Multimodal Model Integration for Sentence Unit Detection

Lei Chen
Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907-1285
chenl@purdue.edu

Yang Liu*
International Computer Science Institute
Berkeley, CA 94704
yangl@icsi.berkeley.edu

Mary P. Harper
Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907-1285
harper@purdue.edu

Elizabeth Shriberg†
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
ees@speech.sri.com

ABSTRACT

In this paper, we adopt a direct modeling approach to utilize conversational gesture cues in detecting sentence boundaries, called SUs, in video taped conversations. We treat the detection of SUs as a classification task such that for each inter-word boundary, the classifier decides whether there is an SU boundary or not. In addition to gesture cues, we also utilize prosody and lexical knowledge sources. In this first investigation, we find that gesture features complement the prosodic and lexical knowledge sources for this task. By using all of the knowledge sources, the model is able to achieve the lowest overall SU detection error rate.

Categories and Subject Descriptors: H.5.1 [Multimedia Information Systems] Audio and Video Input, H.5.5 [Sound and Music Computing] Modeling and Signal Analysis, I.2.7 [Natural Language Processing] Dialog Processing

General Terms: Algorithms, Performance, Experimentation, Languages.

Keywords: multimodal fusion, gesture, prosody, language models, sentence boundary detection, dialog.

1. INTRODUCTION

People, when understanding human-to-human communications, do not simply focus on words and their meaning. They utilize everything they can in order to understand the communication, including information from the visual domain such as other speakers' gesture and gaze. Speech and gesture are known to exhibit a synchronous relationship in human communication [21, 22]; however, how this information is synthesized to reach understanding is currently an important unanswered question. Unlike words, which tend to map more directly to a meaning, the intent of a gaze or a gesture is much harder to interpret. If we can better understand how a multimodal language performance encodes its meaning, that knowledge could be exploited to build a computer model to support higher quality multimodal human-to-computer exchanges. For example, there are a host of reasons that a dialog participant will retract their hands to their lap (e.g., completion of the gesture, completion of an idea unit, giving up the floor, fatigue); if we can better understand those reasons, our human-computer dialog model should be more effective. In this investigation, we will attempt to incorporate gestural information into a dialog processing task.

There are several ways to obtain gestural measurements. The first involves the human coding of some intermediate representation such as gesture stroke [25] or is based on some pre-defined gesture syntax [15]. Though this method may provide insight into the nature of gesture, it is very time consuming to code and is subject to human error. The second method uses highly accurate motion tracking equipment such as a digital glove [13] to track the movements of a hand (e.g., the location of a finger and its joints). This method provides more accurate fine granularity features; however, such equipment may affect the nature of the conversational gestures and, more importantly, a conversational corpus recorded in this fashion is currently unavailable. The third method is to track hand motion directly from the video; we use this method in this investigation. To support the development and evaluation of multimodal models, we have constructed a multimodal corpus of digital temporally synchronized video and audio recordings of human monologs and dialogs [1]. This database can be utilized to conduct a variety of measurement studies and to develop computer models. An important focus of this paper is on methods of fusing gesture and speech information in a human communication task.

There has been a considerable amount of work on constructing computer models for the spoken language portion of a human dialog. For example, Shriberg and Stolcke [31] have pioneered the "direct modeling" approach to exploit prosodic information in a variety of spoken language processing tasks such as speaker recognition, topic segmentation, and sentence segmentation. An advantage of this approach is that no hand segmentation or intermediate labeling of the speech prosody is required. Instead the prosodic features are extracted directly from the speech signal given its time alignment to a human generated transcription or

*Also in Electrical and Computer Engineering at Purdue.

†Also at International Computer Science Institute.

ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.

to automatic speech recognition (ASR) output. A prosody model is trained using these features, and this model is then combined with a language model in the final system.

