or specify a destination for robot travel. Mitsugami et al. [36] found that gaze and speech controls are advantageous because they free the user's hands for other tasks. However, gaze control requires the operator to wear a helmet-mounted display with an eye-tracking device to determine the precise direction of gaze. This presents a challenge in military environments, where eye trackers must be sufficiently rugged to work well in demanding applications such as ground vehicles, which may contain high levels of dust and vibration.

13.2.3 Environment

Having considered the characteristics of the users and the technology, a further factor is the environment in which the technology will be used and the context in which the speech interactions will take place. Military environments are often harsh and not amenable to the use of technology per se, let alone talking and listening to a machine with all the associated nuances. Military personnel are often subject to extremes of temperature and humidity, as units are employed in deserts, jungles and ice-capped regions with physical conditions which are dusty, sandy, wet, etc. Environmental factors such as these could affect the equipment and thus degrade

performance or even make the system inoperable. A summary of the key physical and cognitive factors are given here, but see [2] for a fuller review of ASR in adverse environments.

Noise – The first consideration is ambient noise, which will degrade ASR performance as it interferes with recognition of an individual's utterances. Noise in military environments might be continuous and repetitive (i.e. steady-state), such as vehicle engine noise. Battlefield noise might also be impulse noise, which is non-repetitive and of short (less than 1 s) duration, such as gunfire. Impulse noise usually has a high intensity, abrupt onset and rapid decay, and often rapidly changing spectral composition, which makes this type of background noise a difficult problem for speech recognition systems. Loud battlefield steady-state and impulse noise may also make it difficult for the soldier to hear synthesised speech, such as speech warnings.

Military aircraft, particularly helicopters, are known to be noisy: acoustic noise levels in helicopters are frequently in the region of 90–95 dBA, and the Chinook CH-47 records 115 dBA in cruise [9]. Likewise, the interiors of Armoured Fighting Vehicles (AFVs) are known to have high levels of steady-state ambient noise. However, Haas et al. [19] in their study of the use of ASR in a simulated tank environment reported a recognition rate of over 95%. More recently, Littlefield and Hashemi-Sakhtsari [32] conducted a study of ASR in background noise ranging from 15 to 50 dBA. They concluded that the recognition system performed well in the quiet, but that ambient and environmental noise was a major factor affecting recogniser performance. In their particular application, data entry in a C3I (Command, Control, Communications and Intelligence) environments, the most prominent type of background noise was human conversation. Naval ships are high-noise, industrial environments with a typical background level of 75–90 dBA [47]. The Naval Research Laboratories' SPeech In Noisy Environments (SPINE) Project is concerned with the development of algorithms and strategies for improving ASR performance in noisy environments [49, 58].

Vibration – All aircraft, vehicles and ships generate vibrations. The issue is whether this is sufficient to interfere with the physical production of human speech, which may render speech recognition technology inoperable. In a study of fixed wing aircraft, Leeks [29] concluded the level of vibration was sufficient to interfere with speech production. This was supported by Moore and Bond [37], who found that under low level, high speed flights, vibration levels would lead to changes in people's speech patterns.

Acceleration – Military aircraft often operate at high levels of acceleration (g-forces), and this will make it physically difficult for the crew to speak, and in turn for the ASR technology to recognise what is being said. A further problem relating to military aircraft is the need to wear oxygen masks; these will generate a further set of noises which the recogniser will pick up and attempt to process.

Stress – This is a nebulous topic in terms of definition and individual differences [12] because what is stressful to one person, may not be so to another. Murray et al. [38] attempted to address the problem of definition by referring to three orders of stressors: zero-order, which are unconscious, direct effects on the physical

production of speech (e.g. vibration), second-order, which result in conscious attempts to cope with stress (e.g. managing workload); third-order, a combination of the first two, for example, an emotional response, which is under the person's control to some degree. In terms of speech production, it would appear that some stressors, for example, the zero- and third-order ones, will have an effect over which the individual has little control.

Time pressure – This is likely to be a stressor prevalent in the military domain, that is, the need to think and act quickly within the context of knowing human lives are at risk. As a zero-order stressor, it will lead to changes in speech patterns and as a second-order stressor, this may also be the case.

Workload – Both high and low workload (under load) situations can prove stressful, although the former is more likely in the military domain. In high workload situations such as a cockpit, in an AFV or onboard a ship, voice changes are likely to occur as personnel struggle to keep up with the work. These changes in speech patterns are likely to lead to degradation in ASR performance, which in turn may lead to frustration and stress as the individuals attempt to make themselves understood. Working in a high workload situation is tiring, and it is known that fatigue alters the speech patterns. This may further exacerbate what is already a difficult situation.

In summary, there are many positive aspects of the military domain, for example, the user group and the technology, which support the use of speech-based systems. However, there are many characteristics of the military environment (noise, vibration, acceleration) and the context in which speech-based interactive systems would be employed (stress, time pressure, workload, fatigue), which do not lend themselves to the use of this technology. This is particularly the case with speech recognition rather than synthesis. However, despite the hostility of the operational environment, speech-based technologies have been extensively researched for military application. The next section considers some of these specific applications.

13.3 Applications

13.3.1 Air

The military cockpit can be a high workload environment with crew having to carry out many tasks at the same time as continuing to fly the aircraft. Thus, speech technology provides another input/output channel. Its primary advantage is that it is “hands and eyes free”, which allows the human to be “head up” and to carry out other manual tasks as well as scanning and monitoring visual displays at the same time. This is particularly the case in the single-pilot cockpit, which was one of the earliest scenarios considered for the introduction of speech technology [41]. In terms of disadvantages, these relate to the physical characteristics of the avionics environment with its high noise and vibration levels, and acceleration forces, as already discussed.

Consideration of the use of speech-based systems in avionics applications has a long history dating back to the 1970s when the first working speech recognisers were being developed. In the avionics domain, speech recognition and synthesis are referred to as Direct Voice Input (DVI) and Direct Voice Output (DVO), respectively. White et al. [65] carried out a comprehensive study to assess possible utilisations of ASR on the flight deck. They concluded that potential applications fell into the following five areas in descending order of usefulness: programming; interrogation; data entry; switch and mode selection; continuous/time-critical action control. Weinstein [64] suggested that in the future Multi-Role Fighter aircraft, voice would be used to control radio frequency settings, displays and gauges, and DVI would be used to enter reconnaissance reports. Today, the U.S. Air Force's Air Operations Center, from which combat air operations are planned and executed, utilises the very latest generation of Command and Control (C2) software [66]. It has been suggested that speech recognition technology has "the potential to improve the interface to these new computer applications by augmenting the conventional keyboard and mouse with spoken commands" (p. 10). Draper [13] tested the hypothesis that speech rather than manual input would be better for menu navigation and option selection in UAV applications. Indeed, they found that "speech input was significantly better than manual input in terms of task completion time, task accuracy, flight/navigation measures, and pilot ratings" (p. 109). Calhoun and Draper [5] noted that the advantage of speech input control was that single simple voice commands could substitute for several manual button selections, which enabled operators to maintain their hands on the controls and keep their heads directed toward critical visual displays.

When considering DVO, speech and non-speech warnings have been in operation on both military and civil aircraft flight decks. Auditory speech warnings have the advantage of being distinctive; this is especially the case when a female voice is used in a work situation which is predominantly male. Noyes et al. [44] conducted a review of speech warnings and found that more complex sound patterns are being generated as different temporal, frequency and amplitude characteristics are being combined. This suggests that looking to the future, speech and non-speech warnings will continue to proliferate.

Speech recognition and synthesis have been used in ATC training simulations [64]. The technology allows training to be automated as well as having the key advantage of providing rapid feedback to trainees. When considering ATC communications, online recognition of flight identification information would allow extremely fast access to data relating to a specific flight. In a military context, the gisting of ATC and flight information could be used to detect potential air space conflicts [64].

In summary, ASR has been considered in avionics applications for single actions, for example, switch and mode selection, and for data entry, for example, reconnaissance reports. Progress appears to have been limited to experimental rather than operational applications; this is primarily because the military cockpit is a safety-critical, harsh environment with a number of physical and cognitive characteristics which could impede recognition. In contrast, speech synthesis, for example, speech warnings, has been successfully used for a number of years.

13.3.2 Land

One of the key features of speech-based interactive systems in Army and Marine Corps applications relates to mobility. Although soldiers can be travelling in vehicles, they are often out in open terrain, and thus, their speech technology needs to accommodate this. It needs to be portable, and thus compact, and robust enough to cope with adverse environments. As well as war, soldiers are called in to help in peacetime activities, for example, in the case of natural disasters such as earthquake, flood, fire and tsunami. In these conditions, the operational environment is also likely to be harsh.

Today's armies are moving faster and further than previous generations, and the challenge is to make sure that Command and Control resources can keep pace with this (labelled as Command and Control On The Move – C2OTM). Situation awareness is of paramount importance, and there is a need to be able to make reports easily and efficiently, and for these to be translated and forwarded to the allies. As an example, a foot soldier acting as a forward observer could use ASR to enter a report that could be transmitted to command and control headquarters over a very low rate, jam-resistant channel [64]. In AFVs, commanders need to access battlefield information and then update it. Leggatt and Noyes [30] reported an experiment using ASR in an AFV simulator. They found no significant difference in the total reporting time taken when using the two input modes of ASR and keyboards; however, they did find that with ASR, the amount of communication in the AFV reduced. This is not surprising given that any speech was now directed towards the recogniser, but it could be disadvantageous to the situation if the commander and co-commander do not communicate with each other.

Two operational applications of speech-based technology in the Army include equipment repair and maintenance and Automated Language Translation (ALT) activities. When equipment repair and maintenance is needed in the field, ASR allows voice access to electronic manuals, which can be viewed via head-up displays attached to helmets. One such example is VID (Voice Interactive Display) developed by the U.S. Army [53]. The VID system was intended to reduce the bulk, weight and setup times of military diagnostic systems. It led to the development of the Soldier's On-system Repair Tool (SPORT) which is a hands-free interface between the operator and the repair tool. The VID comprises a microphone, a handheld display and SPORT. Diagnostic information can be gained by the technician by accessing an Interactive Electronic Technical Manual using voice commands. The intention is that this replaces the old system where data is collected manually on paper, and then input to the computer.

The U.S. Department of Defense (Army, Navy, Airforce and Marine Corps) has considered the use of speech-based technologies for ALT, where text is spoken to an ASR that translates it into another language for output via a recorded voice or speech synthesiser [3]. The Tongues System, which was part of the U.S. Army ACT-II programme, comprises a portable, speech-to-speech translation system [15]. This system comprises a recogniser, translator and synthesiser. It is used to support Army chaplains in their work when they need to communicate with local populations but do not know the native language. Frederking et al. [15] reported trials which had

been carried out in Croatia with the local population using English and Croatian. Their subjective impression was that conversations succeeded about half of the time. Due to this and the large amount of error correction required, participants found the system frustrating to use. This was thought to be due to the translation component, which was considered the weakest link, and not the recogniser or the synthesiser. However, Frederking et al. [15] concluded that the translation system was worthy of further development.

The U.S. Department of Defense explored ALT technology devices in a military environment through the Language and Speech Exploitation Resources (LASER) Advanced Concept Technology Demonstration [33]. The United States as well as other nations need processing capabilities in a wide range of human languages to support coalition and joint task forces in operations involving force protection, medical assistance, peacekeeping and humanitarian efforts. The purpose of ALT is to enhance military unit deployment worldwide where there are no linguists present to support mission requirements.

Marshall [33] described several ALT systems, including some speech-to-speech systems (translators initiated by a voice speaking the source language, with a resulting target language translation produced through an audio device such as a loudspeaker). The author described several commercially available systems designed for the military domain, including two that supported more than 30 languages such as Korean, Iraqi and Serb-Croatian. Several of these systems reached prototype demonstration in an evaluation environment in multi-national operations. In observing these prototype trials, Marshall noted that the human element of using ALT technology possesses unique challenges, especially in face-to-face situations. These challenges included social discomfort, where users of the prototype systems became uncomfortable using the technology (i.e. they did not maintain required eye contact with the persons with whom they were “talking”). Another challenge included socio-cultural differences related to body language and gesture, which translators do not perform but are still needed for effective human-to-human communication. Marshall noted that the biggest challenge is for users to accept the need for human language translation technology, because of the shortage of humans to perform the job. She recommended that future ALT research explores designing more effective human training for prospective ALT device users, particularly in face-to-face interactions using speech-to-speech devices.

For the next generation of combat vehicles, the U.S. Army and Marines are considering lightly armoured robotic vehicles controlled by soldiers operating from within moving command vehicles well back from the forward edge of the battle area [40]. These robotic vehicles will help the soldier maintain situation awareness of battlefield conditions. As part of this effort, Neely et al. developed a prototype virtual environment to allow the soldier to interact directly with visual representations of the robotic entities while seated in a vehicle, without the use of a keyboard. They developed a visual, speech and gesture interface that could assist the soldier perform tasks such as creating target zones, phase lines or markers, assigning groups to locations, zooming to an object or location and obtaining weather information. At

present, Neely et al. have not been able to conduct a proper user study with soldiers, but reported anecdotal observations that civilian users could master the speech and gesture interface with few problems. These researchers suggested that a proper user study be performed in the future, using experienced soldiers as participants.

In summary, mobility is a key theme running through the Army applications. Unlike avionics applications, Army soldiers generally have more freedom to move around their environment, either mounted on a ground vehicle or helicopter, or dismounted (on foot). This has implications for the design of speech-based applications. As previously mentioned, Army ASR systems must be robust in battlefield environments that contain high levels of noise and vibration.

13.3.3 Sea

When considering military applications, the maritime environment shares many of the features of the air and land. Ship engine rooms can be noisy, and subject to significant levels of vibration. Just as aircraft experience turbulence and Army vehicles experience off-road conditions, naval vessels can encounter stormy seas. Accordingly, speech-based systems will need to be robust to cope with the physical and cognitive aspects of this environment.

Specific projects include the development of the U.S. Naval Voice Interactive Device (NVID), which is equivalent to the U.S. Army's VID [47]. The aim of this project was "to create a lightweight, portable computing device that uses ASR to enter shipboard environmental survey data into a computer database" [47]. Onboard naval ships, health professionals check the environmental conditions daily to ensure health and safety standards are being met. These checks include temperature and water testing, pest control surveys, routine sanitation and habitability inspections. Data collection conditions often comprise cramped, tight spaces requiring two hands for climbing and stability; since inspectors have to be mobile and wear safety clothing (gloves, glasses, hard hats, hearing protectors), keyboard entry is not considered feasible. Speech-based technology captures, stores, processes and forwards data to a server where it can be easily retrieved; hence, it allows remote access to databases via wireless technology.

One speech-based application has been the SONAR supervisor for command and control [64]. The SONAR supervisor needs to monitor displays, send messages and direct resources, whilst moving around the control centre. Hence, there is a requirement for a portable device that facilitates mobility. It could be concluded that this is a similar concept to the soldier's computer, where an individual is given the equivalent of a portable computer which is accessed via ASR.

The Navy has also used speech-based technology for training purposes. For example, speech technology was used in the training of SONAR supervisors as well as a combat team tactical training application [64]. In a combat scenario, ASR, typing and trackballs were used for individuals to communicate with each other and with the computer system. Like the ATC system, one of the key advantages of this

type of training simulation is that it allows rapid feedback to be given. Training simulations are also useful for trying out new technologies in realistic settings where it does not matter too much, if performance is poor.

In summary, there are parallels between the applications in all branches of the military in terms of personnel working in restricted, hostile environments. Particularly evident are the effects of using ASR in noisy conditions. Moreover, human-machine speech communication in Air Force command and control posts is similar to the Army and Navy. It is apparent that ASR technology transfer occurred between different branches of the military which share common objectives, such as when ASR has been considered for UAV launch, recovery weapon status and maintenance information [64].

13.4 General Discussion

This chapter has attempted to provide a state-of-the-art review of the application of speech-based interactive technology in the military domain. It is quite likely, however, that some developments are not in the open literature and were not discussed due to the nature of the military domain, and defence and security implications. Further, there is often not enough incentive for the industrial sector to publish findings in the same way as academia. With this in mind, the applications will be considered in terms of users, technology and environment.

13.4.1 Users

Military personnel work in an environment where they may have little choice about whether or not they use speech-based systems. Their views, however, concerning the acceptability of the technology should be taken into account during the experimental and testing phases. For example, Navy personnel have expressed dislike at having to train the speaker-dependent recogniser [64]. However, favourable attitudes were expressed towards the NVID system as ASR was seen to save time and improve the quality of the reports [47]. A possible consideration concerns the embarrassment of talking to a machine in the workplace. This is something we have found in work at Bristol on office applications of ASR. The military sector has a rigid hierarchy, and some senior people may find being seen talking to a recogniser demeaning.

A further issue relates to error correction. In the Tongues system, Frederking et al. [15] found that users were frustrated with the ASR performance and the error correction procedures. This could be a significant problem because even with 98% accuracy, there is still a need to spend some time on error detection and correction [20]. In both the studies of [15, 20], ASR was often found to be unpredictable and would often misrecognise words or phrases with no apparent underlying logic. This is in contrast to using a keyboard where errors resulting from mis-keying are very straightforward for users to comprehend.

13.4.2 Technology

There is a need for robust speech-based technology in military applications [45]. Robustness has been defined as the “ability to handle the presence of background noise and to cope with the distortion by the frequency characteristic of the transmission channel” [48]. Background noise is certainly a major concern for both ASR and speech synthesis technologies in military applications. This is exacerbated by the safety-critical aspects of the operating environment, and the need for the technology to achieve high performance rates to facilitate accurate user interactions.

Over the past several years a great deal of research has been conducted in the area of noise-robust speech recognition. The goal of errorless speech recognition has been unattainable in part because of degradation in recognition accuracy when there is any kind of difference between the conditions in which the speech system has been trained and the conditions in which it is finally used [24]. Junqua listed several causes of ASR error, including those that originate from the speaker or from the acoustic environment. Speaker-based errors include user speaking rate (which can change with stress), differences in user dialect and pronunciation, differences between native versus non-native speakers, unexpected speaker behaviour and task-induced speaker variability. Acoustic environmental causes include reverberation and environmental noise. Problems can also be caused by variable transducers (use of a microphone other than the one used for training) or transmission of speech over noisy communications systems such as radio, intercom, telephone or wireless connections.

Several approaches have been taken to make speech recognition more robust, especially in high-noise (e.g. military) environments. One approach is the use of microphone technologies. Another approach uses algorithms to characterise and estimate frequently changing acoustic conditions, and to identify and compensate for major sources of error in recognition performance. Algorithm-based approaches, which include feature-space compensation and model-space compensation, are beyond the scope of this book but are further described in [11, 22].

Microphone-based technologies have been used to mitigate undesirable environmental acoustic features. Microphone-based solutions focus on ASR signal acquisition and work to reduce signal contamination at the source to improve ASR performance. Microphone-based applications include microphone arrays and noise-cancelling microphones such as those described by Strand et al. [59].

Microphone arrays are groups of microphones placed to capture speech from the direction of the speaker while attenuating background noise from other directions. However, because the arrays are usually mounted on a surface facing the speaker, they can be impractical for mobile military applications.

With close-talking noise-cancelling microphones, which are popular for ASR applications, a boom microphone is usually mounted on a headset or helmet and its performance depends on a correct position relative to the user’s mouth. Close-talking microphones can use active or passive noise reduction. Sawhney and Schmandt [56] noted that passive microphones that use good directional microphone design can provide some form of noise cancellation.

Active noise cancellation is a process in which a microphone is used to pick up sound waves from a noise source in the environment, generate an identical waveform 180° out-of-phase, and direct it toward the noise source. When the two sound waves combine, they cancel each other out, thus reducing the noise. Oviatt [46] found that noise-cancelling microphones enhanced ASR performance in hand-held mobile systems in a noisy environment. In military applications, a key benefit of speech as an input/output device is the mobility it affords.

13.4.3 Environment

A key feature of military as opposed to civil applications concerns the adverse operational environment, which places great demands on the technology, and may help explain why user acceptance has been relatively slow. Although “large vocabulary, speaker-dependent, continuous speech recognisers are commercially available” [32, p. 1], the need to attain good performance rates in a noisy environment with high levels of vibration, g-forces and cognitive stress does mean that it is high risk to implement ASR in this domain. In contrast, there are not so many risks attached to the use of speech synthesis and this is evident from its uptake and the military applications for which it is currently being used.

In summary, there are many advantages of using speech-based technologies in military applications. These relate primarily to the “transparent” nature of speech, which allows users to have their “eyes and hands free” for other tasks and not to be burdened carrying bulky computer equipment. Since ASR allows data to be input straight into the system, thus reducing the cycle of taking down information in hard-copy form for later keyboard entry, speech should reduce the resources needed for reporting and communication. Disadvantages stem from the immaturity of current ASR technology to cope with the demands of an extremely tough environment, where misrecognitions could lead to loss of life.

13.5 Future Research

13.5.1 Challenges

Speech-based interactive systems continue to advance; however, in military applications, there are still a number of significant challenges. These challenges relate to speech system performance, robustness, mobility, multimodality, error correction, integration and reversion.

Performance – High levels of performance of speech-based technology are required in military applications and often recognition rates are not sufficient for ASR and speech synthesis, where intelligibility can be an issue. Recognition rates as a measure of performance are now being extended to include “accuracy and robustness, real time, memory foot-print, latency, barge-in detection and rejection of out of

vocabulary events” [18]. Taking this broader view can only benefit the development of this technology.

Robustness – Speech-based technology needs to be resilient in battlefield environments with high levels of steady-state and impulse background noise, under high levels of soldier workload and stress. In addition to noise-resilience, future challenges include rapid system adaptation to the accents of non-native speakers (especially useful for machine translation devices), recognition in environments with other speakers (useful in crew station systems with several talkers), multiple microphone systems [24] and adaptation to rapidly changing noise (such as impulse noise coming from weapons bursts). Researchers are developing speech databases useful for future research, such as the Speech under Simulated and Actual Stress (SUSAS) database [4]. Other research directions in robust speech recognition include robust and selective system learning, dynamic system and design strategies, intelligent, flexible and portable dialog systems; and task and language adaptation. These elements are described in more detail in Junqua [24].

Mobility – Military applications are often suitable for portable speech-based technology. Further researcher should focus on developing robust hand-held and mobile ASR systems. In addition, the development of low weight wearable computers would suit this domain [35].

Stress – Robust speech recognition for soldiers under stressful battlefield conditions is an important requirement. Military operations are often conducted under stressful conditions caused by fear, confusion due to conflicting information, psychological tension, pain, high workload and other conditions that can be encountered in a typical battlefield. Algorithms such as dynamic time warping (a method for measuring similarity between two sequences which may vary in time or speed) are being developed and tested under relevant conditions such as high g-forces [8], while other techniques such as analysis of the airflow structure of speech production under adverse stressful conditions are being used to identify speech under stress [69]. Anderson et al. [1] stated the need for ASR system testing using planned simulations with military personnel using a wide variety of speech technology and addressing factors that cause battlefield stress.

Multimodality – Challenges exist for the integration of speech and gesture, gaze or touch input in multimodal interfaces [70]. The largest challenge is interpreting or understanding the different modalities in the proper context. For example, when someone says, “What about that one” while pointing to an object in the environment, the system would be able to interpret correctly the indirect referencing in the speech signal using information in the other (in this case, the visual) channel. Zue [70] stated that for a multimodal interface to be effective, it is important to develop a unifying linguistic formalism that can describe multimodal interactions (e.g. “move this one over here”) as well as successful integration and delivery strategies.