Computational models that integrate speech and visual (e.g., gesture) features are in their infancy. Several measurement studies have looked at the use of gesture within human communication. Kendon [16, 8] found that the most effortful part of a speech-accompanying gesture (called the stroke phase) tends to occur together with or just before the most prominent syllable of the speech. Chen et al. [10] have conducted a measurement study on gesture patterns used to mark speech repairs. Sharma et al. [30] combined an HMM architecture for continuous gesture recognition with keyword spotting to explore the relationship between gesture and speech in videotaped weather forecasting narrations. They demonstrated that the co-occurrence of different gestures with a selected set of spoken keywords improves the performance of their continuous gesture recognition system. On the same task, Kettebekov et al. [17] used F_0 to increase robustness of their system. In a series of studies, Quek et al. [24, 25, 27, 28] investigated the role of gesture for signaling discourse structure. Several trackable gesture features were investigated including hand position [24], hand symmetry [28], and hand oscillation [27]. Chai et al. [9] used a graph-matching algorithm to exploit semantic, temporal, and contextual constraints from pen based gesture and speech to more effectively resolve different types of references. Although each of these studies enhance understanding of multimodal fusion, none of them have investigated whether a direct modeling approach is effective.

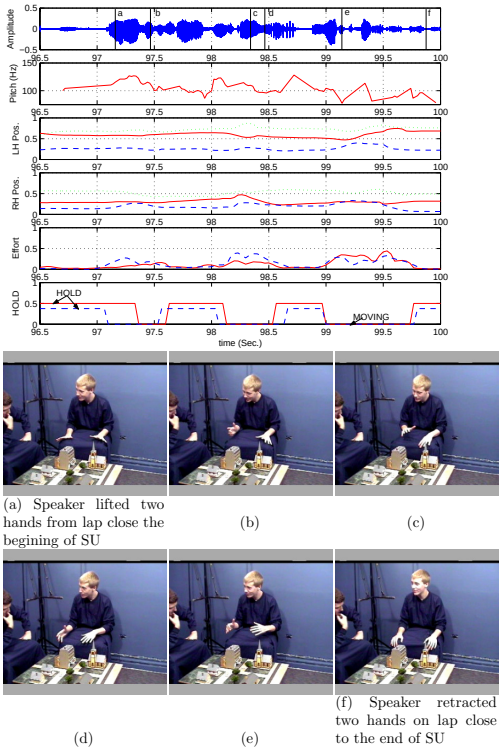
In this paper, we use a direct modeling approach to incorporate gesture features into a natural conversational multimodal model for detecting sentence boundaries in videotaped conversations. This research is highly related to one of the goals of the DARPA EARS program [2]– to enrich speech recognition output by automatically adding metadata events (such as sentence boundaries, disfluencies, and speaker labels). The basic idea is that, although current speech recognition systems output a stream of words, the addition of structural information such as punctuation should aid the human comprehension of the output while enabling more effective downstream natural language processing (e.g., some systems expect sentences as input). Additionally, this information can be helpful for speech recognition, since language models generally use linguistic segments to rescore utterance hypotheses [11]. For this first experiment on direct modeling of gesture, we focus on sentence boundaries since they are fairly frequent in a conversational corpus. This is important since the size of our multimodal corpus is quite small relative to the metadata corpora available within the EARS program.

Sentence boundaries in spontaneous speech are not as well defined as in written text. The sentence unit investigated in the EARS program is called an SU; see [35] for an annotation specification of SUs. An SU typically expresses a speaker's complete thought or idea unless there is some disruption. Sometimes the unit is semantically complete but smaller than a sentence (e.g., a noun phrase in response to a question). An SU can be an entire well-formed sentence, a phrase, or a single word. SU detection algorithms have utilized several types of knowledge [18, 19], in particular, textual and prosodic information. Textual information is based on the words in the transcription generated either by a human or by an ASR system. Prosodic information involves the "rhythm" and "melody" of speech, which provides complementary information to the words in the speech for SU detection and so can add robustness against ASR errors.