Error Correction – A need exists to equip speech recognisers with the ability to learn and correct errors [10]. Error correction in current speech recognition systems is cumbersome and time-consuming, which reduces the user’s incentive to adopt and use speech recognition, especially in challenging environments in which military systems are used. Research should be conducted to define advanced techniques that

could automatically learn and correct errors. Deng and Huang [10] suggested that two main types of errors subject to correction should include new words outside the recogniser's lexicon and errors due to user pronunciations that are different from those in the recogniser's lexicon.

Integration – It does not always work just to link a recogniser, a translator and a synthesiser together (as demonstrated in the Tongues system [15]). Integration of the technology with regard to the various “internal” components as well as with other, external systems is needed. The latter was found in the development of the soldier's computer [64]. Hence, it is necessary to ensure integration and compatibility of speech-based tools with underlying computer operating systems.

Reversion (i.e. backup systems) – Due to the demands on the performance of the technology and the safety-critical aspects of the military domain, there is a need to design reversion techniques into applications.

13.5.2 Recommendations for Future Research

A well-designed speech interface is transparent, well organised and easily comprehensible to the user. In order to achieve this, human factors engineering should be incorporated early into the design process of the military speech recognition application in order to ensure that it is designed with the soldier in mind. As suggested by Junqua [24], steps that should be taken in developing a prototype speech interface include clearly defining the application domain, defining who the users will be and maintaining focus on these users.

It is also important that the human factors engineer be involved after the prototype is designed, to assist in collecting data for incremental design adaptation and to have a hand in iteratively testing and improving system design. At this point in the process, the system should be tested with the intended military users in a realistic military environment, as recommended by Anderson et al. [1].

At all stages of the process, it is important to ensure that all aspects of usability evolve together in an integrated design. Junqua [24] warned that the human factors design process should not underestimate the importance of providing of user feedback. Junqua [24] noted that designing for user and system error is essential, as is allowing the user to understand the semantics of the domain.

An approach much like that suggested by Junqua [24] was described in a case study of the design process of the NVID, the U.S. Navy shipboard speech recognition system described earlier in this chapter. Rodger et al. [55] described the need for a voice interactive technology to improve medical readiness aboard U.S. Navy ships. The goal of this project was to design a lightweight, wearable computing device with voice interactive capability. The authors noted that a focus group was surveyed early in the process about methods they used in reporting health and environmental surveillance inspections. Features important to the focus group were used to develop a device with automated user prompts, enhanced data analysis and presentation and dissemination tools that supported preventive medicine. A human factors analysis

of the prototype found that it enabled quick, efficient and accurate environmental surveillance. In all, the integration of human factors engineering produced a speech interface device that reduced the time needed to complete inspections and supported local reporting requirements and enhanced command-level intelligence.

In conclusion, the challenges facing the use of speech-based interactive systems in military applications are considerable. With increasing globalisation, and developments in the Internet, bandwidth, wireless and microphone technology, the number of speech system applications is growing. However, full implementation of speech-based technology in the military domain is probably still many years away because physical elements (e.g. noise and vibration) and cognitive characteristics (e.g. user stress, time pressure and workload) generated by the battlefield environment provide a challenge to speech recognition technology.

References

1. Anderson, T., Pigeon, S., Swail, C., Geoffrois, E., Bruckner, C. (2004). Implications of multilingual interoperability of speech technology for military use. NATO Research and Technology Organization, Report RTO-TR-IST-011, AC/323(IST-011)TP/26.
2. Baber, C., Noyes, J. M. (1996). Automatic speech recognition in adverse environments. *Hum. Factors*, 38, 142–155.
3. Benincasa, D. S., Smith, S. E., Smith, M. J. (2004). Impacting the war on terrorism with language translation. In: Proc. IEEE Aerospace Conf., Big Sky, MT, USA, 3283–3288.
4. Bolia, R. S., Slyh, R. E. (2003). Perception of stress and speaking style for selected elements of the SUSAS database. *Speech Commun.*, 40, 493–501.
5. Calhoun, G., Draper, M. H. (2006). Multi-sensory interfaces for remotely operated vehicles. In: Cooke, N. J., Pringle, H. L., Pedersen, H. K., Connor, O. (eds) *Advances in Human Performance and Cognitive Engineering Research*, vol. 7: *Human Factors of Remotely Operated Vehicles*, 149–163.
6. Canadian Broadcasting Corporation (CBC) PInews (2006). Women in the military – International. In: CBC News Online, May 30, 2006. Available online <http://www.cbc.ca/news/background/military-international/>
7. Carr, O. (2002). Interfacing COTS speech recognition and synthesis software to a Lotus notes military command and control database. Defence Science and Technology Organisation, Information Sciences Laboratory, Edinburgh, Australia. Research Report AR-012-484. Available online, May 2006: <http://www.dsto.defence.gov.au/corporate/reports/DSTO-TR-1358.pdf>.
8. Chengguo, L., Jiqing, H., Wang, C. (2005). Stressful speech recognition method based on difference subspace integrated with dynamic time warping. *Acta Acoust.*, 30 (3), 229–234.
9. Cresswell, Starr, A. F. (1993). Is control by voice the right answer for the avionics environment? In: Baber, C., Noyes, J. M. (eds) *Interactive Speech Technology: Human Factors Issues in the Application of Speech Input/Output to Computers*. Taylor & Francis, London, 85–97.
10. Deng, L., Huang, X. (2004). Challenges in adopting speech recognition. *Commun. ACM*, 47 (1), 69–73.
11. Deng, L., O’Shaughnessy, D. (2003). *Speech Processing – A Dynamic and Optimization-Oriented Approach*. Marcel Dekker, NY.
12. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D. (1998). SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NEST 1998 speaker recognition evaluation. In: Proc. IEEE Int. Conf. on Spoken Language Processing, ICSLP ‘98 Sydney, Australia, 608–611.

13. Draper, M., Calhoun, G., Ruff, H., Williamson, D., Barry, T. (2003). Manual versus speech input for unmanned aerial vehicle control station operations. In: Proc. 47th Annual Meeting of the Human Factors and Ergonomics Society, Denver, CO, USA, 109–113.
14. Francis, A. L., Nusbaum, H. C. (1999). Evaluating the quality of synthetic speech. In: Gardner-Bonneau, D. (ed) *Human Factors and Voice Interactive Systems*. Kluwer, Norwell, MA, 63–97.
15. Frederking, R. E., Black, A. W., Brown, R. D., Moody, J., Steinbrecher, E. (2002). Field testing the Tongues speech-to-speech machine translation system, 160–164. Available online, May 2006: <http://www.cs.cmu.edu/~awb/papers/lrec2002/tongues-eval.pdf>.
16. Frigola, M., Fernandez, J., Aranda, J. (2003). Visual human machine interface by gestures. In: Proc. IEEE Int. Conf. on Robotics & Automation, Taipei, Taiwan, 386–391.
17. Fuegen, C., Rogina, I. (2000). Integrating dynamic speech modalities into context decision trees. In: Proc. IEEE Int. Conf. of Acoustic Speech Signal Processing, Istanbul, Turkey. ICASSP 2000, vol. 3, 1277–1280.
18. Goffin, V., Allauzen, C., Bocchieri, E., Hakkani-Tür, D., Ljolje, A., Parthasarathy, S., Rahim, M., Riccardi, G., Saraclar, M. (2005). The AT&T Watson speech recogniser. In: Proc. IEEE Int. Conf. on Spoken Language Processing, ICLSP 2005, Philadelphia, PA, I-1033–I-1036.
19. Haas, E., Shankle, R., Murray, H., Travers, D., Wheeler, T. (2000). Issues relating to automatic speech recognition and spatial auditory displays in high noise, stressful tank environments. In: Proc. IEA 2000/HFES 2000 Congress. Human Factors and Ergonomics Society, Santa Monica, CA, vol. 3, 754–757.
20. Halverson, C. A., Horn, D. B., Karat, C. M., Karat, J. (1999). The beauty of errors: Patterns of error correction in desktop speech systems. In: Sasse, M. A., Johnson, C. (eds) *Proc. Human-Computer Interaction – INTERACT '99*. IOS Press, Amsterdam.
21. Hu, C., Meng, M. Q., Liu, P. X., Wang, X. (2003). Visual gesture recognition for human-machine interface of robot teleoperation. In: Proc. 2003 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Las Vegas, NV, USA, 1560–1565.
22. Huang, S. D., Acero, A., Hon, H. (2001). *Spoken Language Processing – A Guide to Theory, Algorithms, and System Development*. Prentice Hall, NY.
23. Jokinen, K. (2006). Constructive dialogue management for speech-based interaction systems. In: Proc. *Intelligent User Interfaces'06*, Sydney, Australia. ACM Press, New York, NY.
24. Junqua, J. (2000). *Robust Speech Recognition in Embedded Systems and PC Applications*. Kluwer, Norwell, MA.
25. Kane, T. (2006). *Who are the recruits? The demographics characteristics of U.S. Military enlistment, 2003–2005*. The Heritage Foundation, Washington, DC.
26. Kirchoff, K., Vegyi, D. (2004). Cross-dialectal acoustic data sharing for Arabic speech recognition. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP 2004, vol. 1, 765–768.
27. Kudo, I., Nakama, T., Watanabe, T., Kameyama, R. (1996). Data collection of Japanese dialects and its influence into speech recognition. In: Proc. 4th Int. Conf. on Spoken Language Processing (ICSLP), vol. 4, 2021–2024.
28. Lai, J., Wood, D., Considine, M. (2000). The effect of task conditions on the comprehensibility of synthetic speech. *CHI Lett.*, 2, 321–328.
29. Leeks, C. (1986). Operation of a speech recogniser under whole body vibration (Technical Memorandum FDS(F) 634). RAE, Farnborough, UK.
30. Leggatt, A. P., Noyes, J. M. (2004). A holistic approach to the introduction of automatic speech recognition technology in ground combat vehicles. *Mil. Psychol.*, 16, 81–97.
31. Lippmann, R. (1997). Speech recognition by machines and humans. *Speech Commun.*, 22, 1–15.
32. Littlefield, J., Hashemi-Sakhtsari, A. (2002). The effects of background noise on the performance of an Automatic Speech Recogniser. Defence Science and Technology Organisation, Information Sciences Laboratory, Edinburgh, Australia. Research Report AR-012-500.

- Available online, May 2006: <http://www.dsto.defence.gov.au/corporate/reports/DSTO-RR-0248.pdf>.
33. Marshall, S. L. (2005). Concept of operations (CONOPS) for foreign language and speech translation technologies in a coalition military environment. Unpublished Master's Thesis, Naval Postgraduate School, Monterey, CA.
 34. McCarty, D. (2000). Building the business case for speech in call centers: Balancing customer experience and cost. In: Proc. SpeechTEK, New York, 15–26.
 35. Minker, W., Bühler, D., Dybkjaer, L. (2005). Spoken Multimodal Human–Computer Dialogue in Mobile Environments. Springer, Dordrecht.
 36. Mitsugami, I., Ukita, N., Kidode, M. (2005). Robot navigation by eye pointing. In: Proc. 4th Int. Conf. on Entertainment Computing (ICEC), Sanda, Japan, 256–267.
 37. Moore, T. J., Bond, Z. S. (1987). Acoustic-phonetic changes in speech due to environmental stressors: Implications for speech recognition in the cockpit. In: Proc. 4th Int. Symp. on Aviation Psychology, Aviation Psychology Laboratory, Columbus, OH.
 38. Murray, I. R., Baber, C., South, A. (1996). Towards a definition and working model of stress and its effects on speech. *Speech Commun.*, 20, 3–12.
 39. Myers, B., Hudson, S. E., Pausch, R. (2000). Past, present, and future of user interface software tools. *ACM Trans. Comput. Hum. Interact.*, 7, 3–28.
 40. Neely, H. E., Belvin, R. S., Fox, J. R., Daily, J. M. (2004). Multimodal interaction techniques for situational awareness and command of robotic combat entities. In: Proc. IEEE Aerospace Conf., Big Sky, MT, USA, 3297–3305.
 41. Newman, D. (2000). Speech interfaces that require less human memory. In: Basson, S. (ed) AVIOS Proc. Speech Technology & Applications Expo, San Jose, CA, 65–69.
 42. North, R. A., Bergeron, H. (1984). Systems concept for speech technology application in general aviation. In: Proc. 6th Digital Avionics Systems Conf. (A85-17801 06-01). American Institute of Aeronautics and Astronautics, New York, AIAA-84-2639, 184–189.
 43. North Atlantic Treaty Organisation (NATO) Committee for Women in the NATO Forces (2006). Personnel comparison in deployments 2006. Available online, December 2006: http://www.nato.int/issues/women_nato/index.html
 44. Noyes, J. M., Hellier, E., Edworthy, J. (2006). Speech warnings: A review. *Theor. Issues Ergonomics Sci.*, 7 (6), 551–571.
 45. Oberteuffer, J. (1994). Commercial applications of speech interface technology: An industry at the threshold. In: Roe, R., Wilpon, J. (eds) Voice Communication Between Humans and Machines. National Academy Press, Washington DC, 347–356.
 46. Oviatt, S. L. (2000). Multimodal system processing in mobile environments. *CHI Lett.*, 2 (2), 21–30.
 47. Paper, D. J., Rodger, J. A., Simon, S. J. (2004). Voice says it all in the Navy. *Commun. ACM*, 47, 97–101.
 48. Pearce, D., Hirsch, H. G. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. 6th Int. Conf. on Spoken Language Processing, ICSLP 2000, Beijing, China.
 49. Pellom, B., Hacıoglu, K. (2003). Recent improvements in the CU sonic ASR system for noisy speech: The SPINE task. In: IEEE Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Hong Kong, China, ICASSP 2003, I-4–I-7.
 50. Perzanowski, D., Brock, D., Blisard, S., Adams, W., Bugajska, M., Schultz, A. (2003). Finding the FOO: A pilot study for a multimodal interface. In: Proc. 2003 IEEE Conf. on Systems, Man and Cybernetics, vol. 4, 3218–3223.
 51. Picone, J. (1990). The demographics of speaker independent digit recognition. In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP 1990, vol. 1, 105–108.

52. Ralston, J. V., Pisoni, D. B., Lively, S. E., Greene, B. G., Mullennix, J. W. (1991). Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Hum. Factors*, 33, 471–491.
53. Rodger, J. A., Pendharkar, P. C., Paper, D. C., Trank, T. V. (2001). Military applications of natural language processing and software. In: *Proc. 7th Americas Conf. on Information Systems*, Boston, MA, USA, 1188–1193.
54. Rodger, J. A., Pendharkar, P. C. (2004). A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *Int. J. Hum. Comput. Studies*, 60 (5–6), 529–544.
55. Rodger, J. A., Trank, T. V., Pendharkar, P. C. (2002). Military applications of natural language processing and software. *Ann. Cases Inf. Technol.*, 5, 12–28.
56. Sawhney, N., Schmandt, C. (2000). Nomadic Radio: Speech and audio interaction for contextual messaging in nomadic environments. *ACM Trans. Comput. Hum. Interact.*, 7 (3), 353–383.
57. Shneiderman, B. (2000). The limits of speech recognition. *Commun. ACM*, 43, 63–65.
58. Singh, R., Seltzer, M. L., Raj, B., Stern, R. M. (2001). Speech in noisy environments: Robust automatic segmentation, feature extraction, and hypothesis combination. In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2001*, Salt Lake City, UT, vol. 1, 273–276.
59. Strand, O. M., Holter, T., Egeberg, A., Stensby, S. (2003). On the feasibility of ASR in extreme noise using the PARAT earplug communication terminal. *IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, Virgin Islands, 315–320.
60. Tashakkori, R., Bowers, C. (2003). Similarity analysis of voice signals using wavelets with dynamic time warping. *Proc. SPIE*, 5102, 168–177.
61. Viswanathan, M., Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. *Comput. Speech Lang.*, 19, 55–83.
62. Wagner, M. (1997). Speaker characteristics in speech and speaker recognition. In: *Proc. 1997 IEEE TENCON Conf.*, Brisbane, Australia, part 2, 626.
63. Weimer, C., Ganapathy, S. K. (1989). A synthetic visual environment with hand gesturing and voice input. In: *HCI International 89: 3rd International Conference on Human-Computer Interaction* September 18–22, 1989, Boston, MA, USA.
64. Weinstein, C. J. (1995). Military and government applications of human-machine communication by voice. *Proc. Natl Acad. Sci. USA*, 92, 10011–10016. (Reprint of *Military and government applications of human-machine communications by voice*. In: Roe, R., Wilpon, J. (eds) *Voice communication between humans and machines*. National Academy Press, Washington DC, 357–370).
65. White, R. W., Parks, D. L., Smith, W. D. (1984). Potential flight applications for voice recognition and synthesis systems. In: *Proc. 6th AIAA/IEEE Digital Avionics System Conf.*, 84-2661-CP.
66. Williamson, D. T., Draper, M. H., Calhoun, G. L., Barry, T. P. (2005). Commercial speech recognition technology in the military domain: Results of two recent research efforts. *Int. J. Speech Technol.*, 8, 9–16.
67. Wilpon, J. G., Jacobsen, C. N. (1996). A study of speech recognition for children and the elderly. In: *Proc. IEEE Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Atlanta, GA, USA, vol. 1, 349–352.
68. Yoshizaki, M., Kuno, Y., Nakamura, A. (2002). Human-robot interface based on the mutual assistance between speech and vision. In: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Swiss Federal Institute of Technology, Lausanne, Switzerland, 1308–1313.
69. Zhou, G., Hansen, J. H. L., Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Trans. Speech Audio Process.*, 9 (3), 201–216.
70. Zue, V. (2004). Eighty challenges facing speech input/output technologies. In: *Proc. from Sound to Sense: 50+ Years of Discovery in Speech Communication*, MIT, Boston, MA, USA, B179–B195.

Chapter 14

Accessibility and Design for All Solutions Through Speech Technology

Diamantino Freitas

14.1 Introduction

The advent of computer-based speech-processing systems like speech synthesizers (SS) and speech recognisers (SR) has brought mankind a promising way of realising the fundamental need for spoken communication by enabling automatic speech-mediated communication. The acoustic medium and speech can be used to implement a high potential communication channel in an alternative, or augmentative, way to improve accessibility to communication for persons with special needs. It takes speech-processing both for speech input and output into consideration, and is presently a well-defined and visible trend in communication technology. Moreover, it is assumed that the solutions for communication difficulties of disabled persons can also bring advantages for non-disabled persons by providing redundancy, and therefore higher comfort, in the use of the communication systems.

This chapter provides a brief presentation of the problems as well as current and future solutions for the specific situations of the visually disabled, the mobility impaired, the speech impaired, the hearing impaired and the elderly. Problems found generally in physical locations, principally in built public sites, as well as in transportation, require special attention, as they require that one can use navigation systems and have access to the information. The chapter discusses accessibility issues also in the telecommunications environment with reference to the main projects in the area. In order to examine what main benefits can be extracted from the use of speech technology today and in the near future, the final chapter discusses issues concerning access to complex documents in e-learning environments, and in instructional games and ebooks.

D. Freitas (✉)

Speech Processing Laboratory, University of Porto, Faculty of Engineering, 4250-066 Porto, Portugal

e-mail: dfreitas@fe.up.pt

14.1.1 Text and Speech Media

It may be surprising to introduce Design-for-All (DFA) speech-processing solutions by discussing text features. However, text is an essential medium in all communication, due to its fundamental role in providing a representation for the communication contents in close relationship with spoken natural language. It is also a light and inherently compressed medium in regard to the technical code used, and thus one of the preferred media for the computer representation of complex information. Text representation has evolved a lot and acquired new dimensions when mark-up languages were created to merge tags and structure information (i.e. document type description – DTD) into the text stream that made it a more complete communication medium. If we consider these characteristics, together with mark-up technology, then it should be understandable that we dedicate a few moments to it in our discussion of speech solutions.

Speech and natural language convey several layers of meaningful features and contents. In speech, besides the localised segmental and less localised supra-segmental features that can be heard and readily detected by technical means, there are other layers of information which are strongly supported by underlying language structures. These are, for example, the pragmatic (linguistic), the intentional (paralinguistic) and the emotional (non-linguistic) contents of speech. They cannot be disregarded in human communication, but they are more difficult to recognise. Interaction through speech thus places a big stress on the naturalness of the medium and on its capability to convey the necessary layers of information. If the channel is not capable of conveying the richness of communication, the well-known result is discomfort with the loss of communicative accuracy and effectiveness.

14.1.2 Multimedia

When communicating, humans usually combine several media as much as possible, and in this way they enrich and bring plasticity to the meaning. This is the reason why multimedia speech communication naturally deserves to be developed: speech and visual communication are tied together and combined in most peoples' daily life communication.

However, for disabled or elderly users, and for users with special needs in general, the multimedia communication approach appears to be simultaneously a challenge and a promise. The challenge deals with the access to one or more of the media possibilities, and there is often a great, or even unsurpassable, barrier of sensorial or physiological origin. The promise comes from the possibility for conversing and combining the various media in order to take into consideration the existing sensorial or cognitive capacities of the users and to help them to augment their cognition, or just to increase the comfort and accuracy of information transactions. Take, for instance, the case of the established text-to-Braille and the text-to-speech medium conversions. An enriched text representation can be better conveyed through speech instead of Braille due to the essential plasticity of speech.

However, in this case accuracy is, in principle, not better because of the cognitive issues in speech perception, comprehension and memorisation.

In this chapter, the simple concept of medium conversion and interaction will be explored like a framework for analysing and describing a set of developments that are available today for technology applications in speech-based interactive systems. The perspective will be that of Design-for-All, i.e. accessibility to digital information in daily life for several disability groups. Such groups include the blind (applications concerning mainly system output but also input), the mobility impaired (applications concerning mainly input, with audio-visual recognition), the speech impaired (applications concerning mainly output in telecommunication), the hearing impaired (applications concerning mainly output in text format as well as facial speech synthesis and sign language) and the elderly. The applications concern human-machine interfaces, navigation systems for built environments and for transportation, learning applications (which require access to complex documents often with mathematical content) and other applications such as instructional games and ebooks. The COST 219 action (<http://www.cost219ter.org>) has been one of the few pan-European research and development actions with relevant worldwide associates to devote attention to the topic of accessibility in telecommunications and information systems.

14.2 Applications for Blind or Partially Sighted Persons

The common computer desktop design metaphor, a Graphical User Interface (GUI), practically leaves blind persons out of consideration, because the metaphor is based on a more or less rich graphic display of icons, windows, pointers and texts which are inaccessible for the blind. In the first place, this situation motivates the need to have alternative ways to access the more or less complex graphic information depicted on the screen. However, a totally different metaphor that complies with non-visual media requirements and brings the *aural* interface to the front, making audio the dominant medium, is the preferred interface for the blind users. The new metaphor is usually called *Aural User Interface* (AUI), based on the terminology supported by many authors, e.g. [1].

14.2.1 Screen-reader

After some trials with special versions of self-voicing software capable of driving a speech synthesiser and thus providing access for the blind, a more general concept appeared in the 1980s and initiated a family of applications concerning screen-readers. Some modern applications, however, such as the web-browsers Opera and Home page reader, still keep the self-voicing approach. The purpose of the screen-readers is to create a vocal rendering of the contents of the screen which the user controls through the keyboard, using a text-to-speech (TTS) converter, see Wikipedia Screen-reader: http://en.wikipedia.org/wiki/Screen_reader. A properly installed screen-reader software stays active in the operating system and

operates in the background, analysing the contents of the screens produced by *any* software.