These previous investigations have been restricted to audio information sources; however, SU events in human conversations typically occur in an environment in which other modalities such as vision (e.g., gesture, facial expression, and body posture) can be used. McNeill proposes that when visual cues are generated by the same thought processes that produce an utterance, that they function together to serve the communication process with each source carrying some aspect of the original thought [22]. If such is the case, we would expect both speech and gesture cues to work together to signal SU boundaries. In this paper, we investigate whether visual cues from videotaped human-to-human dialogs contribute to SU detection accuracy beyond lexical and prosodic features.

The rest of the paper is organized into several sections. In section 2, we briefly describe the multimodal data used in this investigation. In section 3, we describe the model for each knowledge source and discuss model combination strategies. Finally section 4 describes the experimental setup and discusses the experimental results.

2. WOMBAT DATA AND ANNOTATION



(a) Speaker lifted two hands from lap close the begining of SU (b) (c)

(d) (e) (f) Speaker retracted two hands on lap close to the end of SU

Figure 1: Temporally aligned audio and video features for a portion of the multimodal data set (between 96.5 and 100 seconds of the wd20-subj set; see Table 1) with the following transcript: *and$_a$ we need to decide$_b$ who's doing$_c$ what$_d$ and what equipment$_e$ we're gonna$_f$ bring to scare them$_{[SU]}$*. From top to bottom, the features are: (1) the speech waveform, which is marked with six time points indicated in the transcript. Note that video frames for these points appear in the lower part of the figure. (2) pitch (F_0). (3) the left hand's (LH) 3D position (X is a solid line, Y dashed, and Z dotted). All positions are normalized to [0,1]. (4) the right hand's (RH) 3D position. (5) normalized *Effort*. A solid line indicates the LH and a dashed line the RH. (6) *Hold*, such that a non-zero value indicates a gesture *Hold* and a zero value indicates hand movement.

Human dialogs were videotaped using a set of cameras that were both temporally and spatially calibrated [1] and audio recorded using unidirectional boom mounted microphones. These dialogs were recorded in a somewhat noisy laboratory environment and contain frequent speech overlap. In each session, a subject and interlocutor pair who knew each other sat next to each other facing a model of a town (see the keyframes in Figure 1). The subject was shown the following script describing the task to be accomplished:

> A family of intelligent wombats has taken up residence in an abandoned movie theater in the town of Arlee. You and your assistant need to catch the wombats so that you can send them back to Australia. You will be taking the train to Arlee, so be ready to get off at the station right after you pass a church on your right. When you get off the train, go around the station and cut through the adjacent park to meet your assistants: pass between the two trees and you should reach the house number 33. Then go next door and ask the neighbors in 35 (you'll notice the road construction in front of the house) to assist you. The movie theater is across the intersecting street. One of you should go in the front entrance and scare the wombats out the back entrances. With the help of the people in 33 and 35, you should be able to snare the wombats as they exit the rear entrance. Explain the task to your assistant and decide on who does what, and what equipment you will need to bring with you.

The subject first explained the situation to the interlocutor using the town model, and then together they developed a plan.

Data was recorded with a 5-DV camera, 2-microphone setup (1 pair of stereo cameras aimed at each subject and interlocutor, 1 zoomed camera to capture detailed head/gaze information, 1 microphone for each subject and interlocutor fed into different channels of one of the cameras to provide audio separation). We used off-the-shelf consumer-grade miniDV 30 frames-per-second cameras in progressive scan. Using a five foot-wide prism with a known constellation of points, we are able to obtain points with typical average errors within 1 mm in x and y and about 1.5 mm in z (toward the cameras). The maximal errors are within 4 mm. This has been sufficient for measurements involving conversational gesture interaction. The audio for each participant was digitally recorded using a Shure Sm94 unidirectional boom mounted microphone that was placed at a distance of approximately eight inches from the subjects' mouths. The video and audio were time synchronized using a movie clapper device. The video was digitized on an SGI workstation and saved in SGI MPEG format. The audio was initially sampled at $44.1KHz$ and then downsampled to $14.7KHz$ for analysis.

All videos were processed using a fuzzy image processing approach, known as Vector Coherence Mapping, that was used to track the hand motion [23]. VCM applies spatial coherence, momentum (temporal coherence), speed limit, and skin color constraints in the vector field computation by using fuzzy combination strategies, and produces good results for hand gesture tracking. An iterative clustering algorithm was applied that minimizes spatial and temporal vector variance to extract moving hands. The positions of the hands in the stereo images were used to produce 3D motion traces describing the gestures. Three gesture features were extracted for each hand of a speaker: $3D$ hand position, *Hold* (a state when there is no hand motion beyond some adaptive threshold (see [14]); a motion energy-based detector was used to locate places where there was low motion energy [6]), and *Effort* (analogous to the kinetic energy of hand movement [6]).

A transcript of each conversation side of each dialog was prepared, and it was force aligned to the audio and then hand adjusted by an experienced speech scientist using the Praat tool [3]. The time aligned transcription was also annotated with SUs using version 5.0 of the EARS Simple Metadata Annotation Specification [35]. SU boundaries were marked by the second and third author using both the recorded speech and its transcription. All disagreements in annotation were resolved by discussion.

For this experiment, we chose to use three data files from our digital wombat video corpus for which the percentage of hold is lower than 90%, which we call the KDI data set. Table 1 indicates the duration of each data file in seconds (s) along with the number of words spoken and SUs produced. It also indicates the right hand's (RH) hold percentage for each set.

Speaker	Dur. (s)	# Words	# SUs	RH Hold %
wd01-subj	509	1243	116	65.23
wd20-subj	727.24	1255	228	80.82
wd20-int	727.24	1453	250	79.76
total		3951	594	

Table 1: Characteristics of the KDI data set used in this experiment.

3. MULTIMODAL MODELS

We treat SU detection as a classification task such that for each inter-word boundary, the classifier decides whether there is an SU end boundary or not. The SU boundary decision is based on three somewhat independent knowledge sources: prosodic, textual, and gestural cues. We believe that the combination of these three sources should improve the overall accuracy of the system relative to each source alone or in pairwise combinations. Following the direct modeling approach, we utilize the time alignment of words to the speech and visual signals to obtain prosodic and gestural features. All features are extracted with respect to windows around a word boundary; hence, no hand labeling of gesture or prosodic cues is required. Figure 1 depicts a variety of temporally synchronized multimodal features for a portion of the KDI data set.

The schematic diagram in Figure 2 adapted from [31] depicts our approach for combining lexical, prosodic, and gestural knowledge sources. Because the audio and video signals are time synchronized, the word time marks obtained by forced alignment of a speech transcript with the audio can be used to synchronize both the audio and gesture features with word boundaries. To model prosody and gesture, each word boundary has a corresponding vector of features. Given that E denotes the word boundary class sequence (SU or not), W denotes the corresponding word sequence, and F and G denote the corresponding prosodic and gestural features, the goal is to estimate $P(W, F, G, E)$ and then choose the boundary classifications that have the highest probability given the observed words and multimodal features:

$$\begin{aligned} arg\max_E P(E|W,F,G) &= arg\max_E P(W,F,G,E) \\ P(W,F,G,E) &= P(W,E)P(F,G|W,E) \\ &\approx P(W,E)P(F,G|E) \\ &= P(W,E)P(E|F,G)P(F,G)/P(E) \end{aligned}$$

A language model is used to determine $P(W, E)$. $P(F, G|W, E)$ is approximated by making the simplifying assumption that the gesture and prosody features of a boundary depend only on the boundary class. Although this assumption is not always true (e.g., the phonetic makeup of a word can affect