From the initial command-line interface (CLI) to the now ubiquitous GUI, screen reader software has evolved much in 25 years. The pixel information in the screen memory has made the task of extracting the embedded textual content virtually impossible. Therefore, the screen-reader software must interpret the operating system messages to build an *off-screen model* with the configuration of the screen and its textual contents. This is a significantly difficult task if an accurate model is needed. In Fig. 14.1 the relationship between the screen-reader, the GUI and the application is depicted by means of a diagram.

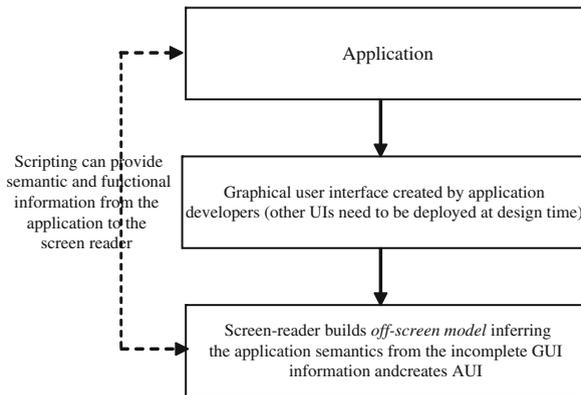


Fig. 14.1 Illustration of the relationship between the screen-reader, the GUI and the application showing the paths for flow of information to build the Aural User Interface. Other user interfaces should be provided by the application developers from the start, enabling relevant input to the screen-reader or a similar program

A wide range of user-programmed modes allow, in general, many kinds of text scanning, from introductory reading of a few characters from the first lines of each paragraph down to the full text reading or even individual character reading possibility. Screen-readers can also analyse many visual constructs and produce speech output from these, such as menus, alerts or dialog boxes which allow interaction with the user. Unfortunately, the power to analyse general graphical objects and produce a voice description still lacks in a major extent.

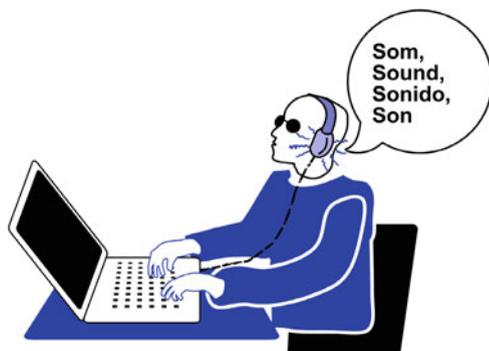
Navigation in the screen is possible as well to allow a non-linear or even random exploration and acquisition of the depicted information. Control of the produced speech is normally given to the user so that fast navigation becomes possible when the user uses shortcuts. A simulation of the use of a screen-reader is available at WebAIM website (<http://www.webaim.org/simulations/screenreader.php>).

Scripting is a technique adopted by some screen-readers to adapt to applications semantics that are not observable in the GUI data, this is the case of, for example, the JAWS screen-reader.

Many screen-reader applications exist (http://en.wikipedia.org/wiki/List_of_screen_readers). Some stand-alone commercial examples are JAWS (Job Access With Speech) from Freedom Scientific (http://www.freedomscientific.com/fs_products/software_jaws.asp) Windows Eyes from GW Micro (<http://www.gwmicro.com/Window-Eyes/>) and Hal from Dolphin (<http://www.yourdolphin.com/productdetail.asp?id=5>). There are a few proprietary operating systems that provide some more or less basic screen-readers, like Narrator from Microsoft (<http://www.microsoft.com/enable/training/windowsxp/narratorturnon.aspx>) and VoiceOver from Apple Computer (<http://www.apple.com/macosx/features/voiceover/>). Emacspeak, from Raman (<http://emacspeak.sourceforge.net/>), is an interesting free screen-reader and aural user interface system among the many available for Linux.

An illustration of a blind person using screen-reader software to access the laptop's content is given in Fig. 14.2. The audio channel is conveyed to the person's ear by means of an earpiece.

Fig. 14.2 A blind person can benefit in a crucial way from the speech output produced by a screen-reader software



There are many limitations, however, that current screen-readers cannot overcome per se. These include, for instance, problems related to images (screen-readers cannot describe images, only output a readout of a textual description of these), to visual layout (the user has no means to realise how the page is organised and go directly to the interesting spot; the screen-reader usually reads linearly and does not skip uninteresting portions) and to data constructs that use positional information in rows and columns, for instance, like tables and plots (data tables can be quite confusing and lengthy when reproduced through speech, placing big demands on the subject's ability for interpretation and memory).

14.2.2 Screen-readers' Technical Requirements

The basic speech-processing requirements for screen-reader applications deal with robust operation of the text-to-speech converter, with the possibilities of spelling and reading random individual characters. All kinds of text elements that may

appear in the text, such as numeric expressions, abbreviations, acronyms and other more or less coded elements, must be taken into account, and punctuation is usually spoken as well, besides its being important in introducing prosodic manipulation in the synthetic voice.

Moreover, formatting information and general meta-information, which are embedded in the object file, also need to be retrieved under user command. This may be signalled by prosodic changes; for instance, the appearance of bold text could be signalled by producing a higher voice tone. Following this idea, the World Wide Web consortium (W3C) introduced the Aural Cascading style sheet (ACSS) in 1998. Within the Cascading Style Sheet 2 (CSS2) recommendation, there is a chapter respective to the acoustical rendering of a web page (<http://www.w3.org/TR/WD-acss>). Acoustical cues or *auditory icons* are acoustical elements that are rendered to the user appropriately by means of loudspeakers or earphones, which are capable of producing some form of surrounding sounds (basically stereo sound). The acoustic elements possess spatial characteristics, and combined with voice segments, they also possess voice properties like speech-rate, voice-family, pitch, pitch-range, stress and richness which are used as command parameters to the speech synthesiser. Compared with linear reading, the use of auditory icons gives the content editor new degrees of freedom in order to describe files and to design his auditory canvas.

Cosmetic features, or cosmetic dynamic changes like changes that are triggered by *on-mouse-over* events, are generally not relevant for information retrieval. However, when relevant they should be made accessible, and a way to do this is to include description of the data in the file's meta-information so that it could be speech-rendered.

14.2.3 Relationship with the TTS Module

Screen-reader software has naturally a close technical connection with the text-to-speech module. However, screen-reader applications are special in that interruptions of the current utterance by the user are always possible, and *barge-in* by the user or by the operating system should thus be provided. In fact, the possibility for interruption may be considered a basic feature of screen-readers. For speech-processing this brings a few challenges. First of all, the signalling of the change of speech stream must be clearly given to the user by means of voice changes and/or prosodic boundaries, or even by additional cues like short acoustic signals, etc. Another aspect to be considered is the *time latency* for the command-to-speech-output. This should be quite small, so that the user can navigate at his/her own pace, and not at the TTS's output pace.

Orthographic error correction tools must also be at the disposal and under strict control of the blind user. Usually, such tools are not used for automatic correction but only to signal errors or irregularities and provide correction hints. Finally, prosody and other supra-segmental speech parameters like rhythm; articulation and

pauses must be easily adjustable by the user. They can also be used automatically by the screen-reader compliant with the ACSS in order to mark acoustically some features of the rendered text based on the design of the creator of the page or the document.

14.2.4 Audio-Browsing Tools

Text or webpage structure analysis and parsing of the text or the navigation web page contents are necessary for speech-rendering. The structure or outline of a web page can be discovered through a careful analysis, and used as a table of contents. For instance, in the Emacspeak package (<http://emacspeak.sourceforge.net/>), a tool for the outline processing of files is available. However, the idea was fully implemented with an interesting success in the project AudioBrowser, funded by the Fundação para a Ciência e a Tecnologia de Portugal (FCT) for 2003–2005 (<https://repositorium.sdum.uminho.pt/bitstream/1822/761/4/iceis04.pdf>). AudioBrowser was developed for Portuguese, but it is applicable for most other languages. It is a special browser which has the capability of displaying separate windows with related information: the created (or existing) table of contents of the webpage is linked to one window, the original page contents is linked to another window and the portion under observation is linked to a third window with magnification (for low-vision users). Figure 14.3 depicts an AudioBrowser screen with the windows 1, 2 and 3.

The user of this application can freely navigate inside the contents of each window, or jump between the windows, i.e. from the contents to the table of contents or vice-versa, in order to scan or navigate through the page in a structured and friendly way. Moreover, blind or low-vision users are helped by the text-to-speech device which follows the navigation accurately. This browser accessibility model is less general but more structured, gaining in terms of speed and clarity of the browsing. It falls back to the traditional approach of page voice rendering, if the user opts not to follow the discovered structure. In this case, voice manipulation is even more important in order to signal the switching between the windows, and it can be achieved with one or two voices in different speech modification styles.

The W3C consortium, through its Web Accessibility Initiative (WAI), has been issuing a relevant set of Web contents accessibility guidelines (WCAG), now in version 2. These guidelines are intended to help in orienting web page design for accessibility (<http://www.w3.org/WAI/>). Authoring Tool Accessibility Guidelines (ATAG), also in version 2.0 (draft), are also important for developers of authoring tools. To ensure that a web page is accessible, one of the best ways to do this is through testing, even if this is done by a sighted developer. Coding errors, images with missing ALT-text, spelling mistakes and grammatical errors, which are often unnoticeable, can be easily detected by testing (http://www.utexas.edu/disability/ai/consult/user_test.html).

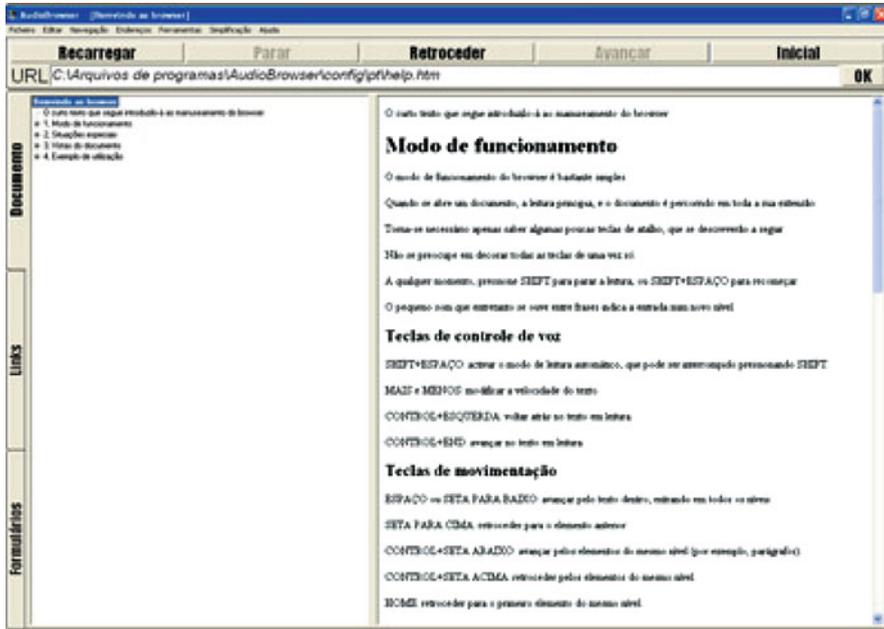


Fig. 14.3 A screen shot of the windows of the AudioBrowser accessibility software. On the left window, the structure or outline of the webpage is presented and on the right the contents of the page are linked to the items on the left. Navigation and speech output can easily be toggled from one window to the other. On top, the third window depicts a magnification of the highlighted text

14.2.5 General Purpose Speech-Enabled Applications

The term screen-reader is in fact quite obsolete nowadays, since the context for the piece of information that is conveyed through voice is not the screen anymore but a window or just some information or a dialog box. Windows-based software is normally associated with a set of menus and dialogue situations which complicate non-visual navigation and operation tremendously. The main problem is to keep track of the events that pop-up in order to *barge-in* the speech stream, and to provide adequate alerts and messaging to the user. The use of auditory icons and of an auditory user interface is an interesting approach to address this problem.

The reading of documents from text files is promptly available through screen-readers. However other formats are also available, like the DAISY format for talking books (http://www.daisy.org/about_us/default.asp). This format (ANSI/NISO Z39.86 Specifications for the Digital Talking Book), currently in version 3, includes digital audio files containing human narrations of the source text, a synchronisation file to relate markings in the text file with time points in the audio file, and a navigation control file that enables the user to browse with synchronised text and audio. The audio files can be produced through a text-to-speech system (e.g. Dolphin

Producer <http://www.dolphinuk.co.uk/education/products/producer.htm>) as well as from natural voice recordings. If documents are in a printed form, the OCR (Optical Character Recognition) software can be used as a front-end, and the resulting text can be treated as a regular text file.

Authoring of documents, on the contrary, can be done by means of dictation, i.e. using automatic speech recognition software (ASR) and speaking the resulting text back with the text-to-speech system. Text can also be entered through the keyboard and spoken back. In these cases, as well as when inputting text through the OCR, an orthographic error correction must be performed on the recognised or entered text, and the errors be aurally highlighted.

A number of software programs for productivity, programming, communication or leisure are already speech-enabled and they offer an extensive set of possibilities for the blind users. Search facilities, which use aural formatting properties, complement the normal GUI search in addition to common speech-enabled search engines. Also programming is possible using speech-enabled development system software in a variety of languages, with the assistance of modules like a speech-enabled inspector and wizards. On the other hand, distraction software and games pose interesting problems for an AUI, and some relevant results have been discussed for example in [2].

14.2.6 Ambient and Security Problems for the Blind User

Privacy and, most of all, security of personal information are important issues when employing some form of speech output in non-controlled environments, such as public or similar places (Rui Almeida, Consultideias – Lisbon). In Fig. 14.4, such a scenario with a blind person paying some goods at the cashier in a shopping facility is depicted. For these scenarios, a private acoustic voice channel is thus necessary,

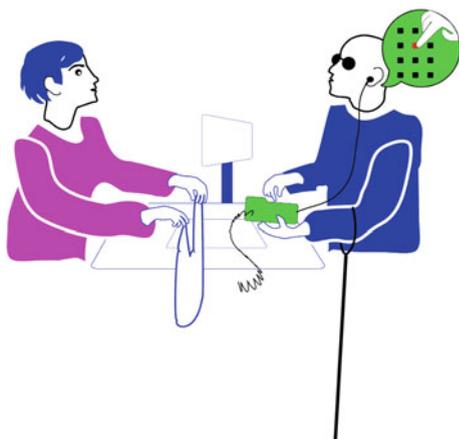


Fig. 14.4 At a cashier, a blind person may use a payment keypad with an adequate speech channel that conveys the information about the amount to pay and the keys pressed. The connection could preferably be over a wireless channel

instead of the one based on sound emission from a localised loudspeaker. An earphone would be a convenient solution, but it has some problems if the physical connection is based on a plug and a socket: this would require insertion or positioning of the plug into a certain location, and the information about this must be passed onto the user. Hygiene problems require the blind person to use his/her own earpiece that he/she must carry all the time.

When the text-to-speech device speaks aloud the amount of payment and the pressed keys, the required privacy in the audio transmission can be kept by means of an earpiece and a cable that the blind person uses. A wireless connection with the earpiece would be best if only automatic login is available. The relevant data could then be recorded on the person's banking card or identified by other network-based means of identification. For instance, a wireless earphone can be identified in the network or in the special terminal through its *MAC address*. MAC (Medium Access Control) address is a universal code that electronic communication devices possess. With an automatic authentication, the setting-up of the connection would be quite practical and the privacy issue could be resolved.

It must be noticed that in public places the terminals are often randomly located, and this causes the first barrier for the blind user, namely finding a terminal without the common visual aids. However, this issue belongs to the scope of general accessibility, so we do not deal with it at this point.

Concerning the input device, the keyboard is quite safe and convenient for short transactions, as far as the device is conveniently located, out of sight of other persons, as well as reachable and identifiable in terms of type and layout. In private environments or inside individual booths, the keyboard is thus a good solution. The identification of the type of keyboard remains the main difficulty, but an alternative means of input can be used based on automatic speech recognition (ASR). In other environments, both the keyboard and the ASR become inconvenient. On one hand, the identification of the keyboard type is more difficult, and on the other hand, the speech could be overheard by others breaking the needed privacy. One possibility to alleviate the problem is to use speech through a combined wireless microphone/earphone associated with a special dialogue system prepared to randomise as much as possible of the required voice input that can be overheard together with quite a simple keyboard input.

14.3 Applications for the Mobility Impaired

Mobility impaired users are generally prevented from (or severely limited in) accessing the common information systems' input interfaces. The basic physical inaccessibility to the site, room or sidewalk where the terminal is, again falls out of the scope of this chapter, but there are difficulties that arise in various other ways due to the general "keyboard and mouse" approach.

Of special importance is the impossibility to reach the interface devices due to an elevated height of its placement. Although the location is adequate for a standing

person, it is not so for a wheelchair user or a user with upper limbs impairments, as it is difficult or impossible to reach the body position that is required in order to operate the interface correctly. Eventually, it is often difficult to do the mechanical manipulation of the parts of the device, see Fig. 14.5. Examples may be found in several situations.

For instance, access to kiosk terminals or to most of the present day ATMs is frequently difficult for persons with mobility impairments. Besides the elevated position relative to the ground, the re-entrant shape, the placement of different buttons all over the front panel and the force needed to produce a valid keystroke cause big problems for these persons. The required arm movements can be difficult to achieve due to the amplitudes, the elevation and the force required.

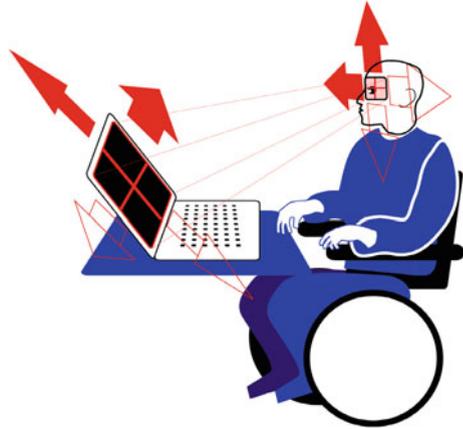
Also the use of conventional console-based tabletop interfaces is difficult for persons with upper limbs impairments: Manipulation of a keyboard and/or a mouse can be impossible because of the reduced strength and/or the lack of control of the required movements, as is frequently the case with patients of neuromuscular diseases or cerebral palsy.

As the information stream in the applications mainly consists of text that is employed for the data and for the commands, some available solutions concern virtual keyboards so as to avoid the need for manual manipulation. Virtual keyboards are displayed on the screen with mechanisms based on key scanning and selection, and triggered by a switch commanded by one finger or by a small air jet expelled from the mouth and picked up with a microphone. Also eye control with an eye-gaze system using the fixation of the gaze and/or the blinking action



Fig. 14.5 At a cash-machine a wheelchair user usually has tremendous difficulties in using the keyboard and display, unless special attention is paid to the design

Fig. 14.6 An eye-gaze system may be very useful for a mobility disabled person at the upper limbs for handling the terminal



can be used. Figure 14.6 depicts such a situation. Random access to the virtual keyboard may be possible using only one finger by means of a miniature joystick. Electro-myographic (EMG) and electro-encephalographic (EEG) detectors may also provide triggers that a specific transducer can use to produce the desired switch action. Image analysis of the head position can also be used to point to the screen, although this requires more precise movements of the head which are tiring or even impossible for mobility impaired persons to do.

The importance of speech-processing in this wide class of problems thus appears as an alternative medium and alternative channel that can be used to overcome both the physical gap problem and the functional difficulty in handling the console. Let us start by considering the input. If the user has good enough vocal capabilities, the environment is not very noisy, and privacy is not an issue, a normal automatic speech recogniser (ASR) available in the terminal may provide a voice-to-text input which enables both commanding the console and inputting data. In a public terminal a speaker-independent system is required, which may be difficult to obtain for many languages with a good enough quality. In an ambient-intelligence environment, once the user is identified, the system could load the customised speech models of the user, if available, for the recognition to be as accurate as possible.

If the user has lost some of the speech capabilities, the common ASR becomes less effective and some add-ons must be looked for. In the case of reduced vocal abilities, the first issues to consider is whether the user is capable of producing a reasonably consistent set of recognisable words or part of words like syllables. Many persons with cerebral palsy can only produce sequences of articulated syllables due to their difficulties to control their vocal system. The question is then whether these utterances are consistent and stable enough in order to be used to generate good speech models for a special customised ASR.

This can be assessed by means of recognition trials. If the answer to the question is negative, then recognition with a smaller search space may bring improvement to the recognition rate and speed. This can be done, for instance, by using eye-gaze or eye-wink strategies together with a scan system, a virtual keyboard and a text

predictor to obtain a text input. Spatial context driven speech recognition, by which a small part of the screen is at any moment considered the target search space for acoustical recognition, is therefore a possible solution that combines use of speech and the pointing mechanism, such as eye-gaze or head position (Vitor Pera, FEUP).

However, if the answer is positive, a set of words may be used as a code of indexes for accessing the row and column of a virtual keyboard. Thus a random input can be produced in combination with, or even replacing the scan input mode. Together with a text predictor, the audio recognition of the words can work as an effective speech-to-text based input system. Persons with cerebral palsy often possess some vocal abilities that permit such solutions (Vitor Pera, FEUP).

The robustness of the recognition of even a small set of words or syllables may still be disturbed by many reasons. For example, the ambient noise and difficulties in voice pick-up frequently cause problems. To overcome these problems, the use of customised headsets for holding the microphone, as well as techniques to improve robustness of the processing, is needed. A wide range of such techniques is available.

The main rationale for achieving robustness is to understand how the interference signals affect the main characteristics of the speech signal. From this understanding several types of techniques can be used mainly for signal acquisition and for acoustical modelling [3]. Besides the careful choice of microphones, the main methods are connected to the extraction of robust acoustic speech parameters [4, 5], like in spectral subtraction, cepstral mean normalisation, RASTA, signal bias removal, etc., and to the use of special modelling techniques [6] like parallel model combination, auditory modelling, multi-band and multi-stream processing.

Multi-stream speech recognition is a particularly interesting approach because it offers a possibility to combine different information streams in order to supply the most reliable features at any given moment. Several streams may be extracted from speech and of special interest are the streams originating from different related media, like in audio-visual speech recognition. The facial video is captured together with audio and combined together with it as an additional stream. The choice of different streams must follow some criteria. First, each stream must convey discriminative information in linguistic terms, to enable the recognition task. This can be discovered by measuring the mutual information between each stream and the linguistic classes to be recognised. Moreover, each stream must convey redundant information as little as possible. For instance, the mutual information between the linguistic classes and the visual stream conditioned to the acoustic stream must be significant.

Finally, the relative contribution of each stream to the decision process must be specified as a function of some indicators of their reliabilities. The combination of the facial expressions with the voice, in a multi-stream speech recognition task, may bring up the recognition rate value to a usable level. This is the basis of the experiments that were done in the author's laboratory in the recent years [7]. Although there are face-tracking techniques available, the video signal of the dynamic facial image is picked up by a special miniature camera placed whenever possible in a

fixed position relative to the user's face, and delivered to the facial image processor in real-time, together with the speech signal from the microphone. Experiments done with a neural muscular patient, concerning simple audio-visual recognition task consisting in the use of a virtual calculator, revealed that the use of the facial expression can enhance the audio recognition in the case of acoustic noise, even if the region of interest (RoI) only concerns the lips region like in the present experiment [7, 8]. The audio-visual approach thus deserves to be continued in spite of its computational load and of many challenges that concern the extraction of facial features and their recombination with the somewhat optimised audio stream [9].

14.4 Applications for the Speech Impaired

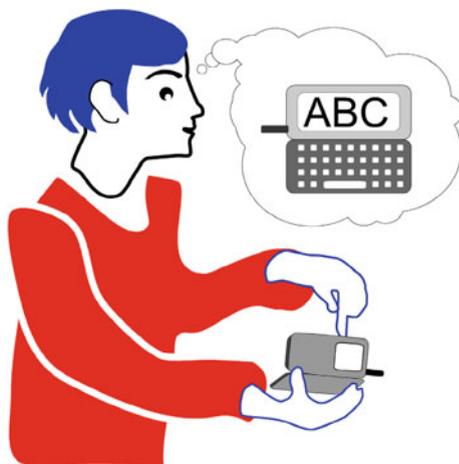
Whenever the bi-directional speech channel is hampered due to a limited capability either in producing or receiving speech, a text-telephone is a common solution. Text-telephones must comply with the technical standards for telecommunication (ITU-T V18 is the latest international standard, see also Gunnar Hellström: http://tap.gallaudet.edu/access_standards.htm) which are, unfortunately, far from being commonly accepted in the global market. The main problem, however, is the lack of communication between the non-disabled users' audio phones and the text-telephones, due to the fact that there is no speech-to-text and text-to-speech conversion available to perform the necessary transformation.

As for the speech-impaired persons, there are specific issues to address when they use a text-telephone. First of all, there is a need to produce an output into the telephone line. However, with appropriate software and a modem hardware, a text output could be easily generated and transmitted from a text-telephone or from a personal computer through the PSTN (Public Switched Telephone Network). The reception of text is also readily available by means of another text-telephone or another PC, so there should be no additional difficulty to communicate between text-telephone users.

Second, if the interlocutor does not possess a text-telephone, problems obviously appear, although several possibilities also occur. For instance, the speech-impaired person or the interlocutor may possess a text-to-speech device that can produce synthetic speech output. This can then be either transmitted through the network or synthesised locally from the received text. A third possibility is the use of a relay service. This is a facility that some telephone service providers offer to their clients in order to convert the received text into speech. Earlier the relay service was a human-mediated conversion service, done by an operator who read the incoming text to the interlocutor. The current trend is to make relay services automatic.

Present day mobile phones and PDAs offer reasonably good text writing facilities through a keyboard, combined often with text prediction as well. Figure 14.7 depicts a situation where a speech-impaired person uses a text channel on a PDA. Internet Protocol data transaction and messaging or a chat software with audio and video channels may function as the text-telephone for users that are computer literate.

Fig. 14.7 A mobile phone or a PDA may provide the needed text communication channel for a speech-impaired person to communicate with others



When speech is used for output, there is also a need for text-to-intelligible-speech communicator function. The starting point is the text to be communicated. A text-to-speech system can handle input text and produce the necessary speech output, but for a good quality communication, careful control of the prosodic features of the speech output is needed.

In the SPEECH-AID project (1994–1997), a speech communicator was developed and tested on a laptop PC platform. The purpose was to allow, among other possibilities, telephone communication for severely speech-impaired persons. Test runs with a small group of laryngectomised persons at IPO (Portuguese Institute of Oncology, at Porto, Portugal) showed great interest in this application although the laptop would be too heavy for a convenient daily use. The system was built on the MULTIVOX multilingual speech synthesis system, using European Portuguese. The text-to-speech was formant-based, using four formants and a database specifically constructed for the purpose. The synthesis strategy was based on diphones, and the short-time signals were formant-coded with 10 ms coding frames. Phonetic and phonologic aspects of the synthesiser followed a rule-based approach.

Several products that follow the general communicator concept exist in the international market, for instance, a mobile phone equipped with a text-to-speech synthesiser. Another point to notice is the choice of synthetic voice that is used in the replacement communication aids. The loss of one's natural personal voice can often be predicted in advance so that a sufficient set of audio files can be recorded for the realisation of a voice-customised speech synthesiser. The idea is then to process the audio tracks by general software which, by means of a subsequent segmentation and labelling of the speech material, produces a voice font for the speech synthesiser. Later, after surgery, the laryngectomised patient may, by this way, keep doing part of the communication with a voice similar to the original.

14.5 Applications for the Hearing Impaired

Hearing impairments can vary much in their severity. Light-to-moderate hearing loss is quite common in today's population, but it can be considered a minor problem compared with deafness that occurs together with ageing. In general, more profound hearing losses should be considered together with deafness.

Language and speech acquisition of the deaf depends on the age of occurrence of deafness: the earlier in childhood deafness appears, the more difficult is language and speech acquisition. A child that is born deaf may never be able to speak fluently. Sign language is a natural alternative that the deaf persons use for communication.

Speech communication in the case of hearing impaired users is therefore related to the availability of speech conversion systems, i.e. producing speech output from text input (text-to-speech function) and producing text output from speech input (speech-to-text function). Persons with early profound deafness therefore require a bi-directional converter between speech and text.

Earlier, the main solution for telecommunication for the deaf was human-mediated relay services offered by the telephone network operator; cf. a similar case for speech-impaired people discussed above. Unfortunately, this has not been available universally, and therefore there have been serious limitations to the communication of speech disabled persons. Video-telephony has thus been explored as a viable multimedia channel for communication in sign language, but it has, however, been too expensive due to high equipment cost and large bandwidth requirements in order to allow live sign language communication. The development of Internet

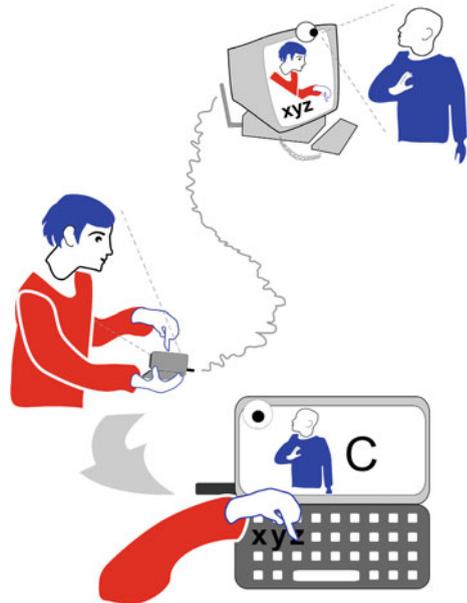


Fig. 14.8 Sign language is important for the deaf and speech-impaired people's communication. When the interlocutor cannot use sign language text communication is needed as a replacement of the video or superimposed to it

as well as improvements in video pick-up and transmission offer a convenient solution with cheaper IP calls and a video-telephony on PC equipment with special high-quality cameras.

A communication scenario between two people is depicted in Fig. 14.8. The availability of video, audio and text communication channels together with speech and text converters makes this scenario almost universal. In fact, it is possible for two deaf persons to communicate in sign language, and if either of them cannot use their hands, e.g. because of holding a palm terminal, text communication is possible on the screen. However, most non-deaf people are not capable of communicating in sign language, so there is a need to use speech/text-converted channels. Communication is thus possible with people who cannot use sign language, by using the speech-to-text (ASR) converter that produces a text output at the deaf person's terminal screen and by using the text-to-speech converter (TTS) in the opposite direction, to convert the text into speech at the interlocutor's terminal whenever the deaf person is not capable of speaking and video sign language communication is not possible.

14.6 Applications for the Elderly

Specific requirements concerning elder people's interaction with information and communication systems deal with the decrease of sensory acuteness (hearing and sight) combined with the loss of fine control of hand and limb movements. Moreover, gradually increasing cognitive difficulties due to memory degradation are manifested in the management of daily tasks and in responding to other people's requests not to mention task requirements imposed by various devices.

In such complex situations speech technology can be used to contribute to a desirable simplification of the interaction to decrease cognitive load and to increase impression of security. To achieve these goals, one can, first of all, get inspiration from human behaviour. One of the main observations is the redundancy among the several media involved in the communication. For instance, the combination of speech and vision is a basic one which can be effectively used for this purpose. Due to their experience in watching TV, sub-titling of image communication is in general a good communication channel for the elderly people.

Another important observation is related to speed and rhythm of communication. Elderly people are slower while communicating, and this poses some restrictions on the parameters of speech conversion devices, particularly on output devices, like the TTS.

Cognitive load and memory demands must also be taken into serious consideration. This means that at each step of the interaction, an elderly user need not clearly remember all the premises that affect the present situation, and may feel quite insecure when dealing with the concepts involved in that interaction. The information and communication systems must thus possess good knowledge of the elderly users' characteristics, i.e. the system must have the necessary technical

capability to acquire and store relevant user data. The data must also be organised in such a way that the communication management can access the relevant parameter values and decide how to use the different media and media combinations, without loss of security and privacy. During interpersonal communication it is also good to provide as much redundancy as possible with combination and association of media. Human interlocutors can react appropriately in order to lower the stress of the communication whenever cognitive overload is detected, while an automatic communication system can use the available data on the user's characteristics to decide whether video-telephone communication and real-time sub-titling is needed.

Let us consider, for instance, the case of an elderly user accessing a health care service through the web. The user terminal gives a range of information of the user characteristics available to the remote service server, which can then switch on the video-telephone mode and run the service with appropriate dialogues in this mode. A synthetic face agent (avatar) with synthetic speech would then personalise the service agent and interact with the elderly user. Guidelines to orient the design of accessible software agents are published, for instance, in the User Agent Accessibility Guidelines 1.0 from W3C (<http://www.w3.org/TR/2002/REC-UAAG10-20021217/>). Synthetic speech should also be adjusted so that the user's preferred voice fonts with an appropriate speech rhythm and prosody are used. Sub-titling would also be used on the screen if the user characteristics indicate this, and the use of pictures to represent objects, places and ideas should be eventually selected.

If the reason for the elderly person to call to the health care service is to recall what medicines are prescribed for that day, a careful check of the user's identity and characteristics is necessary. If the system discovers that the user's age is quite advanced and has sight and hearing problems, it can then decide that the right way is thus to use the video-telephony mode and the dialogue system should follow the picture-based approach for interaction. After a series of interactive steps in which the user gives and receives information about the medicines, a printout is also sent to the user's printer with pictograms and images of the medicines. The local terminal alarm with a speaking agent is also programmed with the data just defined for the day takes. The interaction is thus based on speech recognition, which has been customised to the user's speech, and on an advanced dialogue system which takes care of the whole interaction.

The most important aspects for the development of present day's technology towards this scenario are related to the dialogue system and agent manager and to the speech engines' modalities of operation and customisations to the user. The development of such systems has been under continuous research. For instance, [10] describes a range of techniques and experiments and proposes a system architecture, which has been used for instance in [11]. An important aspect to consider in the development of the interactive interface systems is their robustness and adequateness. There is also a need to have a back-up from a human operator in case of dialogue failures so as to prevent frustrating experiences that would draw the user away from future use of the system.

14.7 Accessibility and Application

14.7.1 Navigation in Built Environments and Transportation

Built environments and transportation are particularly important issues to consider from the point of view of accessibility because, in them, the disabled person is confronted with artificial barriers derived from the space restrictions and the technological devices. The complexity of the space also causes health and injury risks. On the other hand, these locations are unavoidable and needed for all purposes of daily life, so accessibility is important for social inclusion and the individual's needs and rights.

Navigation is a natural activity that requires support in various ways. The main items that are provided concern information about the locations, and guidance in these locations. In the planning phase of navigation, information resources are the most important ones, but in the travelling phase, the guidance resources become dominant.

In order to see how system interface can help in these situations, it is useful to discuss three related daily life paradigms. First, in a travel guide paradigm, a travel guide allows the user to access information about a certain location and thus enables such activities as the planning of a trip, building of an itinerary and making of a list of actions to accomplish. Another paradigm is the human tour guide who acts as a guide and companion in the travel. The guide needs to possess excellent communication skills and a fluent knowledge of the features of the location. The third paradigm is a virtual tour. The relevant information is used to reconstruct a simulated space or location by computational means and artificial representations so that the user can explore the space with as much sensorial completeness as possible.

The main difficulties in travelling in the built environment and in transportation systems come from the need to identify, at any moment, the sites that are referenced in the travel plan and the directions to reach the next one. In general, each site has a number of signs and plates which indicate directions and other information, and which help sighted people to navigate. Disabled persons like the blind and the cognitively impaired have total or partial difficulty in accessing this visual information.

Another difficulty is to access the complex information involved in each location. Generally, this is solved by means of a travel guide, a tour guide or local information panels. Disabled persons could benefit from these solutions if they were made accessible. It must be noticed that neither travel guidebooks nor tour guides are a solution, so an alternative should be found. This can be achieved by implementing the ideas and concepts concerning a virtual tour, a tour guide and a travel guide into a system that combines them into a new type of information, guidance and navigation system.

In Portugal, a long-awaited opportunity to apply speech interface concepts in the field of transportation appeared with the NAVMETRO and INFOMETRO projects that were funded by the Portuguese government program POSC (Operational Programme for the Knowledge Society). The consortium consisted of the Faculty

of Engineering at the University of Porto, the metro transport operator company Metro do Porto, S.A. and ACAPO (Associação de Cegos e Amblíopes de Portugal, the Portuguese Association of the Blind). The aim was to introduce accessibility features in the city metro network to allow blind users to travel as independently as possible.

The main target was to provide access to the printed information and visual information signs at the metro stations, and also to give guidance during the travel inside the station buildings, through the mobile phone, and by employing dialogue and speech technology. The requirements for the system thus are that it should produce speech-mediated access to information and guidance, while a dialogue system control, speech input and speech output are essential interface elements of the system. In terms of privacy, speech input may need an alternative so that the user need not speak to his/her communicator in public places.

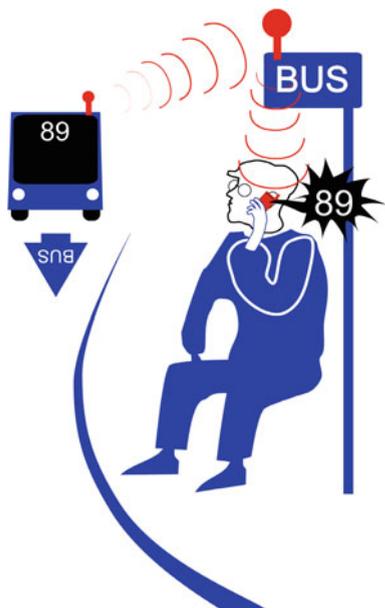
The tour guide is a virtual agent, interacting with the user through speech and providing guidance for the actual travelling in different locations. A virtual tour is also accessible through speech, and it allows the user to discover detailed navigation information relevant for the intended travel. Speech directions are not always unambiguous, and some other physical cues may thus be necessary, depending on the complexity of the information to be presented. Acoustical or sound cues for the blind and visual or light cues for the cognitively impaired users are an example of this kind of multimodal signalling. In the NAVMETRO system a network of sound “buoys” was created and used in combination with the dialogue system (a patent request was submitted for this system). NAVMETRO was publicly demonstrated on the 30th June 2008 and entered into public service on the 2nd December 2009 at Trindade station, Porto, Portugal.

Another crucial aspect in this application is the user tracking. This is the most effective way to coordinate the guide-book, the tour guide and the virtual tour. Such user-tracking technologies exist in the market that can provide the required connectivity between the system’s infrastructure and the wireless user. In the NAVMETRO system an acoustic user tracking system was developed as well.

Besides the metro stations and similar locations, other suitable target locations include public places away from the normally used spaces that the user attends. The users’ home and workplace are generally too familiar to the user to need guidance from the system. Although built environments, including transportation facilities and vehicles, are the most relevant target locations, the concept may be extended to outdoor locations as well. In Fig. 14.9, an example is given of a situation in which the personal area network allows the exchange of transport information that is essential for a blind person waiting at the bus stop. (This idea was developed from interesting discussions with Dr. Rui Almeida, Consultideias, Lisbon.)

Personal area network shown in Fig. 14.9 is one of the ways of creating the required wireless connectivity. Some international R&D projects have targeted these general problems. For instance, the EU FP6-IST e-Inclusion project Ask-IT (Ambient Intelligence System of Agents for Knowledge-based and Integrated Services for Mobility Impaired users) (see <http://cordis.europa.eu/search>) is of particular interest as it is comprehensive and addresses many needs of the disabled

Fig. 14.9 Personal area networking communication allows the exchange of transport information that is essential when waiting at the bus stop



users. The approach to the design of the dialogue systems is global effectiveness which requires the integration of user data with advanced adaptive dialogue interaction and speech-mediated transactions of the relevant data. Examples of usage situations for the Ask-IT kind of a system can be found every day in professional or private functions.

14.7.2 Access to Complex Documents

The main objective for introducing this sub-chapter is to describe what is happening in the field of accessibility to scientific and mathematical contents for persons with special needs. In general, the digital representation of mathematics has been an object of research for decades. The first usable results appeared in the context of the edition of scientific documents in the 1980s. LaTeX coding of mathematical formulae or expressions is a linear structured code that allows a semantic approach into the formula or expression contents (<http://www.latex-project.org/>, see also <http://en.wikipedia.org/wiki/LaTeX>).

A general form for description of a formula can be done using basic mathematical entities followed with a series of comma separated arguments. For instance:

$$X = \text{int}(0, \pi, \sin(x)dx), \text{ that represents } X = \int_{-\pi}^{\pi} \sin(x)dx$$

is the code necessary for the print representation of the “integral of $\sin(x)$ taken between 0 and π .. dx ” of an hypothetical text processor’s math module. As the main purpose of this codification was the printing of the document, that meant a conversion of the formula to an image, it took a while before other conversions started taking place. However, since mathematical objects are graphically two-dimensional and the relative and absolute positions of symbols and operators are essential, the codification thus evolved by approaches which introduced a series of rows and columns to address the positioning of the symbols.

The advent of the Internet brought up the need to convey the code information in a coherent and universal way, so efforts were directed towards obtaining a different and modern codification that could function as a compatible and interoperable platform for the various players in the field, namely, for those who generate mathematical contents and render software web-browsers’ add-ons.

MathML (<http://www.w3.org/Math/>, see also http://www.w3.org/Math/Software/mathml_software_cat_accessibility.html) emerged in the context of the W3C as a response to this need. It is a de facto standard and some software houses already adhere to it in order to make their products compatible. However, as pointed out by Jan Engelen (ESAT, K.U. Leuven, Belgium), the existing e-learning system lacks available software that enables accessibility to mathematical contents. Thus, to allow disabled students in the campus to use e-learning software, a research project was started at FEUP’s Speech Processing Laboratory (LSS) in 2000, with the objective of producing an audio rendering of mathematical expressions. This was to be done by means of a software module suitable for machine–human communication with blind students in the faculty. Taking into consideration the on-going initiatives over the LaTeX base, for example, the work of T.V. Raman in the development of ASTER (<http://www.cs.cornell.edu/home/raman/aster/aster-toplevel.html>), and a few other R&D projects, the Audiomath 2005 system was designed and implemented for Portuguese [12]. The system is presently in a pre-release phase, and a demonstration page is available (http://lpf-esi.fe.up.pt/~audiomath/demo/index_en.html).

Audiomath 2005 starts from the assumption that mathematical expressions can be entirely and unambiguously transmitted through speech only. Although this assumption seems reasonable, it unfolds into lengthy descriptive phrases even for moderately complex formulas. The user’s hearing memory risks to be overloaded and the user will not be able to retain a good mental representation of the target formula unless something is done to ease up the process.

A few critical remarks must take place in this context. First of all disambiguation requires that formulas’ elements be kept together with their fellow neighbours in sub-expressions and not be connected with other elements that belong to other sub-expression. A solution for this is to employ words that signal the existence of boundaries between sub-expressions.

$$\sqrt{a^3 + b^2}$$

Consider, for instance, the expression above that can be rendered through speech as: “square root of a to the third, end of radicand, plus b squared”. If the boundary-signalling element “end of radicand” is omitted, the textual description produces a different meaning (the square root is extended to cover the sum of a to the third and b squared). Of course, the full textual representation is just an intermediate phase before obtaining the speech waveform from the TTS. The prosodic cues that are added to the text are also important and should not be disregarded.

In Audiomath 2005, studies on the speech prosody and the distribution of prosodic cues were conducted when speaking two-level nested expressions. A professional speaker who had understood the mathematical expression, read the textual description and used prosodic cues as much as possible, so as to enhance the understanding of the description. The research results showed that the prosodic cues mainly deal with pauses and corresponding intonation movements in order to signal the formula’s internal boundaries. Two classes of pauses (with different durations) and two classes of intonation movements (two different patterns) were identified to signal two types of boundaries, i.e. major and intermediate boundaries, and the results of the analysis were then organised as rules for the TTS. A more effective and understandable TTS output could be produced by taking these rules into consideration.

Another important aspect to consider is a possible listener’s memory overload during the production of long descriptions of mathematical expressions. The description should thus be broken down into short enough chunks that make semantic sense and an intra-formula navigation mechanism should be introduced. This is to say, the description should follow a meaningful organisational tree but the progression up to the tree leaves should be left to the decision and rhythm that the user desires. For the blind users, the definition and use of a set of arrow keys is sufficient for this kind of navigation. In Audiomath 2005, such an intra-formula navigation mechanism was created and tested with the blind users, and it showed good results in comparison with other audio rendering approaches which did not have a navigation mechanism.

MathML code may be produced in two modes, presentation and content mode, according to the recommendations. While the first is more directed to visual rendering of the expressions, the second is more adequate for audio rendering. Unfortunately, the conversion between the two types is not straightforward. The Audiomath 2005 system was developed for the presentation version of MathML, due to the fact that this is the more frequently encountered mode of output presentation in the different software commonly used for editing mathematical expressions. However, user customisation (user-defined modes) of the reading mechanism is possible in certain ways in order to adjust the system behaviour to the user’s cognitive requirements and to obtain the best possible comfort. The reading of mathematical expressions can thus be considered as a special case of structured text generation for speech production.

Besides the reading of documents with mathematical expressions, another important activity that must be considered is writing or editing mathematical expressions. A screen-reader can help blind users to handle the editing by reading the

result of the current keyboard writing simultaneously with the writing. However, upper limb mobility impaired users, who cannot handle a keyboard easily enough, are prevented from editing their mathematical expressions. In order to overcome this situation, the Speakmath project was recently conducted at LSS. The project was funded by the portuguese POSConhecimento fund under the FCT-RIPD program and coordinated by Vitor Pera (FEUP). The main objective of the project was to study and design a speech recognition interface that would allow the control of software that is used for editing mathematical expressions. In order to cope with the voice difficulties, a previously developed multi-stream approach was used with the combination of speech and facial video.

14.7.3 Applications for Instructional Games

Speech is becoming used more and more in instructional games and it seems that in combination with graphics, speech is an excellent medium for presenting enriched displays of objects, and facilitating situations where the objects are used in learning or playing. For instance, in math applications, conversion between written and spoken representations of mathematical objects, such as numbers and their basic arithmetic operations, can be quite easily handled. This observation, together with the fact that no math application could be found for the children at initial school age, to help them to learn numbers in Portuguese, lead to a project which started in 2001, and used a text-to-speech (TTS) engine and a Visual Basic multimedia application in order to help to teach children how to read numbers. A prototype was produced, called “A Quinta dos Números” (AQN) (The Farm of Numbers), and based on a scenario of a farm with animals which intend to build a bridge. The scientific basis followed psycho-pedagogic concepts, inherent to the teaching of maths, artificial intelligence, speech/text processing and multimedia systems [13, 14]. AQN offers a multimedia interaction with speech synthesis and text boxes, and it is designed to assist the teacher to consolidate the student’s knowledge, and to support progressive evaluation of their performance. AQN offers a choice of didactic games with learning and evaluation aims, and simultaneously it also allows accessibility for disabled students [15].