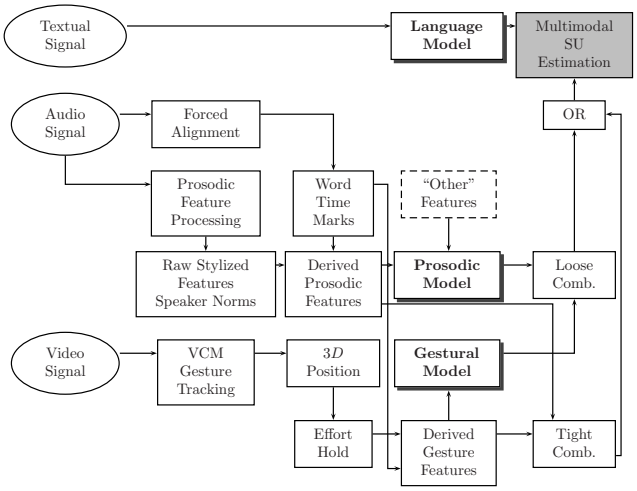


Figure 2: A schematic diagram of our system.

prosodic features), this independence assumption is a reasonable practicality. In the final step, we use Bayes rule to make use of a decision tree gesture and prosody model, which provides $P(E|F,G)$. Since $P(F,G)$ is a constant, we can ignore it when carrying the maximization.

Let W_i denote the the ith word, E_i denote the boundary event after W_i, and F_i and G_i denote W_i's prosodic and gestural features, respectively. Then prosody and gesture can be modeled by constructing separate decision tree classifiers that output posterior probability estimates $P(E_i|F_i)$ and $P(E_i|G_i)$, respectively, with the posterior probability $P(E_i|F_i,G_i)$ obtained as follows:

$$P(E_i|F_i,G_i) \approx \lambda P_{DT}(E_i|F_i) + (1-\lambda)P_{DT}(E_i|G_i) \quad (1)$$

Note that λ is set to a fixed value of 0.5 in these experiments since there is only the small amount of multimodal data available. It is also possible to train a single decision tree that uses both prosodic and gestural features to obtain $P(E_i|F_i,G_i)$ directly. As shown in Figure 2, both of these options are evaluated in this paper.

To calculate the argmax, we use the forward-backward algorithm for HMMs [29]. Training of the HMM is supervised since the SU-labeled data described in Section 2 is used. There are two sets of parameters to estimate. The state transition probabilities are estimated using a hidden event N-gram language model [34] described in the next subsection. The second set of HMM parameters are the observation likelihoods estimated given the prosodic and gesture features.

3.1 The SU Language Model (LM)

Words are a very useful knowledge source for the sentence segmentation task. These can be obtained from automatic recognition or from human transcripts. For this first investigation applying the direct modeling approach to gesture, we use human transcripts rather than ASR output. We chose to do so for several reasons. First, the focus of this paper is on the impact of gesture features in the sentence boundary task. Second, there is only a small amount of multimodal data and it was recorded in a noisy environment quite challenging for ASR (in insufficient quantities to train an ASR system). Third, if gesture is able to add beyond these other sources even when the other sources are perfect, this makes a strong case for the impact of gesture. Finally, previous studies involving SU detection in speech have shown that insights gained using human-generated transcripts generalize well to the ASR transcript case [18].

The word-level information is incorporated using a hidden-event word language model [31] that models the joint distribution of the SU event sequence E and the word string W, $P(W,E)$. This word/event LM is trained from the transcriptions, hand-labeled with the SU events. The N-gram LM parameter estimation optimizes the joint likelihood of $P(W,E)$.

In addition to the word-based hidden event LM, we also trained several class-based hidden event LMs. We used two different types of classes, part of speech (POS) tags and a set of automatically induced classes. The POS tagging of the word stream, obtained using the TnT tagger [4] trained using the Switchboard Treebank data, supports generalizations based on syntactic structure and smooths possibly undertrained word-based probability estimates. The automatically induced classes, obtained using the algorithm described in [5] from bigram word distributions, similarly supports generalization based on word usage patterns. Parameters for these class-based hidden event LMs are estimated from the joint class and event sequence $C_1\ E_1\ C_2\ E_2\ ...\ C_i\ E_i\ ...\ C_n\ E_n$. Linear interpolation is used to combine the probabilities from these LMs as in [18].