The main concept explored in AQN is the simultaneous integration of the three media: text, graphics/animation and sound/speech. This allows the student to try and practice reading of numbers and to play in free activities while performing evaluation. Moreover, the combination of text, graphics and speech provides rich interaction and a comprehensive teaching environment for the students. The students can watch the number on the screen and listen to its pronunciation, and thus relate the object and its characters to words and specific sounds. The use of TTS has been shown in tests to allow extraordinary interactivity and comfort.

The reading of numbers is developed in a bi-directional structure of numeric categories and their readings, associated with graphic representation, animation and speech synthesis, with a clear correlation between text and numeric representation.

AQN was built as a disguised interface based on Visual Basic, with soft animations, intuitive sounds and non-aggressive colours. All the buttons and icons have a tool tip text that can be read by the text-to-speech engine, while dynamic messages are produced with the help of a TTS engine in European Portuguese. All number categories are marked with a custom mark-up language called Lab205ML which identifies the type of the number and directs the TTS's text converter in the conversion. Finally, when the students input their name at the session start, the system produces a warm reception through a greeting sentence which includes the student's name. This is gratifying for the students.

Although several multimedia math applications exist in the market to support teaching of the basic math operations, geometric analysis, algebra as well as more complex operations, AQN is an interesting application in its realm in that it combines pedagogic aspects, a text-to-speech engine and a basic expert system to guide and evaluate the student knowledge. The system is planned to be connected to a more advanced expert system that will try to understand the student's progress on the proposed exercises, detect difficult levels and advise the next steps.

Figure 14.10 illustrates some screen shots of the AQN application. Tests were held with students in the classroom, and they show an interesting reception to the application. The moderately low quality of the TTS's voice at the time caused some slow adhesion to the software, but was rapidly overcome because of the interest that the activities created in general.

14.7.4 Accessibility to Ebooks

The dream of making books accessible to everyone has materialised quite a lot since the advent of the ebook. (It must be noted that in this context the name refers to the work in digital format, and not to the electronic reading device that has the same name.) The Open eBook movement aims at bringing significant enhancement to the use of ebooks with speech technology specification, and e.g. the International Digital Publishing Forum has activities in the standardisation of ebooks. Specifications and documents on accessibility can be found in <http://www.idpf.org/specs.htm>. Also the National Information Standard Organization (NISO) at the USA has worked in relevant standards, namely, the Digital Talking Book Standard (<http://www.loc.gov/nls/z3986/index.html>).

The possibility of using a text-to-speech device to read the book brings forward two new aspects compared to the reading of a paper book: combination of text and the rendered speech on one hand, and the replacement of text by speech on the other hand. A speech-to-text (STT) device based on the speech recogniser technology can also help interfacing the software to the user, as the user may command browsing and other operations through voice. Let us examine these possibilities in the light of accessibility.

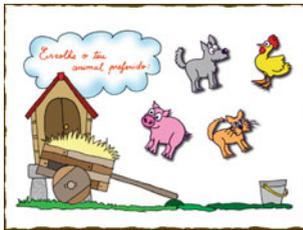
Learning disabilities (LD) is a general expression that refers to the difficulties and disorders in the acquisition and development of language, speaking and reading, and



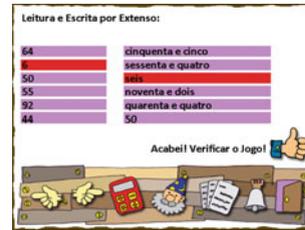
Introduction Window



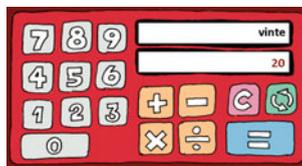
Input for student's name or team name



Choosing a pal for games and learning



Association of number and text formats.



The basic talking calculator.

Fig. 14.10 The screen shots depict phases of the use of the AQN application

of skills related to communication. Sensory deficits such as blindness or deafness should not be included in this group. Combination of speech and text is very important for enriching the mental representation that the reader can have. A stronger image is possible through this combination. Therefore it is foreseen that an ebook equipped with a good TTS may be much more than either the text or the speech when separated. Specifically, reading software is required, that can bring the full dimension of the text into the reader's mind. For this purpose, a flexible control of the TTS operation is needed in order to allow a sufficient integration with the reading gestures. As visual reading of a displayed text is significantly non-linear and information about the reading point is difficult to obtain, although an eye-gaze device could be used for this purpose.

The combination of text and speech can be governed in a coordinated manner by the speech output and reader interface actions (i.e. keyboard or speech input). In this way, the speech stream can be smoothly produced if the reader so desires, but it can also be abruptly stopped and directed to another part of the text by keyboard action. The appropriate behaviour would be such where the TTS module stops and empties the memory buffer at the end of the current word, and re-starts speaking at the new

location which the reader has commanded. This kind of behaviour is similar to the one that a screen-reader has, see above in Section 14.2.1.

Another important aspect is the need to highlight the text being read at each moment on the screen. This way the human reader can have real redundancy as the highlighting concerns both text and speech media. However, the feature is optional because a good human reader might find it too guiding. In the case of children with learning disabilities, including dyslexia, the synchronisation between text highlighting and speaking is believed to be crucial for the good apprehension of the text as well as also for the good leaning of reading and speaking more correctly.

A project is currently under way at LSS to explore the use of ebooks with TTS and apply them to facilitate the learning of language communication for young students with dyslexia. It is believed that both reading and speaking abilities may improve with the use of talking ebooks. User mode text pre-processing and prosody control are important aspects for the success of talking ebooks. Reading of punctuation and text-embedded signs is possible and a careful control of intonation, rhythm and pauses of synthetic speech is needed.

As mentioned in the beginning of this section, another approach for ebooks is medium conversion, i.e. to replace text by speech. Audio books form a different approach, because in them speech is pre-recorded and reproduction is done using the DAISY standard (see above Section 14.2.5). Replacement presents a special interest for readers who, although not being blind, are print-disabled and not able to read. However, in audio books it is generally impossible to change prosody, which can introduce cognitive difficulties in the listening due to unclear segmentation.

TTS-enabled talking books are different from audio books which are based on recorded speech. In TTS-enabled ebooks, a complete flexibility for reading is available, including rhythm and speed as well as intonation control. The speed control is crucial for some users so as to slow down the reading due to cognitive or other difficulties. Some non-profit organisations take care of publishing TTS-enabled ebooks which may be obtained in reasonable conditions (see, e.g. <http://www.bookshare.org/web/Welcome.html>). Many publishers and other organisations have interest in this market, too (see, e.g. http://www.pearsoned.com/pr_2006/101906.htm).

14.8 Conclusion

It is clear that the use of speech technology is useful to enable accessibility to digital terminals and digital information for a wide range of disabled people, elderly people and children with learning difficulties. Speech also has high potential to improve usability of applications dealing with navigation in the built environments and transportation travel guides. This chapter has summarised some characteristics of accessibility for different groups of disabled persons, as well as discussed challenges and solutions for various applications using speech technology.

It is foreseen that in the near future a diversity of possibilities for spoken medium management in interpersonal or person–machine communication will continue to increase, becoming integrated in the daily life of all citizens. This presupposes higher inclusion of disabled people in the information technology society, through improved accessibility in digital applications, which will be beneficial for all. Speech technology should thus be developed with this type of issues in mind in order to allow creation of successful communication solutions and effective applications for the daily use.

Acknowledgments Projects MULTIVOX and SPEECH-AID were developed with collaboration of LSS’s member João Paulo Teixeira from the Polytechnic Institute of Bragança, Portugal, in cooperation with the Technical University of Budapest and the Academy of Science of Budapest and the colleagues Geza Nemeth and Gabor Olaszy.

Project Audiobrowser was developed with collaboration of Fernando Lopes, in consortium with the Department of Informatics (DI) of the University of Minho, and the colleague António Fernandes.

Projects INFOMETRO and NAVMETRO are developed by a consortium formed by ACAPO, Metro do Porto, S. A. and FEUP and is sponsored by the Portuguese Government Program POSConhecimento.

Project SPEAKMATH is sponsored by the Portuguese Government Program, POSConhecimento.

Project AQN was developed by Helder Ferreira and Vitor Carvalho (both from FEUP) with collaboration of Dárida Fernandes and Fernando Pedrosa (both from ESE-IPP, Centro Calculus) and was coordinated by the author. It was started after an idea from the former LSS’s member, Maria Barros.

The author was a member of Action COST 219 ter – Accessibility for All to Services and Terminals for Next Generation Networks. COST 219 ter was a forum that allowed the author for many years to learn about the “whys” and “hows” of design for all in accessibility to telecommunications.

Pedro Marcolino Freitas, the elder of the author’s sons, to whom deep gratitude is expressed, designed the chapter’s pictures.

References

1. Raman, T. V. (1997). *Auditory User Interfaces*. kluwer, Dordrecht.
2. Raman, T. V. (1998). Conversational gestures for direct manipulation on the audio desktop., In: Proc. 3rd Int. ACM SIGACCESS Conf. on Assistive Technologies: Marina del Rey, CA, 51–58.
3. Potamianos, A., Narayanan, S. S. (2003). Robust recognition of children’s speech. In: *IEEE Transactions on Speech and Audio Processing* 11(6), 603–616.
4. Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech Signal Process.*
5. Junqua, J. C., Haton, J. P. (1996). *Robutness in Automatic Speech Recognition – Fundamentals and Applications*. Kluwer Academic Publishers, Dordrecht.
6. Gales, M., Young, S. (1996). Robust Continuous Speech Recognition using Parallel Model Combination. In: *IEEE Trans Speech and Audio Processing* 4(5), 352–359.
7. Moura, A., Pêra, V., Freitas, D. (2006). An automatic speech recognition system for persons with disability. In: Proc. Conf. IBERDISCAP’06: Vitória-ES, Brasil, 20–22.
8. Roe, P. (ed) (2007). *Towards an Inclusive Future – Impact and Wider Potential of Information and Communication Technologies*. COST, European Commission.

9. Neti, C., Potamianos, G., Luettin, J., Mattheus, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., Zhou, J. (2000). Audio-visual Speech Recognition. Final Workshop 2000 report. Baltimore MD Centre for Language and Speech Processing, The Johns Hopkins University.
10. Turunen, M., Hakulinen, J., Rähkä, K.-J., Salonen, E.-P., Kainulainen, A., Prusi, P. (2005). An architecture and applications for speech-based accessibility systems, In: IBM Systems Journal, 44(3), 485–504.
11. Jokinen, K., Kerminen, A., Kaipainen, M., Jauhiainen, T., Wilcock, G., Turunen, M., Hakulinen, J., Kuusisto, J., Lagus, K. (2002). Adaptive dialogue systems – interaction with Interact. In: Jokinen, K., McRoy, S. (eds) Proc. 3rd SIGdial Workshop on Discourse and Dialogue, Philadelphia, 64–73.
12. Ferreira, H. (2005). Audiomath 2005, developed as part of the Graduation Thesis and MSc thesis, LSS, FEUP.
13. Freitas, D., Ferreira, H., Carvalho, V., Fernandes, D., Pedrosa, F. (2003). A prototype application for teaching numbers. Proc. 10th Int. Conf. on Human–Computer, HCII-2003, Crete, Greece.
14. Freitas, D., Ferreira, H., Fernandes, D. (2003). A. Q. N., *A Quinta dos Números, Um projecto em desenvolvimento*. Proc. “8o Baú da Matemática”: Ermesinde, Portuguese.
15. Roe, P. (ed) (2001). Bridging the Gap? COST219bis, European Commission.

Chapter 15

Assessment and Evaluation of Speech-Based Interactive Systems: From Manual Annotation to Automatic Usability Evaluation

Sebastian Möller

15.1 Introduction

Due to the improvements of speech and language technologies during the last few decades, the demand for assessment and evaluation of such technologies increased significantly. Starting from the assessment of individual system components such as automatic speech recognition (ASR) or text-to-speech (TTS) synthesis, evaluation methods are now required to address the system – and the service which is based on it – as a whole. Both individual component assessment and entire system evaluation are worth being considered here, depending on the question which shall be answered by the assessment or evaluation.

In the following, the term *assessment* will be used for the measurement of system (component) performance with respect to one or more criteria set by the assessor, and the term *evaluation* for the determination of the fitness of a system for a specific purpose by the prospective user of the system. This classification is similar to “performance evaluation” vs. “adequacy evaluation” proposed by Hirschman and Thompson [1]. Both types of evaluation may be analytical or utilitarian in nature, i.e., they may provide a diagnostic profile of system quality or performance (sometimes called “diagnostic evaluation”), or a global measure related to overall quality, usability, or acceptability.

Both assessment and evaluation are *measurement* processes in a larger sense: The aim is to obtain quantitative descriptions of system performance or of user-perceived quality. Such measurement processes differ from, e.g., physical ones in that the measurement instrument needs to be complemented by a measurement organ, namely a human test participant. In this sense, they are “subjective” measurements, in contrast to instrumental measurements, which rely only on a physical measurement instrument. The notion of “subjectivity,” however, does not imply that such methods would be less reliable or valid. On the opposite, valid and reliable quality measurements can only be obtained with the help of human test participants.

S. Möller (✉)

Deutsche Telekom Laboratories, Technische Universität, Berlin, Germany
e-mail: sebastian.moeller@telekom.de

Both subjective and instrumental measurements have to fulfill general requirements to measurements, such as

- validity (the method should be able to measure what is intended to measure),
- reliability (the method should be able to provide stable results across repeated administrations of the same measurement),
- objectivity (the method should reach inter-individual¹ agreement on the measurement results),
- sensitivity (the method should be able to measure small variations of what is intended to measure), and
- robustness (the method should be able to provide results independent from variables that are extraneous to the construct being measured).

In the next section, a brief overview of assessment and evaluation applied to speech-based interactive systems is provided, followed by a more thorough review of the terms “quality” and “performance.” Assessment principles for individual components of speech-based interactive systems are presented in Section 15.2 and evaluation principles for entire systems in Section 15.3. Section 15.4 describes common approaches for predicting quality judgments on the basis of performance metrics. The overview in these sections will be mainly limited to task-oriented and speech-based interactive systems; other system types and novel principles for semi-automatic quality assessment, evaluation, and prediction will be presented in Section 15.5.

15.2 A Brief History of Assessment and Evaluation

Following the development of speech and natural language processing systems, first assessment and evaluation efforts were directed toward individual system components (mainly speech recognizers, speech and natural language understanding modules, and speech synthesizers), or toward systems with a limited functionality (such as simple question–answering systems). Putting these modules together, the evaluation expanded toward entire systems, frequently addressing one specific task (such as flight information).

In the United States, competitively organized ARPA/DARPA assessment campaigns were established since the late 1980s. In Europe, efforts were organized in collaborative, but partly also large-scale projects like SAM (Multi-Lingual Speech Input/Output Assessment, Methodology and Standardization; ESPRIT projects 2589 and 6819), EAGLES (Expert Advisory Group on Language Engineering Standards), Aupelf-Uref [2], SQUALE [3, 4], Class [5], or DISC [6].

¹With respect to the test administration, not with respect to the test participants whose individual perception and judgment processes may differ.

The results of these campaigns and projects were made available to researchers and practitioners in the form of guidelines or de facto standards. The first de facto standard was the so-called EAGLES Handbook published in 1997 [7], which includes a chapter on the evaluation of speech-based interactive systems [8], as well as chapters on the assessment of speech recognition [9], speaker verification [10], and speech synthesis [11]. A second part of this handbook was issued in 2000 [12] and focused on multimodal systems and consumer off-the-shelf product and service evaluation. In the telecommunication context, the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T) established two recommendations and one supplement for the assessment of speech output components [13], the subjective evaluation of speech-based interactive systems [14], and the parametric quantification of interactions with such systems [15]. These standards will be briefly reviewed in Sections 15.4.2 and 15.4.3.

15.2.1 Performance and Quality

It is the task of assessment and evaluation to determine in how far a service and the system it relies on fulfill the user's or the system designer's requirements. Two terms are commonly used for this purpose: *performance* and *quality*.

When it comes to individual system components, it is more appropriate to speak about *performance*, i.e., the ability of the module to provide the function it has been designed for. Performance can easily be quantified when a measure² exists which is closely linked to the function under consideration. For example, the function of a speech recognizer is to transcribe a spoken utterance into a sequence of words, and a quantifiable measure associated with this is, e.g., the word error rate. An indication of performance is less viable when the function is not completely apparent, or when there is no "natural" measure linked to this function. Consider, e.g., a TTS synthesizer: Its function is to provide a spoken utterance from a sequence of written words; however, the success in providing this function may be linked to different measures such as segmental intelligibility, comprehensibility, listening effort, naturalness (whether it is human-like or not), expressivity, and adequacy for the given task.

For an interactive system as a whole, it is also not straightforward to specify one precise function. Although the performance of a train timetable information system may primarily be specified in terms of its effectiveness for the task (i.e., whether it provides the desired information), there are other important functions and corresponding measures, such as its efficiency (measured, e.g., via the time needed for task completion or via the effort required from the user), the satisfaction of the user (related to the experienced comfort, the pleasantness, or joy-of-use), its utility (involving cost measures), as well as its acceptability (i.e., the question whether a potential user would be willing to use the system). Measures for most of

² The number or category assigned to an attribute of an entity by making a measurement, see [30].

these quality aspects can only be defined on the basis of subjective experiments with human test participants.

The *quality* of a speech-based interactive system is determined by the perceptions of its users. It results from a perception and a judgment process, in which the perceiving subject (e.g., a test user of the system) establishes a relationship between the perceptive event, and what he/she expects or desires from the service. This result can also be called a “quality event,” because it is determined in space, time, and character (as all events are). Such a user-centric point of view is reflected in the definition of quality given by Jekosch [16, 17]:

Result of judgment of the perceived composition of an entity with respect to its desired composition. [17, p. 15]

As it is the user who judges on quality, user factors like attitude, emotions, experience, and task/domain knowledge will influence the perception of quality.

An important aspect of the quality of a speech-based interactive system is its *usability*. Usability is defined in [18] as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.” It can be measured on the three “layers” effectiveness (i.e., the accuracy and completeness with which specified users can achieve specified goals in specified environments), efficiency (i.e., the resources expended in relation to the accuracy and completeness of the goals achieved), and satisfaction (comfort and acceptability of the system to its users and other people affected by its use).

In order to differentiate between the multitude of quality aspects, a taxonomy has been set up [19, 20], see Fig. 15.1. In its upper part, the figure shows the factors of the system and of the context of use, which exercise an influence on quality (*quality factors*). In its lower part, it shows the *quality aspects*, i.e., categories of quality, and their composing perceptual dimensions, *quality features*, from a user’s point of view. The taxonomy will not be further discussed here, and the interested reader is referred to an extended description in [19]. Nevertheless, the picture reveals the multi-dimensionality of quality, both from the system developer’s and from the user’s perspective. These dimensions have to be taken into account in the assessment and evaluation process.

15.3 Assessment of Speech-System Components

There is no universally agreed architecture for speech-based interactive systems. Most systems contain components for speech recognition, natural language understanding, dialog management, and speech output. Additional (optional) components include telephone interfaces, speaker identification or verification, database access, response generation, and speech synthesis. In the following paragraphs, a brief review of assessment principles will be presented for speech recognition, for the interpretation of semantic units on the basis of signals (speech understanding) or

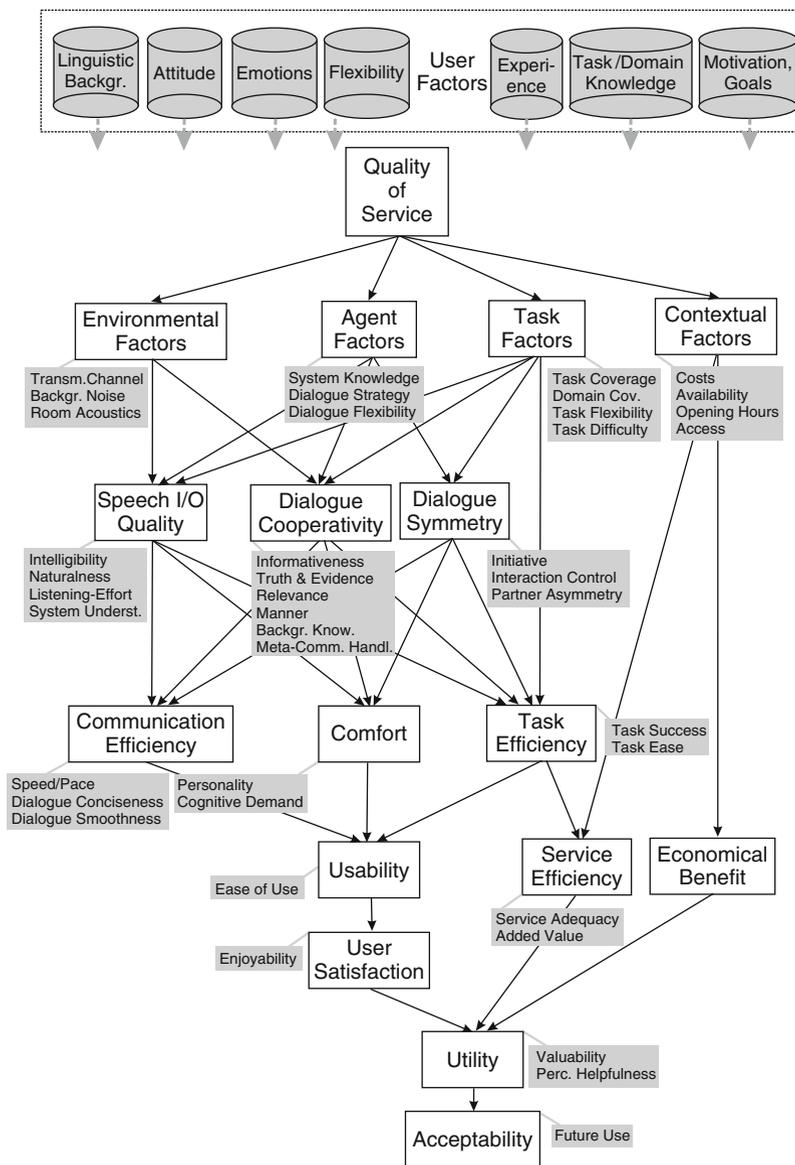


Fig. 15.1 Taxonomy of quality aspects for speech-based interactive systems [19, 20]. Depicted are quality factors and quality aspects indicated by the *white boxes*, as well as their composing dimensions quality elements and quality features (*gray boxes*), and their mutual relationships (*arrows*). As the user is the decision point for each quality aspect, user factors are distributed over the whole diagram. This fact has tentatively been illustrated in the cover panel (*gray cans*)

symbols (natural language understanding), for dialog management, and for the output of spoken language (including response generation and speech generation). A more thorough overview can be found, e.g., in the EAGLES Handbook [7] or in [19].

15.3.1 Assessment of Speech Recognition

Although speech recognition assessment seems to be quite straightforward – namely to compare the hypothesized output of the recognizer to a reference output produced by a human expert – there are a number of factors complicating this task [21], namely, speech recognition assessment results differ with respect to the type and linguistic categories of the input speech, the speakers, the task or domain the recognizer is designed for, as well as the usage environment. For benchmark assessment campaigns, these conditions have to be held constant and controllable, cf. the DARPA evaluation campaigns [22–24], or the European SQALE project [3, 4].

In order to compare the word string hypothesis of a continuous speech recognizer and the reference transcription, hypothesis and reference first have to be aligned, e.g., using a dynamic programming matching algorithm which assigns different penalties to deleted, substituted, or inserted words, see [25, 26]. On the basis of the alignment, the number of correctly identified, substituted, deleted, and inserted words is counted, and word error rate or word accuracy is determined. For an isolated word recognizer, the number of insertions is replaced by a false alarm rate. Formulae for these metrics can be found, e.g., in [9].

Similar metrics can also be determined on a sentence level, resulting in sentence error rates or sentence accuracies. Alternative metrics on a sentence level include the average number of errors per sentence, and the word error per sentence, see [27, 28]. Usually the sentence accuracy is lower than the word accuracy, because a single misrecognized word may impact an entire sentence. However, many recognition errors do not really impact the dialog flow; they may be ruled out by applying robust partial parsing techniques or they may be completely ignored in case that no “keyword” necessary for the semantic interpretation of the system is affected. Such cases are better taken into account by the understanding-related metrics discussed hereafter.

15.3.2 Assessment of Speech and Natural Language Understanding

The task of the “understanding” component is to extract the semantic content of the user utterance, either directly from the speech signal (so-called “speech understanding”) or from a previously transcribed utterance (“natural language understanding”). The assessment thus consists of determining the accuracy of the extracted semantic

content. Depending on the extraction task and the domain covered by the system, different principles have been applied.

In the DARPA ATIS (Air Travel Information System) domain, understanding accuracy has been assessed by monitoring the system's ability to generate appropriate database queries from spoken input questions [29]. User utterances first had to be classified as to whether they are interpretable without additional knowledge, and then described in terms or reference database answers. From a comparison between the system's and the reference database answer, specific scores called the DARPA score and the DARPA weighted error were calculated. This approach was limited in that many user utterances were not interpretable per se, but only in the light of the dialog history which was not taken into account.

A better approach for speech-based interactive systems is to directly compare the semantic concepts, which can be extracted from the user utterance. Most frequently, the semantic concepts can be formalized in terms of attribute-value pairs (AVPs). Similar to the approach taken for speech recognition scoring, semantic concepts which have been substituted, deleted, or inserted are counted, and then a concept accuracy or concept error rate is calculated. Alternatively, each user utterance can be labeled as to whether it has been correctly, partially correctly, or incorrectly parsed, and then an understanding accuracy (in relation to the total number of user utterances) is calculated. In addition to these metrics, Glass et al. [30] proposed two new ones on the dialog level, namely the query density (the rate of new concepts understood by the system per dialog) and the concept accuracy (the average number of turns which is necessary for each concept to be understood by the system). To the author's knowledge, these two metrics have only rarely been used in the assessment of speech-based interactive systems so far.

15.3.3 Assessment of Dialog Management

Whereas the tasks of speech recognition and understanding are clear and relatively well-defined, dialog management requires a large number of ill-defined and partially contradicting functions to be covered. Core functions of a dialog manager include the collection of all information from the user which is necessary for accomplishing the task, the distribution of dialog initiative, the management of knowledge sources involved in the dialog, the provision of feedback and verification of information, the provision of help, the correction of misunderstandings, the interpretation of complex phenomena like ellipses or anaphora, and the organization of information output to the user. The performance on such a multitude of functions cannot easily be quantified.

The most widely used metrics which refer mainly to the dialog manager quantify its meta-communication capabilities. Based on an annotation of the dialog flow by a human assessor, the number or percentage of help messages, of time-out prompts, of ASR rejections, or of other diagnostic error messages from the system can be measured. From the user's perspective, parameters related to

meta-communication include the number of help requests and of cancel attempts. In a more general way, meta-communication may also be quantified by counting the number or percentage of system and user turns which are primarily concerned with rectifying some kind of “trouble.” The corresponding metrics are known as the system or user correction rates. The system’s ability to recover from instances of misunderstanding can be quantified by means of the “implicit recovery” parameter, counting the number of appropriate system’s answers in case of concept errors.

It has to be emphasized that the system’s meta-communication behavior is not determined by the dialog manager alone. On the contrary, it is also affected by speech recognition and understanding components, as well as by response generation.

15.3.4 Assessment of Speech Output

In order to maintain a spoken interaction, a natural language response and a corresponding speech signal need to be generated by the system. For assessing the natural language response, prompt design guidelines have been set up, see, e.g. [8, p. 592]. A quantitative description of the success in applying these guidelines is the “contextual appropriateness” parameter. It is based on Grice’s maxims for cooperative behavior in human-to-human communication scenarios [31]. These maxims have been extended for interactions with speech-based interactive systems by Bernsen et al. [32]. Following these maxims, the appropriateness of a system response in the light of the immediate dialog context can be rated by an external human assessor. In this approach, the system is mainly regarded as a black box, and the system response will be influenced by a number of system components.

The focus of speech output assessment is currently mainly on the (naturally produced or synthesized) speech signal itself, thus on the speech generation part. A number of auditory test methods have been developed which focus on different aspects of the speech signal, such as articulation, intelligibility, comprehensibility, naturalness, pleasantness, global acceptance in a specific application scenario. An overview of such methods can be found, e.g., in [11, 17, 33].

In such auditory tests, participants are usually asked to listen to test stimuli which are somehow “typical” for the application context, but which may differ from this with respect to their length and complexity, linguistic level, meaningfulness, and linguistic representativeness. Test participants have to either identify/verify what they have heard, using an open or closed set of answers, or to judge different aspects of what they heard, mostly through questionnaires or interviews. A good example of such a test is the procedure recommended in ITU-T Rec. P.85 [13] for assessing the speech output component of a telephone-based system.

Participants have to listen to exemplary samples, reproduce some information contained in this sample, and afterwards judge upon different quality aspects on a set of 5-point category-rating scales (acceptance, overall impression, listening effort,

comprehension problems, articulation, pronunciation, speaking rate, voice pleasantness). This method has partially been criticized [34, 35], but is still frequently used, as it is quick and provides still reliable and meaningful results.

15.4 Evaluation of Entire Systems

Figures on the performance of individual system components are useful optimization criteria for the developer of such components, and they support the implementation of an interactive system by providing criteria to select one component out of a range of similar ones offered by the market. However, they may be bad indicators for the quality and usability of the entire system, because the performance of each system component depends on its context of use. Thus, a glass-box approach considering the individual components should be complemented by black-box testing of the entire system in a typical application setting.

Such a black-box evaluation requires (test) interactions between the user and the system to take place. The interactions are commonly performed in the laboratory to guarantee controlled experimental conditions and to obtain quantitative data from the test participants which otherwise would be difficult to get. In particular, questionnaires, interviews, or heuristic evaluations can much better be carried out in a laboratory setting. On the other hand, field tests involve a more realistic setting for the test participants. The contextual factors are more similar to the ones of the later service, and as a result, the behavior of and the judgments obtained from the test participants may be more realistic (ecological validity), and more valid with respect to the evaluation task (see Section 15.1).

Test interactions are usually incited by providing usage scenarios to the participants. In this way, a more-or-less meaningful purpose of the interaction can be achieved. However, such a given purpose is not identical to a real purpose a user might have in using the system. This frequently results in effectiveness, e.g., in terms of task success, playing a minor role in the evaluation process.

The interactions are usually logged, and the audio, video, and/or textual log files can be used to identify and classify interaction problems. A record of observed interaction problems is useful for the developer to identify weak parts of the system or unpredicted behavior from the user which should be catered for by the system. Interaction problems are further discussed in Section 15.4.1. From the log files, a number of parameters can be extracted which provide quantitative descriptions of the performance of system components as well as of the behavior of the system and the user. Such interaction parameters are briefly reviewed in Section 15.4.2, and an in-depth discussion can be found in [19]. After each test interaction, qualitative and quantitative judgements can be obtained from the test participants via questionnaires or interviews, see Section 15.4.3.

Apart from such formal evaluation experiments, usability inspection methods can be applied to obtain less formal – but often more varied – information on

system behavior and user problems. These methods are briefly summarized in Section 15.4.4, and the interested reader is referred to Nielsen and Mack [36] for an in-depth description.

15.4.1 Detection and Classification of Interaction Problems

The logged interactions are transcribed and annotated by a human expert in order to detect and classify interaction problems. The annotation should be formalized as far as possible to obtain quantifiable data.

Bernsen et al. [32] describe a classification of cases where the user does not perform in accordance with the “normative model” provided by the system. Such cases are frequently called “user errors,” although no fault can be attributed to the user – it is in fact the system which has not been designed to adequately respect the user’s behavior! According to this classification, “user errors” can be distinguished, e.g., in terms of ignoring clear system feedback, responding to a question different from the one asked by the system, answering several questions at a time, asking unattended questions, changing a request through comments, or thinking aloud.

A different classification scheme for “user errors” has been proposed by Oulasvirta et al. [37]. Following their approach, “user errors” can occur on a goal level (e.g., limitations of the system’s capability), a task level (e.g., the user issuing a command which is not valid in the present state of the dialog, but which would be valid in a different state), a command level (e.g., vocabulary and grammar errors), or a modeling level (e.g., the user issuing a command which would require the “world” of the system to be represented in a different way). This classification of errors is “phenotypical,” as it refers to the surface form of the interaction, and not to assumed causes of interaction problems.

A more “genotypic” classification of interaction problems comes from general usability analyses following, e.g., ISO Standard 9241 Part 110 [38]. Following this standard, interaction logs may be labeled for so-called critical incidents. The critical incidents can be linked to the violation of dialog principles, such as the suitability of the system for the task, the self-descriptiveness of the system, its controllability, its conformity with user expectations, its error tolerance, its suitability for individualization, as well as its suitability for learning. Annotation and classification according to this scheme often requires some additional involvement of the user, e.g., in terms of intermitting or post-experimental thinking-aloud.

Another “genotypic” classification of interaction problems consists in identifying the system component, which is responsible for the observed problem. Manual annotation along this line of thinking is facilitated, e.g., by the “fishbone” diagram [39], which has been adopted from the analysis of safety-critical systems: Following the “bones” of the diagram, an interaction problem is classified according to the assumed source of the problem (recognition, understanding, dialog, system output, task, and system failure), distinguishing in each case between different sub-problems (for the recognition, e.g., language model, dictionary, user accent, or background noise).

15.4.2 Parametric Description of Interactions

Interaction parameters support system developers in providing quantitative data for system development and optimization. They quantify the flow of the interaction, the behavior of the user and the system, and the performance of the speech technology devices involved in the interaction. Interaction parameters address system performance from a system developer's and/or service operator's point of view, and thus provide complementary information to subjective evaluation data.

For extracting such parameters, interaction experiments have to be carried out. As described before, the interactions are logged, and from the log files, parameters can be calculated either instrumentally or with the help of a transcribing and annotating expert. Parameters which relate to the surface form of the utterances exchanged between user and system, like the duration of the interaction or the number of turns, can usually be measured fully instrumentally. In turn, human transcription and annotation is needed when not only the surface form (speech signals), but also the contents and meaning of system or user utterances (e.g., to determine a word or concept accuracy) are addressed. Both (instrumental and expert-based) ways of collecting interaction parameters should be combined in order to obtain as much information as possible.

Based on a broad literature survey, a large number of parameters were identified which have been used in different assessment and evaluation experiments during the past 15 years, see [41]. The respective literature includes [8, 9, 27, 28, 30, 40–56]. The parameters can broadly be classified into

- dialog- and communication-related parameters,
- meta-communication-related parameters,
- co-operativity-related parameters,
- task-related parameters, and
- speech-input-related parameters.

Definitions of each of these parameters can also be found in [19]. The parameters have recently been recommended for the evaluation of telephone-based interactive systems, see ITU-T Suppl. 24 to P-Series Rec. [15].

15.4.3 Subjective Quality Evaluation

Following the definition of the term “quality” given in Section 15.2.1, measurements of quality have to rely on subjective judgments given by human users. The judgments are commonly collected in a quantifiable form, e.g., on a questionnaire with a number of rating scales. Questionnaires are distributed to a group of test participants both before the first interaction with the system to obtain unbiased information on the users' background and expectations, and after some interaction experience to reflect the current impression of using the system.

A questionnaire which can be distributed to test participants directly after an interaction with the system was developed by Hone and Graham [57, 58], called “SASSI” (Subjective Assessment of Speech System Interfaces). It has been designed on the basis of subjective experiments with eight different systems, all showing speech input capability, and some also speech output capability. The questionnaire contains 44 declarative statements (e.g., “the system is easy to use”) with which respondents rate their agreement on 7-point Likert scales. A factor analysis of judgments from 214 questionnaires revealed six underlying perceptive dimensions which were termed “system response accuracy,” “likeability,” “cognitive demand,” “annoyance,” “habitability,” and “speed” [57]. Some of these dimensions correspond to quality aspects or sub-aspects in Fig. 15.1 (e.g., cognitive demand or speed), while others express more general user perceptions (likeability, annoyance) which may be somehow related to user satisfaction. It has to be emphasized that the questionnaire has been developed for systems with speech input capability; judgments related to speech output quality have been removed during the design process, in order to retain only judgments which are relevant for all systems under test.

A different – but partly overlapping – list of questions is proposed in ITU-T Rec. P.851 [14] for the evaluation of speech-based telephone services. This recommendation distinguishes between three types of questionnaires: (1) questionnaires collecting information on the user’s background, and distributed at the beginning of an evaluation experiment; (2) questionnaires with questions related to individual interactions with the system under test; and (3) questionnaires related to the user’s overall impression of the system, to be answered after a number of interactions with the system (e.g., at the end of an experiment). For each type of questionnaire, an open list of topics is proposed; the topics then have to be translated into precise questions or statements according to the purpose of the evaluation and the system/service under test. Exemplary questions and statements are provided which are rated on 5-point Likert scales or on continuous rating scales. In addition, general guidelines are given for the experimental set-up, the test scenarios, as well as the selection of test participants.

15.4.4 Usability Inspection

In addition to the controlled user interaction tests described above, usability inspection methods enable developers to detect usability problems which would be overlooked in standard user testing. The aim of usability inspection methods is to find usability problems in an existing user interface design, to rate the severity of problems, and to make recommendations on how to improve the system design.

A good overview of usability inspection methods – not necessarily applied to speech-based systems – can be found, e.g., in [36]. They include

- heuristic evaluation (usability experts judge whether a dialog element conforms to established usability principles),

- guideline reviews,
- pluralistic walkthroughs (meetings where users, developers, and human factors specialists step through a given scenario, discussing usability issues for each dialog step),
- consistency inspections,
- standards inspections,
- cognitive walkthroughs (simulation of a user’s problem-solving process at each step, and checking whether the user’s goal and action memory can be expected to lead to the next correct step or not),
- formal usability inspections (formalized meetings involving a usability expert team), and
- feature inspections (analysis of operational system functions).

15.5 Prediction of Quality Judgments

In the preceding section, two types of metrics have been presented: (1) User-internal metrics directly obtained from the user, quantifying *quality* aspects, and (2) user-external metrics, determined with the help of a measuring instrument and/or a human assessor, quantifying the *performance* of the system (components) and of the user in the interaction, namely in terms of interaction parameters. The question arises whether it is possible to relate these metrics to each other, and potentially to predict user judgments on the basis of interaction parameters.

Correlations between user judgments and interaction parameters have been documented in literature, see, e.g. [19, pp. 278–279], for an analysis of a restaurant information system or [59] for a smart-home system. The results show that correlations are mostly weak, usually not higher than 0.4. Surprisingly low values have been observed, e.g., between the user-perceived length of an interaction and its measured length or between the user’s perception of task success and a corresponding expert-labeling [60]. These findings may be interpreted in that both types of metrics provide data from two complementary points of view: The one of the system developer and the one of the user. They underline that it is necessary to assess and evaluate systems according to both principles.

Nevertheless, models have been developed to estimate quality on the basis of interaction parameters. The most popular approach is the PARADISE framework proposed by Walker et al. [54]. It combines a set of input parameters (interaction parameters) to predict a target variable called “user satisfaction,” which is an estimate of the arithmetic mean over several user quality judgments. The algorithm supposes a linear superposition of the (normalized) input parameters in the following way:

$$US_w = \alpha N(\kappa) - \sum_{i=1}^n w_i N(c_i), \quad (15.1)$$

with κ an interaction parameter or a subjective judgment related to task success (i.e., whether the user achieved his/her task goals), c_i a set of further interaction parameters, and N the z -score normalization function. α and w_i are weighting coefficients which can be determined on the basis of a controlled laboratory experiment, collecting both user judgments and interaction parameters, by using a multivariate linear regression analysis. Once the coefficients and the relevant interaction parameters have been determined, Eq. 15.1 can be used to predict “user satisfaction” for other interactions and systems, without directly asking the user any more.

Although the PARADISE model is very helpful in the system design cycle, its predictive power is relatively limited. Usually, about 40–60% of the variance of user judgments used for training can be covered by the linear model. Only few examples have been reported where a model derived from one specific system and user group has been tested on a different system and/or user group (see [61] for an example). In most of these investigations, the extrapolation was limited to telephone-based systems developed at the same laboratory, presumably with similar system components. Analyses described in [62] showed that a cross-user extrapolation significantly reduced the prediction accuracy and that a cross-system extrapolation did not provide any meaningful results. In conclusion, approaches to predict user judgments on the basis of interaction parameters alone should be considered with care, as their predictive power still seems to be quite limited.

15.6 Conclusions and Future Trends

In the previous sections, the most important and most widely applied principles for assessing and evaluating speech-based interactive systems have been summarized. Assessment methods have been pointed out for the standard components of most dialog systems, including speech recognition, natural language understanding, dialog management, and speech output. These methods allow the performance of the respective modules to be quantified. However, the obtained performance measures are not necessarily linked to a good overall system performance, or to a high quality perceived by the user.

Evaluation of the entire system requires interaction experiments in which test users have to rate the quality (and its composing sub-aspects) in a subjective way. In parallel, interaction parameters can be collected, providing quantitative data of system (component) performance, and of user and system behavior. The log files can also be annotated for interaction problems, according to different annotation standards. Usability inspection methods allow complementary problems to be identified and system improvements to be proposed.

The methods described so far are limited in several ways. First, they mainly cover static task-oriented systems where speech is the only interaction modality. Although this class is by far the most important one from a commercial point of view, new multimodal, adaptive or non-task-oriented systems come up which

require the assessment and evaluation methods at best to be extended or at worst to be completely redefined [63].

Second, even the available methods are not as frequently applied as they should be, mainly because of time and money constraints which prevent subjective tests to be carried out before a new system gets operational. In order to not endanger the acceptance of interactive systems as a whole, the available assessment and evaluation methods should be complemented with automatic or semi-automatic ones, which allow system weaknesses and usability problems to be detected early in the design process, and design improvements to be made quickly and at low costs, without requiring human test participants at each stage. Such methods should be able to provide valid and reliable estimations of user-perceived quality. For this purpose, it is necessary to improve the quality prediction approaches discussed above. Research ideas and trends in these three directions will be discussed in the following paragraphs.

15.6.1 Multimodal, Adaptive, and Non-task-Oriented Systems

Interactive systems which allow several input and output modalities to be used – either alternatively or in conjunction with each other – will have additional modules for recognition and interpretation (e.g., gesture, pointing devices), for the fusion of modalities, for the distribution of output to different modalities, as well as for the display of information (e.g., a talking head). The performance of these additional modules has to be quantified, but this is not an easy task, in particular when several modalities are used conjointly.

Interaction data with multimodal systems have to be collected in realistic environments and usage scenarios. The physical characteristics of the environment may play a significant role, e.g., with respect to its acoustic and illumination conditions. Multimodal systems will frequently be used in mobile scenarios, where the user is moving (requiring robust sensing and putting restrictions on the weight and power consumption of devices), and may be confronted with other (parallel) tasks, which require cognitive resources. Examples include car navigation systems or applications running on mobile personal digital assistants (PDAs). Under these circumstances it is difficult to obtain constant and controlled test conditions.

Provided that data on the behavior of the user with respect to the additional devices as well as on the performance of the supplementary modules are available, it is still unclear how such data should be annotated and transformed into quantitative interaction parameters. For additional input modules, it may be possible to define error rates as it is done for speech recognition and understanding. However, the options offered by different modalities may result in a large variability of the behavior of users, and even within a certain user. As a consequence, dialog- and communication-related measures may become meaningless.

It may be difficult to define the effectiveness and efficiency of a speech-based or multimodal interactive system. Beringer et al. [64] reported on the evaluation of

a smart-home assistant and found difficulties to code the success of film-selection tasks related to a program guide. As the approach and the number of steps necessary to accomplish a given task vary widely, it may be meaningless to associate a simple task success or efficiency measure.

Effectiveness and efficiency measures are also problematic measurement objects when it comes to non-task-oriented dialog systems, like the Hans-Christian-Andersen “edutainment” system described in [65]. Traditional notions of task success and interaction duration seem to be meaningless for such systems. However, there are indirect tasks related to an “edutainment” system, like education and entertainment. Corresponding metrics may thus aim at quantifying education success (similar to the ones applied in tutoring dialog systems) or joy-of-use. For the latter, it is not yet clear how it can be measured subjectively. Possible solutions include physiological data, like facial muscle activity, skin conductance, and so on.

A further challenge is the assessment and evaluation of non-static systems. For example, a dialog system may adapt to the behavior of the user, by providing tutorial guidance to newcomers and more efficient interaction capabilities to frequent users. Although learning and adaptation behavior on the part of the user has been observed and quantified in terms of interaction parameters, adaptation on the part of the system will reduce the frequency of occurrence of specific phenomena and thus will reduce the reliability of conclusions. Many users may not expect their system to adapt, so they may be confused by (from their point-of-view) inconsistent system behavior. The systems may also provide the possibility of personalization incited by the user. Evaluation of such features usually requires a long-term exposure to the system under test. Data collected in this way may then be very specific to the usage situation.

15.6.2 Semi-automatic Evaluation

It has already been stated that the evaluation of speech-based interactive systems suffers from the expenses of subjective testing. As a result, efforts have been made to automatize at least part of the evaluation process. Evident ways include the support of transcription and annotation via specifically designed tools.

The next step is to replace the human user in the human–system interaction loop by some type of simulation. For example, Araki and Doshita [66] propose a system-to-system evaluation via a mediator program. The human user is replaced by a system with similar characteristics of the system to be evaluated, and the mediator introduces acoustic noise in the communication channel, simulating speech recognition errors. In this way, the system’s robustness against recognition errors can be assessed, and its ability to repair or manage such misrecognized sentences with the help of a robust linguistic processor. López-Cozar et al. [67] propose a rule-based “user simulator” which feeds the dialog system under test. It generates user prompts from a corpus of utterances previously collected in a human–human interaction paradigm, and re-recorded by a number of speakers.

On a language level, Walker [68] reports on experiments with two simulated agents carrying out a room-design task. Agents are modeled with scalable attention/working memory, and their communicative strategies can be selected according to the desired communication style. In this way, the effect of task, communication strategy and of cognitive demand can be analyzed separately, without direct involvement of human test participants. Similar experiments have been described by Walker [69] with agents which can be parameterized according to their communicative and error recovery strategies.

A recent step toward a semi-automatic evaluation of interactive systems is the MeMo workbench described in [70]. In this workbench, the user is simulated on the basis of an explicit task and interaction model, which is partially derived from the corresponding system task and interaction model. A direct adaptation of the system model would result in a kind of “optimal” simulated user, who provides the system-desired information at each step of the dialog. However, users behave differently, because of misconceptions of the system and its capabilities. In order to generate such “erroneous” behavior, deviations from the ideal user model are generated, using the error classification described in [37], see Section 15.4.1. The system and the user model provide input to an automatic testing unit, which generates an arbitrary number of simulated dialogs, potentially for different groups of users. The simulated interactions are logged, and on the basis of the log files a usability profile is provided, making use of quality prediction algorithms like PARADISE.

15.6.3 Quality Prediction

The limitations of quality prediction algorithms for speech-based interactive systems have already been pointed out. There are two reasons assumed to be responsible. First, the interaction parameters do not seem to measure the “right” type of information, i.e., the information which is relevant for quality from a user’s perspective. This assumption is supported by the low correlation values between interaction parameters and user judgments; see Section 15.5. Second, the linear superposition of interaction parameters in the PARADISE framework (see Eq. 15.1) is most likely too simplistic.

In order to overcome the first limitation, additional interaction parameters have to be extracted. It is preferable that such interaction parameters be instrumentally measurable, so that the resulting model can be applied without a tedious and time-consuming annotation process. As an example, simple analyses of the speech signals exchanged between user and system may be carried out. More sophisticated measures include an automatic quality estimation of the system’s speech signal, e.g., using single-ended quality prediction approaches recently developed for monitoring telephone quality. First results reported in [71] show correlations of up to 0.74 on specific databases, without a prior optimization of the models for the given task. Further improvements may be possible, leading to an adequate prediction performance. Such estimated quality indices provide additional information as an input to

PARADISE-style models; they are, however, limited to the surface level, and cannot extract the semantics of the system output.

The second limitation may be overcome by applying more complex modeling algorithms. A first idea would be a non-linear regression, but the exact shape of a non-linear function is not apparent from empirical data. A second idea would be to use automatic classification algorithms, like classification or regression trees, or neural networks. Results reported in [72] show that the prediction accuracy can be slightly increased compared to PARADISE; the major advantage, however, seems to be that such models are more generic, in that they allow for predictions for unknown users and systems, with a stable (but still low) level of reliability. Generic nature is a prerequisite for meaningful quality prediction.

Further improvements may be possible by including temporal context in the prediction models. Instead of using one input vector of interaction parameters for each dialog, it may be possible to apply a sequence of feature vectors, one for each exchange (user–system utterance pair). Potential model candidates are neural networks or Hidden-Markov Models, which include a description of the temporal sequence of an interaction.

Acknowledgments The work described in this chapter has partially been carried out at the Institute of Communication Acoustics, Ruhr-University Bochum, and partially at Deutsche Telekom Laboratories, Technische Universität Berlin. The author would like to thank all colleagues and students who contributed to the mentioned work, as well as Robert Schleicher and Klaus-Peter Engelbrecht for their comments on an earlier version of the chapter.

References

1. Hirschman L., Thompson, H. (1997). Overview of evaluation in speech and natural language processing. In: *Survey of the State of the Art in Human Language Technology*, Cambridge University Press and Giardini Editori, Pisa, 409–414.
2. Mariani, J. (2002). The Aupelf-Uref evaluation-based language engineering actions and related projects. In: *Proc. 1st Int. Conf. on Language Resources and Evaluation (LREC'98)*, Granada, 123–128.
3. Steeneken, H., van Leeuwen, D. (1995). Multi-lingual assessment of speaker independent large vocabulary speech-recognition systems: The SQALE-project. In: *Proc. 4th Eur. Conf. on Speech Communication and Technology (EUROSPEECH'95)*, Madrid, 1271–1274.
4. Young, S., Adda-Decker, M., Aubert, X., Dugast, C., Gauvain, J., Kershaw, D., Lamel, L., Leeuwen, D., Pye, D., Robinson, A., Steeneken, H., Woodland, P. (1997). Multilingual large vocabulary speech recognition: The European SQALE project. *Comput. Speech Lang.* 11(1), 73–89.
5. Jacquemin, C., Mariani, J., Paroubek, P. (eds) (2005). Parameters describing the interaction with spoken dialogue systems using evaluation within HLT programs: Results and trends. In: *Proc. CLASS Pre-Conf. Workshop to LREC 2000*, Geneva, Athens.
6. Ernsen, N., Dybkjær, L. (1997). The DISC concerted action. In: *Proc. Speech and Language Technology (SALT) Club Workshop on Evaluation in Speech and Language Technology*, Sheffield, 35–42.
7. Gibbon, D., Moore, R., Winski, R. (eds) (1997). *Handbook on Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin.
8. Fraser, N. (1997). Assessment of Interactive Systems. *Handbook on Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin, 564–615.

9. Leeuwen, D., van Steeneken, H. (1997). Assessment of Recognition Systems. Handbook on Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin, 381–407.
10. Bimbot, F., Chollet, G. (1997). Assessment of Speaker Verification Systems. Handbook on Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin, 408–480.
11. van Bezooijen, R., van Heuven, V. (1997). Assessment of Synthesis Systems. Handbook on Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin, 481–563.
12. Gibbon, D., Mertins, I., Moore, R. (2000). Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation. Kluwer, Boston, MA.
13. ITU-T Recommendation P. 85 (1994). A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices. International Telecommunication Union, Geneva.
14. ITU-T Recommendation P. 851 (2003). Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. International Telecommunication Union, Geneva.
15. ITU-T Supplement 24 to P-Series Recommendations (2005). Parameters Describing the Interaction With Spoken Dialogue Systems. International Telecommunication Union, Geneva.
16. Jekosch, U. (2000). Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung. Habilitation thesis (unpublished), Universität/Gesamthochschule Essen.
17. Jekosch, U. (2005). Voice and Speech Quality Perception. Assessment and Evaluation. Springer, Berlin.
18. ISO 9241-11 (1998). Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 11: Guidance on Usability. International Organization for Standardization, Geneva.
19. Möller, S. (2005). Quality of Telephone-based Spoken Dialogue Systems. Springer, New York, NY.
20. Möller, S. (2002). A new taxonomy for the quality of telephone services based on spoken dialogue systems. In: Proc. 3rd SIGdial Workshop on Discourse and Dialogue. Philadelphia, PA, 142–153.
21. Pallett, D., Fourcin, A. (1997). Speech input: Assessment and evaluation. In: Survey of the State of the Art in Human Language Technology, Cambridge University Press and Giardini Editori, Pisa, 425–429.
22. Pallett, D., Fiscus, J., Fisher, W., Garofolo, J. (1993). Benchmark tests for the DARPA spoken language program. In: Proc. DARPA Human Language Technology Workshop, Princeton, NJ, 7–18.
23. Young, S. (1997). Speech recognition evaluation: A review of the ARPA CSR programme. In: Proc. Speech and Language Technology (SALT) Club Workshop on Evaluation in Speech and Language Technology, Sheffield, 197–205.
24. Pallett, D. (1998). The NIST role in automatic speech recognition benchmark tests. In: Proc. 1st Int. Conf. on Language Resources and Evaluation (LREC'98), Granada, 327–330.
25. Picone, J., Goudie-Marshall, K., Doddington, G., Fisher, W. (1986). Automatic text alignment for speech system evaluation. IEEE Trans. Acoust., Speech, Signal Process. 34(4), 780–784.
26. Picone, J., Doddington, G., Pallett, D. (1990). Phone-mediated word alignment for speech recognition evaluation. IEEE Trans. Acoust., Speech, Signal Process. 38(3), 559–562.
27. Strik, H., Cucchiari, C., Kessens, J. (2000). Comparing the recognition performance of CSRs: In search of an adequate metric and statistical significance test. In: Proc. 6th Int. Conf. on Spoken Language Processing (ICSLP2000), Beijing, 740–743.
28. Strik, H., Cucchiari, C., Kessens, J. (2001). Comparing the performance of two CSRs: How to determine the significance level of the differences. In: Proc. 7th Eur. Conf. on Speech Communication and Technology (EUROSPEECH 2001 – Scandinavia), Aalborg, 2091–2094.
29. Price, P. (1990). Evaluation of spoken language systems: The ATIS domain. In: Proc. DARPA Speech and Natural Language Workshop, Hidden Valley, PA, 91–95.

30. Glass, J., Polifroni, J., Seneff, S., Zue, V. (2000). Data collection and performance evaluation of spoken dialogue systems: The MIT experience. In: Proc. 6th Int. Conf. on Spoken Language Processing (ICSLP 2000), Beijing, 1–4.
31. Grice, H. (1975). *Logic and Conversation. Syntax and Semantics*. Academic, New York, NY, 41–58.
32. Bernsen, N., Dybkjær, H., Dybkjær, L. (1998). *Designing Interactive Speech Systems: From First Ideas to User Testing*. Springer, Berlin.
33. Francis, A., Nusbaum, H. (1999). Evaluating the Quality of Synthetic Speech. *Human Factors and Voice Interactive Systems*. Kluwer, Boston, MA, 63–97.
34. Sityaev, D., Knill, K., Burrows, T. (2006). Comparison of the ITU-T P.85 standard to other methods for the evaluation of Text-to-Speech systems. In: Proc. 9th Int. Conf. on Spoken Language Processing (Interspeech 2006 – ICSLP), Pittsburgh, PA, 1077–1080.
35. Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. *Comput. Speech Lang.* 19(1), 55–83.
36. Nielsen, J., Mack, R. (eds) (1994). *Usability Inspection Methods*. Wiley, New York, NY.
37. Oulasvirta, A., Möller, S., Engelbrecht, K., Jameson, A. (2006). The relationship of user errors to perceived usability of a spoken dialogue system. In: Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, Berlin, 61–67.
38. ISO 9241-110 (2006). *Ergonomics of human–system interaction. Part 110: Dialogue principles*. International Organization for Standardization, Geneva.
39. Constantinides, P., Rudnicky, A. (1999). Dialog analysis in the Carnegie Mellon Communicator. In: Proc. 6th Eur. Conf. on Speech Communication and Technology (EUROSPEECH'99), Budapest, 243–246.
40. Billi, R., Castagneri, G., Danieli, M. (1996). Field trial evaluations of two different information inquiry systems. In: Proc. 3rd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA'96), Basking Ridge, NJ, 129–134.
41. Boros, M., Eckert, W., Gallwitz, F., Gorz, G., Hanrieder, G., Niemann, H. (1996). Towards understanding spontaneous speech: Word accuracy vs. concept accuracy. In: Proc. 4th Int. Conf. on Spoken Language Processing (ICSLP'96) IEEE, Piscataway, NJ, 1009–1012.
42. Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistics. *Comput. Linguist.* 22(2), 249–254.
43. Cookson, S. (1988). Final evaluation of VODIS – Voice Operated Database Inquiry System. In: Proc. SPEECH'88, 7th FASE Symposium, Edinburgh, 1311–1320.
44. Danieli, M., Gerbino, E. (1995). Metrics for evaluating dialogue strategies in a spoken language system. *Empirical Methods in Discourse Interpretation and Generation. Papers from the 1995 AAAI Symposium*, Stanford, CA. AAAI Press, Menlo Park, CA, 34–39.
45. Gerbino, E., Baggia, P., Ciaramella, A., Rullent, C. (1993). Test and evaluation of a spoken dialogue system. In: Proc. Int. Conf. on Acoustics Speech and Signal Processing (ICASSP'93), IEEE, Piscataway, NJ, 135–138.
46. Goodine, D., Hirschman, L., Polifroni, J., Seneff, S., Zue, V. (1992). Evaluating interactive spoken language systems. In: Proc. 2nd Int. Conf. on Spoken Language Processing (ICSLP'92), Banff, 201–204.
47. Hirschman, L., Pao, C. (1993). The cost of errors in a spoken language system. In: Proc. 3rd Eur. Conf. on Speech Communication and Technology (EUROSPEECH'93), Berlin, 1419–1422.
48. Kamm, C., Litman, D., Walker, M. (1998). From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems. In: Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98), Sydney, 1211–1214.
49. Polifroni, J., Hirschman, L., Seneff, S., Zue, V. (1992). Experiments in evaluating interactive spoken language systems. In: Proc. DARPA Speech and Natural Language Workshop, Harriman, CA, 28–33.

50. Price, P., Hirschman, L., Shriberg, E., Wade, E. (1992). Subject-based evaluation measures for interactive spoken language systems. In: Proc. DARPA Speech and Natural Language Workshop, Harriman, CA, 34–39.
51. San-Segundo, R., Montero, J., Colás, J., Gutiérrez, J., Ramos, J., Pardo, J. (2001). Methodology for dialogue design in telephone-based spoken dialogue systems: A Spanish train information system. In: Proc. 7th Eur. Conf. on Speech Communication and Technology (EUROSPEECH 2001–Scandinavia), Aalborg, 2165–2168.
52. Simpson, A., Fraser, N. (1993). Black box and glass box evaluation of the SUNDIAL system. In: Proc. 3rd Eur. Conf. on Speech Communication and Technology (EUROSPEECH'93), Berlin, 1423–1426.
53. Skowronek, J. (2002). Entwicklung von Modellierungsansätzen zur Vorhersage der Dienstqualität bei der Interaktion mit einem natürlichsprachlichen Dialogsystem. Diploma thesis (unpublished), Institut für Kommunikationsakustik, Ruhr-Universität Bochum.
54. Walker, M., Litman, D., Kamm, C., Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. In: Proc. of the ACL/EACL 35th Ann. Meeting of the Assoc. for Computational Linguistics, Madrid, 271–280.
55. Walker, M., Litman, D., Kamm, C., Abella, A. (1998). Evaluating spoken dialogue agents with PARADISE: Two case studies. *Comput. Speech Lang.* 12(4), 317–347.
56. Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., Hetherington, L. (2000). JUPITER: A telephone-based conversational interface for weather information. *IEEE Trans. Speech Audio Process.* 8(1), 85–96.
57. Hone, K., Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Nat. Lang. Eng.* 6(3–4), 287–303.
58. Hone, K. S., Graham, R. (2001). Subjective assessment of speech-system interface usability. In: Proc. 7th Eur. Conf. on Speech Communication and Technology (EUROSPEECH 2001–Scandinavia), Aalborg, 2083–2086.
59. Möller, S., Smeele, P., Boland, H., Krebber, J. (2007). Evaluating spoken dialogue systems according to de-facto standards: A case study. *Comput. Speech Lang.* 21(1), 26–53.
60. Möller, S., Smeele, P., Boland, H., Krebber, J. (2006). Messung und Vorhersage der Effizienz bei der Interaktion mit Sprachdialogdiensten. In: Fortschritte der Akustik - DAGA 2006: Plenarvortr., Braunschweig, 463–464.
61. Walker, M., Kamm, C., Litman, D. (2000). Towards developing general models of usability with PARADISE. *Nat. Lang. Eng.* 6(3–4), 363–377.
62. Walker, M., Kamm, C., Litman, D. (2005). Towards generic quality prediction models for spoken dialogue systems – A case study. In: Proc. 9th Eur. Conf. on Speech Communication and Technology (Interspeech 2005), Lisboa, 2489–2492.
63. Dybkjær, L., Bernsen, N. O., Minker, W. (2004). Evaluation and usability of multimodal spoken language dialogue systems. *Speech Commun.* 43(1–2), 33–54.
64. Beringer, N., Louka, K., Penide-Lopez, V., Türk, U. (2002). End-to-end evaluation of multimodal dialogue systems: Can we transfer established methods? In: Proc. 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002), Las Palmas, 558–563.
65. Bernsen, N., Dybkjær, L., Kiilerich, S. (2004). Evaluating conversation with Hans Christian Andersen. In: Proc. 4th Int. Conf. on Language Resources and Evaluation (LREC 2004), Lisbon, 1011–1014.
66. Araki, M., Doshita, S. (1997). Automatic evaluation environment for spoken dialogue systems. In: *Dialogue Processing in Spoken Language Systems*. Proc. ECAI'96 Workshop, Budapest. Springer, Berlin, 183–194.
67. López-Cozar, R., de la Torre, A., Segura, J., Rubio, A. (2003). Assessment of dialogue systems by means of a new simulation technique. *Speech Commun.* 40(3), 387–407.
68. Walker, M. (1994). Experimentally evaluating communicative strategies: The effect of the task. In: Proc. Conf. Am. Assoc. Artificial Intelligence (AAAI'94), Assoc. for Computing Machinery (ACM), New York, NY, 86–93.

69. Walker, M. (1992). Risk Taking and Recovery in Task-Oriented Dialogue. PhD thesis, University of Edinburgh.
70. Möller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A., Oulasvirta, A., Raake, A., Reithinger, N. (2006). MeMo: Towards automatic usability evaluation of spoken dialogue services by user error simulations. In: Proc. 9th Int. Conf. on Spoken Language Processing (Interspeech 2006 – ICSLP), Pittsburgh, PA, 1786–1789.
71. Möller, S., Heimansberg, J. Estimation of TTS quality in telephone environments using a reference-free quality prediction model. In: Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, Berlin, 56–60.
72. Compagnoni, B. (2006). Development of Prediction Models for the Quality of Spoken Dialogue Systems. Diploma thesis (unpublished), IfN, TU Braunschweig.

Index

A

- Acceleration, 135, 200, 256–257
Acceptability, 254, 262, 301, 303–305
Acceptance, 43, 171, 214–215, 264, 308, 315
Accept/reject decision, 234–235, 238
Acoustic environment, 263
Acoustic model, 7, 11–12, 28, 169, 171, 187
Acoustic noise, 256, 284, 316
Action, 36, 41–42, 44–45, 48–49, 53, 91–93, 95–97, 128, 133–134, 138, 154, 159, 178–179, 199, 205, 239–242, 258, 273, 281–282, 296, 313
Active noise cancellation, 264
Active noise reduction, 263
Adaptivity, adaptation, 7, 9, 12, 14, 161, 171, 189, 205, 222, 265–266, 316–317
Advanced Drive Assistance Systems (ADAS), 197, 200, 203
Adverse environments, 256, 259
Affect, 109–113, 116–118, 138, 151–152, 154–158, 160–162, 198–199, 209–212, 252, 255, 283, 287
Affective computing, 117, 151–162
Affective HCI, 152
Agent
 Belief-Desire-Intention (BDI), 36, 38, 49, 96
 deceptive, 134, 144
Air force, 251, 258, 262
Air Traffic Control (ATC), 251, 253, 258, 261
Air Travel Information Service (ATIS), 6, 14, 62, 174, 176, 307
Algorithm-based approaches, 263
Annotation of corpora, 140
Anthropomorphic, 63
Anytime algorithm, 176
Application, 1, 6–9, 11, 13–15, 23, 27–28, 30, 34, 37, 39–40, 44–45, 50–51, 61–76, 89–90, 105–106, 116, 119, 123–125, 127, 129, 137, 152, 155, 157–159, 161, 162, 172, 180–181, 183, 185, 187, 189, 195–215, 221–249, 251–267, 273–298, 308–309, 315
Appraisal, 92, 133–134, 159
Architecture
 modular, 42, 185
 Open Agent Architecture (OAA), 38, 223
 rule-based, 186
Armoured Fighting Vehicles (AFVs), 256–257, 259
Army personnel, 252, 255, 262, 265
Artificial cognitive systems, 91, 96–98, 100
Artificial intelligence (AI), 34–36, 45, 49, 92, 96, 100, 157–158, 162, 294
Artificial speech, 21, 254
Assessment, *see* Evaluation
Asymmetric error, 237, 249
Attentive behaviour, 137, 143
Attribute-value pairs, 178, 307
Auditory information, 97, 253
Authoring, 65–66, 68–69, 277, 279
Automated Language Translation (ALT), 259–260, 277
Automatic Speech Recognition (ASR), 1–3, 7, 9–11, 15, 62, 66–69, 75, 85, 89–91, 98, 99, 168–169, 182, 185–189, 198, 202–203, 251–254, 256–259, 261–265, 279–280, 282, 287, 301, 307
Awareness, 49, 91, 204, 259–260
- ## B
- Back-translation, 184
Bag-of-words model, 186, 236
Bandwidth, 267, 286
Battlefield conditions, 260, 265
Battlefield environment, 261, 265, 267

- Bayes' theorem, bayesian, 28
- BDI (Belief-Desire-Intention) agent, 36, 38, 49, 96
- Behaviour, 35, 41, 43, 45, 47–49, 90–100, 123–124, 129–131, 133–137, 139–144, 152, 156–157, 162, 233, 239, 263, 287, 293, 296–297
- Behaviour expressivity, 135
- Beliefs, 34, 36–37, 44–45, 49, 96–97, 127, 130, 212
- Bilingual text, 170–171, 186–188
- Blizzard Challenge, 25–26, 28–29
- Body posture, 41, 48, 130–131
- Brazen head, 21
- Broadcast news (BN), 6
- C**
- Calibration, 236, 245, 249
- Call centers, 64, 75, 79
- Call-flow, 71–72, 74
- Call-routing, 39, 64
- Catalyst, 170
- Categorical representation, 133
- Cepstral, 5, 8, 12, 188, 201, 283
- Δ Cepstrum, 5, 8
- Chart structure, 176
- Chatbot, 35, 40
- Chunking, 119, 187–188, 293
- Cockpit, 254, 257–258
- Cognitive architecture, ACT-R, 96
- Cognitive architecture, BDI, 96
- Cognitive architecture, Soar, 96
- Cognitive informatics, 100–101
- Cognitive load, 50, 82, 198–199, 206–208, 212, 287
- Cognitive models, 95, 158
- Cognitive process, 91, 96, 100, 137, 198, 254
- Cognitive psychology, 92
- Cognitive Science, 52, 92, 94, 100, 139
- Collaboration, 49, 172, 174
- Combat roles, 252
- Combat scenario, 261
- Command, 8, 33–34, 36, 38, 50–51, 94, 117, 119, 205–208, 222–224, 233–234, 238–239, 241, 243–244, 247, 253–256, 258–262, 267, 274, 276, 281–282, 295, 297, 310
- Command and Control (C2), 8, 247, 253, 258–259, 261–262
- Command, Control, Communications and Intelligence (C3I), 256
- Command and Control On The Move (C2OTM), 259
- Command and control posts, 262
- Commercial Off-The-Shelf (COTS), 253–254
- Communication, 1, 3, 7, 10, 14–15, 34–38, 41–42, 44–45, 48, 51–52, 61, 63, 69, 90, 96, 106, 113, 120, 123–124, 126, 129, 136, 142, 156–157, 159, 169, 173, 183–184, 196–197, 206, 210, 212, 259–260, 262, 264, 271–272, 279–280, 284–289, 291–292, 296–297, 305, 307–308, 311, 315–317
- non-verbal, 41, 51, 156, 159
- verbal, 35
- Communicative act, 44
- Communicative capability, 34
- Communicative goal, 129
- Communicative principles, 34, 37, 51
- Communicative signals, 124, 130–133
- Comprehensiveness, 254
- Concatenation, 27, 29, 254
- Confidence scores, 169, 187, 224, 234–238
- Consortium for Speech Translation Advanced Research (C-STAR), 172, 174
- Context free grammar (CFG), 44, 63–64, 68, 177, 230
- Conversational Maxims, Grice's, 48, 308
- Conversational principles, *see* Communicative principles
- Conversational strategies, 33
- Conversational system, *see* Dialogue system
- Conversation Analysis (CA), 47
- Cooperation, cooperativity, 34, 36, 44–45, 47–49, 52, 75, 96, 158, 206, 221, 305, 308
- Core language engine (CLE), 174–175, 228
- Core technology, 68–70, 72, 89
- Corpus, bilingual, 187
- Corpus, Columbia-SRI-Colorado (CSC), 85
- Corpus, of Spontaneous Japanese (CSJ), 7
- Correction moves, 223, 240–241
- Crew station systems, 265
- Criteria-based content analysis (CBCA), 81
- Cross-channel early lexical learning (CELL), 97, 205, 226
- Cross-talk, 223, 233–234, 237, 248–249
- Cue-phrase(s), 39
- D**
- DARPA Communicator, 37
- DARPA Research Management, 5
- Database, 3, 5, 14–15, 24–25, 27–29, 35, 73–74, 160, 171–172, 222, 253, 261, 265, 285, 304, 307, 317

- Database, CMU Arctic, 25
 - Data-driven approach, 47, 124, 141, 222
 - Data glove, 255
 - Deception, 79–86, 134
 - Decision list, 229
 - Deep parsing, 187
 - Defence, 251, 262
 - Design, 5, 11, 14–15, 23–25, 29–30, 34, 38, 40, 47–48, 50–51, 62–63, 65, 70–71, 74, 119, 124, 139, 141, 144, 156, 176, 180–182, 188, 196, 198, 203–204, 215, 225–226, 238, 245–246, 261, 263, 265–266, 271–298, 308, 312, 314–315, 317
 - DiaLeague, 38
 - Dialect, 107, 172, 252, 263
 - Dialog, *see* Dialogue
 - Dialogue
 - act, 44–47, 124
 - agent, 51–52
 - applications, 221–249
 - context, 34, 238, 246
 - control, 42–43, 45
 - act, 45
 - design, 40, 47–48, 203
 - development, 34, 36, 45
 - evaluation, 39
 - grammar, 43–44
 - management, 33, 37–38, 41–44, 46–47, 127–130, 202, 205–207, 223, 238–242, 245
 - agent-based, 42–43
 - distributed, 43
 - frame-based, 42
 - intention-baseddialogue management, plan-based, 38
 - models, 42–43
 - script-based, 44
 - side-effect free, 238–242
 - statistical, 41
 - manager, 33, 37, 129, 204–205, 224, 227, 231, 238–239
 - model/modelling, 33–53, 127
 - moves, 129, 224–225, 231, 239–241
 - participant, 43
 - plan, 45, 48
 - strategy, 305
 - system, 33–43, 45, 47–49, 51–53, 79, 89, 126, 153, 157, 159, 196, 202–208, 221–222, 238, 242–245, 280, 288, 290–291
 - technology, 34–35, 38–39, 41, 52
 - Dimensional representation, 10, 133
 - Directed dialog, 63–65
 - Direct Voice Input (DVI), 258
 - Direct Voice Output (DVO), 258
 - Disambiguation, 38, 292
 - Discourse relations, 36
 - Distraction, 196–200, 205–207, 279
 - Doctor/patient examination dialogues, 177, 190
 - Document classification, 234–235
 - Domain, 8, 19, 24, 34, 40, 42, 45, 51–52, 64–65, 86, 92, 105, 117, 137, 144, 158, 169–170, 172, 174, 176–183, 185–186, 188–190, 204, 207, 212, 214, 228–233, 237, 245, 248–249, 251–258, 260, 262, 264–267, 304–307
 - Dynamic Interpretation Theory (DIT), 45
 - Dynamic time warping (DTW), 2–3, 29, 168–169, 265
- ## E
- EAGLES (Expert Advisory Group on Language Engineering Standards), 38, 302–303, 306
 - ECESS (European Center of Excellence in Speech Synthesis), 26
 - Edit distance, *see* Levenshtein distance
 - Effectiveness, 10, 13, 38, 64–65, 74, 272, 291, 303–304, 309, 315–316
 - Efficiency, 50, 63, 162, 187, 303–305, 315–316
 - Elektronische Musik, 23
 - Eliza, 35, 126
 - Ellipsis processing, 178
 - Embodied cognition, 96–97
 - Embodied conversational agent, 48, 123–144
 - Emotion
 - content, 254
 - signal, 124, 133–134
 - Empathy, 96, 98, 123
 - Emulation mechanisms, 92, 94
 - Enrolment, 252–253
 - Ergonomy, 50
 - Error correction, 260, 262, 264–265, 276, 279
 - Error handling, 39, 41
 - Errors, speaker-based, 263
 - Evaluation
 - automatic, 41, 316–317
 - heuristic, 50, 309, 312
 - PARADISE, 317
 - task-based, 183
 - Example-based machine translation (EBMT), 170–171, 182–184

Explanation Based Learning (EBL), 175,
177, 228

Eye contact, 81, 260

F

Facial expression, 33, 41, 48, 79, 123, 128,
130–131, 133, 135, 137, 139,
151–152, 154–156, 159, 161, 209,
255, 283–284

Fatigue, 183, 257

Feedback, 10, 38–39, 43, 47, 92–95, 97–98,
117, 127, 142, 184, 203, 206, 258,
262, 266, 307, 310

Festvox, 183

Field test, 183–184, 309

Fifth Generation Computer Systems
programme, 36

Finite state network (FSN), 3

Flight status, 63–64, 73

Forced alignment, 29

Form Interpretation Algorithm (FIA), 67, 73

G

Gaussian mixture model (GMM), 29, 156

Gaze, 41, 81, 130–133, 136–137, 139,
142–143, 156, 254–255, 265,
281–283, 296

Gender, 172, 209, 214, 252

Generalized EBMT (G-EMBT), 182

Generalized probabilistic descent (GPD), 6, 8

General public, 253

Gesture, 33–34, 36, 41, 48, 79, 81–82, 95,
106, 123, 126, 128–133, 135–142,
151–152, 156, 159, 161–162, 209,
254–255, 260–261, 265, 296, 315

G-forces, 256, 264–265

Globalisation, 267

GMM-based voice conversion, 30

Grammar, 5, 13, 35, 43–45, 63–69, 71–72, 118,
169–170, 173–178, 182, 185, 224,
228–232, 244, 247–248, 253, 310

Grammar-based language model (GLM), 169,
224, 228, 233, 237, 247–249

Grammar specification language (GSL), 228

Graphical user interface (GUI), 34, 50, 182,
185, 273–274

Grounding, 38, 47, 91, 103, 136–137, 142

H

Handheld platform, 181

Headset, 182, 188, 201, 263, 283

Helicopters, 256

Hidden Markov Model (HMM), 4, 13, 28,
155–156, 169, 201, 318

Hierarchical control, 93

Higher-order logic, 174

HMM-based speech synthesis, 29

Hosting, 69–70

How may I help you (HMIHY), 40, 64

Human–computer interaction (HCI), 33–35,
50, 100, 142, 151–157, 209

Human factors, 181–182, 198, 251–267, 313

Human–human communication, 123

Human–machine interface, 1, 34, 273

Human–machine speech communication, 262

Human–robot communication, 41

Human speech, 1, 6, 10–11, 15, 90, 106,
117–119, 171, 251, 256

Hybrid architecture, 173, 185

I

Ideal Cooperation, *see* Cooperation,
cooperativity

Imitation, 94–96, 98

Inflection, 175, 201

Information extraction, 6, 186

Information State Update (ISU), 204

Infotainment, 195, 197, 199, 206

Integration, 74, 99, 105, 154, 159, 161, 168,
185, 197, 233, 244, 264–267, 291,
294, 296

Intelligent interactive systems, 52

Intelligibility, 7, 20, 29, 254, 264, 303–304,
308

Intention, 36, 43–45, 93, 98, 117, 259

Interaction, 13–14, 33–53, 63, 66–67, 69,
71–74, 89–92, 97–100, 105–108,
116–117, 123, 125, 128, 130,
136, 142–143, 151–157, 159–162,
182, 184, 195–199, 202, 204–205,
207–210, 212, 215, 225, 272–274,
287–288, 291, 294, 305, 308–318
parameter, 309, 311, 313–318
problem, 309–310, 314

Interactive Electronic Technical Manual, 259

Interactive system, 33–35, 50, 52, 123, 181,
251, 257, 259, 264, 267, 273,
301–318

Interface

design, 50, 312

speech-enabled command, 33

tactile, 34

Interlingua, 170, 178–179, 189

International Standardisation Organisation
(ISO), 40, 310

Internet, 30, 35, 66, 70, 74, 162, 189, 197, 267,
284, 286, 292

- In-vehicle, 195–196, 198–215
- In-vehicle information systems (IVIS), 195, 197–200, 209, 212–215
- K**
- Kernel functions, 236
- Kernel methods, 236
- Kernels, linear, 237, 248
- Kernels, quadratic, 236–237, 248
- Keyboard, 23–24, 41, 50, 52, 66, 128, 143, 182, 225, 254, 258–262, 273, 279–284, 294, 296
- K-means, 186
- Knowledge-based machine translation (KBMT), 170, 182, 184
- L**
- Language model, 5, 7, 11, 13–14, 28, 118, 169, 176–177, 185–188, 224, 228, 230, 232–233, 247–248, 252–253, 310
- Language model, grammar-based (GLM), 169, 224, 228, 247–248
- Language pair, 168, 170–171, 173, 177–178, 180, 182
- Language and Speech Exploitation Resources (LASER), 3, 260
- Large Vocabulary Continuous Speech Recognition (LVCSR), 89
- Learning, 10, 14, 40–41, 43, 46, 52, 79, 84–86, 89–91, 96–97, 99, 128, 157–158, 175, 177, 228, 247, 265, 271, 273, 292, 294–296, 310, 316
- Learning by imitation, 96
- Left-corner parser, 175, 228
- Levenshtein distance, 20
- Lexical selection, 170
- Lexicon, 28, 97, 170, 172, 178, 228–230, 266
- Lexicon, bilingual, 170
- Limited english proficiency, 190
- Linear Predictive Coding (LPC), 3, 21, 24, 156, 201
- Linguistic Inquiry and Word Count (LIWC), 84
- Loebner Prize Competition, 52
- Logical form (LF), 224, 228, 231–233, 246
- M**
- Machine-learning technique, 40, 43, 46, 85
- Machine translation, example-based (EBMT), 170–171, 182–184
- Machine translation, knowledge-based (KBMT), 170, 182, 184
- Machine translation (MT), 19–20, 24, 38, 168–171, 182–184, 186–189, 265
- Machine translation, multi-engine, 176
- Machine translation, statistical (SMT), 170–171, 182–189
- Machine translation, transfer-based, 170
- Maintenance, 49, 61, 70, 205, 226, 259, 262
- Manual operations, 254
- Manual tasks, 257
- Maximum Likelihood Linear Regression (MLLR), 6, 12
- Maximum Mutual Information (MMI), 6, 8
- Meaning representation, 170
- Measure, measurement, 82, 155, 200, 301–303, 311, 316
- Medical applications, 189
- Medical domain, 179, 189–190
- Memory, 19, 27, 46, 83–84, 91–92, 94–96, 198, 253–254, 264, 274–275, 287, 292–293, 296, 313, 317
- Meta-communication, 307–308, 311
- Microphone arrays, 202, 205, 263
- Microphone-based technologies, 263
- Microphones, close-talking noise cancelling, 263
- Microtremors, 84
- Military applications, 251–267
- Military domain, 189–190, 251–257, 260, 262, 266–267
- Military personnel, 252, 255, 262, 265
- Military police, 251
- Mimics, 123
- Minimum Classification Error (MCE), 6, 8
- Mirror neurons, 92, 94–96
- Misrecognition, 176, 202, 223, 264
- Mixed initiative, 38, 49, 63–65, 67, 73, 204
- Mobility, 259, 261, 264–265, 271, 273, 280–282, 290, 294
- Mock-theft paradigm, 80
- Model/modelling
- excitation, 28
 - harmonic + noise, 28
 - probabilistic, 13, 39
 - statistical, 1, 4, 8, 13, 29, 61, 89, 169
- Mood, 110, 114, 157, 162, 208–210
- Moore's law, 189
- Multimodal information, 151–162
- Multimodal interaction, 34, 37, 41, 51, 204, 207–208, 265
- Multimodality, 205, 207–208, 264–265
- Multimodal system, 36, 151–152, 156–157, 207–208, 315
- Multiparty dialogue, 41
- Multiparty interaction, 41

- Multi-Role Fighter aircraft, 258
MUMIN annotation scheme, 48
- N**
- Native speakers, 85, 168, 172, 182, 263, 265
Natural communication, 34, 41
Natural disasters, 70, 259
Natural interaction, 33, 42, 50–53
Natural language, 6, 33, 35–37, 40, 42, 50, 62–65, 73–75, 91, 100, 157, 170, 198, 204, 253, 272, 302, 304, 306–308
Natural language understanding, 62, 64–65, 198, 204, 302, 304, 306–307, 314
Naturalness, 20, 41, 64, 153, 182, 202, 254, 272, 303, 305, 308
Navy personnel, 262
Neural net, 5, 43, 46, 92, 155–156, 169, 186, 318
N-gram, 5, 8, 13, 46, 177, 186, 231–233, 236, 246–248
Noise, 6, 10–12, 23, 28, 135, 155, 161, 182, 195–196, 200–202, 205, 243–244, 256–257, 261, 263–265, 267, 283, 284, 305, 310, 316
Noisy environment, 7, 90, 198, 202, 254, 256, 264
Non-native speakers, 263, 265
Non-speech warnings, 258
Non-Verbal Behaviour (NVB) coding, 81, 130, 143
- O**
- Objectivity, 20, 302
Office applications, 262
Open Agent Architecture (OAA), 38, 223
Open Source, 41, 177, 228–229
- P**
- Parallel corpus, 171, 182, 186–187
Parallel model composition (PMC), 6
Parameter, 9, 11–12, 99, 124, 135, 153, 176, 235–237, 248–249, 288, 308, 314
Paraphrase, 111, 179, 184, 186, 188, 225
Parse preferences, 175
Parser, 13, 173, 175, 186–187, 228
Parser, robust, 187
Parse tree, 170
Passive noise reduction, 263
Perception, 15, 51, 80–81, 86, 91–93, 97, 137, 141–142, 151–152, 156, 160, 162, 199, 212, 273, 302, 304, 312–313
Perceptive behaviour, 136–137
Perceptual control system, single-layered, 93
Performance, 6–7, 9–15, 20, 24, 26, 38–40, 48, 51, 61–64, 79, 90, 101, 108, 119, 126, 128, 135, 141, 155, 157, 169, 175–177, 184, 188, 198–203, 209–215, 222–223, 231–234, 236–238, 243–249, 252–253, 256–257, 262–266, 294, 301–303, 307, 309, 311, 313–315, 317–318
Personality, 127, 129, 140, 142, 208–209, 211, 213–214, 305
Personification, 123
Phonetic analysis, 252–253
Phrase alignment, 171
Planning, 36–38, 45, 48–49, 91, 129, 137–139, 204, 289
Planning operator, 45
Platform, 21, 26, 41, 65–66, 68–71, 73, 172, 174, 177, 181–183, 188–189, 206, 224, 228, 235, 292
Politeness, 48, 112, 137–140, 143
Portability, 40, 172, 204
Portable devices, 197
PREdictive SENsorimotor Control and Emulation (PRESENCE), 98–99
Preference, 141
Prepackaged applications, 69
Probabilistic training, 177, 230
Procedures, navigation of, 221, 225
Procedures, rendering in voice, 244
Professional services, 69
Prolog, 224–225, 240
Prototype, 75, 176–177, 184, 204–205, 208, 222, 245, 260, 266–267, 294
Pulse Code Modulation (PCM), 23, 28
Push-to-talk, 195, 202, 206, 254
Put-That-There, 36
- Q**
- Quality
 assurance, 71–72
 prediction, 315, 317–318
Quasi Logical Form (QLF), 174–176
Question-Answering (QA), 41, 51, 72, 302
Questionnaire, 215, 311–312
- R**
- Rational cooperation, *see* Cooperation, cooperativity
Reality Monitoring (RM), 81
Reasoning, 34, 37, 44, 49, 91, 227
Reconnaissance, 254, 258
Regression testing, 238, 242, 245
Reinforcement learning, 40–41, 43
Reliability, 187, 213, 302, 316, 318

- Re-ordering, 187
- Reversion, 264, 266
- Robot, 38, 41, 51–52, 97, 126, 137, 142, 158, 162, 238, 254–255
- Robustness, 6–8, 11–12, 48, 90, 176–177, 186, 196, 198, 222–223, 244, 263–265, 283, 288, 302, 316
- Role-playing exercises, 183

- S**
- Safety-critical tasks, 263, 266
- Scale, 2, 25, 35, 37–39, 157, 179, 213, 222, 302, 308, 311–312
- Scenario, 71, 80, 86, 90, 127–128, 130, 134, 144, 156, 161–162, 184, 188, 200, 257, 261, 279, 287–288, 294, 308–309, 312–313, 315
- Search algorithm, 187–188
- Search task, 254–255
- Security agencies, 251
- Semantics
 - atom, 229, 231
 - content, 45, 132, 306
 - Error Rate (SemER), 177, 232–233, 247
 - form, 174
 - representation, 173, 178, 186, 229, 231
 - type, 178
- Sensitivity, 155, 302
- Sentence Error Rate (SER), 233, 247, 306
- Shallow analysis, 129
- Simulation-based approach, 141
- Situation awareness, 259–260
- Social dialogue, 137–138
- Social dynamics, 41
- Sortal constraints, 178
- Source language, 170, 175, 177, 179, 184, 260
- Space, 13, 23, 27, 29, 37, 52, 96, 108–109, 128, 130, 132–133, 136, 151, 155, 167, 171, 196, 221–249, 258, 261, 263, 282–283, 289–290, 304
- Speaker
 - adaptation, 9
 - behaviour, 263
 - dependent, 81, 85, 252–253, 262, 264
 - independent, 252
 - variability, 263
- Speech
 - affective, 79
 - application lifecycle, 70–73
 - computer-generated, 254
 - corpus, 7, 27, 84
 - emotional, 137, 151–154, 196
 - enhancement, 9, 201–202, 205
 - expressive, 105–120, 153–154
 - feature extraction, 201
 - input, 34, 91, 105, 185, 224, 252, 254–255, 258, 271, 286, 290, 296, 302, 311–312
 - output, 185, 224, 252, 274–276, 278–279, 284–286, 290, 296, 303–304, 308–309, 312, 314
 - recognition
 - grammar-based, 176–177, 224, 228–233
 - robust, 6–7, 12, 202, 263, 265
 - spontaneous, 7–8
 - recognizer, DECIIPHER, 174
 - synthesis, 19–30, 66, 68, 71, 89, 90, 118–119, 152–154, 168, 171–174, 182, 188, 189, 200, 251–254, 258–259, 263–264, 271, 273, 276, 285, 294, 303–304
 - synthesized, 254, 256
 - understanding, 3, 6, 8–10, 231–235, 244, 246–248, 304, 306
 - understanding, evaluation of, 231–233
 - warnings, 256, 258
- SPeech In Noisy Environments (SPINE), 256
- Speech Under Simulated and Actual Stress (SUSAS), 265
- Spoken conversation, 33–34
- Spoken dialog, *see* Spoken dialogue modelling
- Spoken dialogue modelling, 33, 35, 38, 42–44, 50
- Spoken dialogue system, *see* Dialogue system
- Spoken language dialogue system (SLDS), 89
- Spoken language interactive systems, *see* Dialogue system
- Spoken language technology (SLT), 89–101
- Stack decoder, 182
- Standards, 10, 20, 24, 26, 30, 40, 62, 65–69, 73–74, 85, 107, 111, 114, 155, 157, 172, 184, 199, 206, 215, 223, 233–234, 236, 247, 261, 284, 292, 295, 297, 302–303, 310, 312–314
- Standard writing system, 172
- Statement Analysis, 84, 86
- Statistical Language Model (SLM), 169, 176–177, 224, 228, 233, 237, 247–249
- Statistical language understanding (SLU), 64–65, 73
- Stress, 10, 27, 79–80, 84–85, 153, 196, 201, 204–205, 215, 256–257, 263–265, 267, 272, 276, 288

- Structural maximum a posteriori (SMAP), 6
- Subjective Assessment of Speech System Interfaces (SASSI), 312
- Sub-symbolic, 92
- Support Vector Machine (SVM), 152, 157, 186, 225, 235–238, 244, 248–249
- Surface parsing, 186
- Symbolic, 35, 40, 43, 92, 96, 132, 241
- Synchronization, 129
- Syntax, 2, 9–10, 39, 167, 171, 175, 243
- System
 - interactive, 33–35, 50, 52, 123, 181, 251, 257, 259, 264, 267, 273, 301–318
 - prompts, 40
 - speech-based, 33–34, 51, 198, 210, 212, 215, 251–266, 312
 - telephone-based, 308, 314
 - tutoring, 51, 79, 138
- T**
- Target language, 170–171, 173, 175, 178, 180–181, 184, 186, 188, 260
- Task-oriented act, 45
- Tasks, control and data entry, 253
- Teamwork, 49
- Teleoperation, 255
- Telephone service, 284, 312
- Text generation, 4, 37, 293
- Text, parallel, 183
- Text-to-speech synthesis (TTS), 23, 26–28, 66, 68–69, 71, 89–91, 98, 153, 188, 198, 223–224, 273, 276–277, 285, 287, 293–297, 301, 303
- Thinking Machine, 35–36, 52
- Topic segmentation, 39
- Touch input, 254–255, 265
- Touch screen, 182, 207, 254
- Touch-tone, 63, 65, 70–71, 73, 75
- Trackballs, 261
- Training, 5–6, 11, 14, 28–29, 40, 43, 47, 63, 75, 80, 86, 113, 155, 169, 171, 177, 183–184, 186–187, 228–231, 233, 236–237, 243, 245, 251–253, 258, 260–263, 275, 314
- Training simulations, 258, 262
- Trait, prototypical, 208
- Transfer rules, 170, 176
- Translation
 - engine, 169–171, 185–187, 189
 - one-way, 180
 - speech-to-speech, 24, 38, 173, 181, 259
 - spoken language, 167–190
 - two-way, 181
- Translator, universal, 167–168
- Treebank, 228
- The Turk, 21
- Turn-taking, 33, 39, 43, 47, 99, 119
- U**
- UAV (Unmanned Aerial Vehicle), 253, 258, 262
- Ubiquitous computing context, 51
- Unit selection, 19, 27–29, 89, 156
- Update function, 225, 239–240
- Usability, 38, 40, 50, 71–72, 74–75, 159, 169, 182, 196, 199–200, 207, 214–215, 245, 266, 297, 301–318
- Usability inspection, 309, 312–314
- User
 - centred design, 50
 - experience, 51, 62, 64, 70, 74–75
 - input, 34, 67, 71, 73, 126
 - satisfaction, 39–40, 51, 65, 305, 312–314
- U.S. Naval Voice Interactive Device (NVID), 261–262, 266
- Utterance, 2–3, 7, 10, 13–14, 19, 23–25, 27, 29, 33, 35, 37–41, 43–47, 64, 72, 106, 108–112, 114, 116–118, 128, 130, 139, 142–143, 152, 169, 175–179, 182, 184–185, 188–189, 202, 205–206, 229–239, 242, 247–249, 254–256, 276, 282, 303, 306–307, 311, 316, 318
- V**
- Validity, 199, 302, 309
- Vibration, 255–257, 261, 264, 267
- VICO project (Virtual Intelligent CO-driver), 204
- Virtual environment, 127, 260
- Visual displays, 198, 208, 225, 241, 257–258
- Vocal jitter, 84
- Voice
 - characteristics, 211
 - commands, 8, 206, 255, 258–259
 - correcting, 224
 - conversion, 29–30
 - cross-language, 30
 - text-independent, 29
 - input, 255, 258, 280
- Voice Interactive Display (VID), 259, 261
- Voice Stress Analysis (VSA), 79, 84–85
- Voice User Interface (VUI), 34, 62–63, 70–71, 74, 188
- VoiceXML, 40, 42, 66–69, 73
- Voting, 186

W

- W3C, 40, 68, 276–277, 288, 292
- White lies, 80
- Wireless technology, 261
- Wizard-of-Oz experiments, 182
- Word Error Rate (WER), 169, 177, 233, 247, 252, 303, 306
- Workload, 198–201, 203, 212, 257, 265, 267