3.2 The Prosody Model

To model the prosody of sentence boundaries, we extract prosodic features around each word boundary, based on forced alignments of the transcripts to the audio. These features capture duration, pitch, and energy patterns in regions very near the word boundaries [32]. A crucial aspect of many of these features is that they are highly correlated (e.g., derived from the same raw measurements via various normalizations), real-valued (non-discrete), and in some cases undefined (e.g., unvoiced speech regions have no pitch). The prosodic features are modeled by a decision tree classifier that outputs posterior probability estimates $P(E_i|F_i)$, where E_i is the boundary event after W_i, and F_i is the corresponding prosodic feature vector. By using a decision tree as the probabilistic classifier, we can automatically select features that are most relevant to the task. Furthermore, the decision tree makes no assumptions about the shape of feature distributions and offers the distinct advantage of interpretability.

We briefly describe the prosodic features we investigate and how they are computed.

- **Duration Features**

  Pause duration after each word boundary is extracted based on the alignment of human transcriptions or recognition output. We also include the duration of the pause preceding the word before the boundary, to reflect whether speech right before the boundary is just starting up or is a continuation of previous speech. Phone durations are also computed. One possible indicator of an SU boundary in speech is preboundary lengthening, which typically affects the nucleus and coda of syllables. To capture such lengthening, we measure vowel and rhyme duration. We extract features such as the duration of the last vowel or the stressed vowel in a multisyllabic word, as well as their normalization.

- **F0 Features**

  To obtain F0 features, we first use an autocorrelation-based pitch tracker (get_f0 function in the ESPS package) to calculate frame-level F0 estimates. These raw F0 values are then post-processed to remedy some tracking errors, to use speaker-dependent parameters, and to simplify the F0 features. For each speaker, the F0 distribution is fitted to a lognormal tied mixture model (LTM) [33], whose mixture weights are found using an expectation maximization (EM) algorithm. The model returns an estimated pitch baseline value, which represents the lowest non-halved pitch value that we use for pitch normalization. We also apply a median filter to smooth voicing onsets for which the pitch tracker is unreliable. The frame level F0 values are then stylized to simplify tonal contours, shapes, and slopes. A piecewise linear fit (PWL) algorithm based on [33] is used to create line estimates for the median-filtered F0 values. On a particular voiced region, the PWL algorithm attempts to fit lines by minimizing the mean squared error between the linearized pitch estimates and the raw F0 values using a greedy algorithm. Using the stylized pitch contour, we compute several different types of F0 features:

  Range features: These features reflect the pitch range of a single word or window relative to the speaker-specific baseline F0 value computed in the LTM model. These include the minimum, maximum, mean, and last F0 values for each word boundary, excluding values which are unvoiced, halved, or doubled. These features are normalized by baseline F0 values using a linear difference, log difference, and log ratio. It is expected that speakers are more likely to fall near the bottom of their pitch range at a phrase, sentence, or topic boundary.

  Movement features: These features are obtained from the stylized F0 contours for the voiced regions of the word preceding and the word following a boundary. Examples of such movement features are the minimum, maximum, and mean F0 values, and the starting or ending stylized F0 values, using various normalization methods.

  Slope features: The stylized pitch values generate pitch slope within a word or a predefined length of window. We also consider the slope across a boundary to capture local pitch variation. A continuous trajectory is more likely to correlate with non-boundaries; whereas, a broken trajectory tends to indicate a boundary of some type.

- **Energy Features**

  Speakers tend to start an utterance loudly and taper off over time. We first generate the frame level RMS energy values (using the ESPS package), and then compute the minimum, maximum, and the mean RMS values over the word and over the voiced frames. As in stylized F0 processing, the raw energy values are fit to a linear model to capture the slope change of energy. We also compute the difference of energy values across a word boundary.

- **Additional Features**

  We include additional features, such as turn-related features and gender features, that may interact with aforementioned prosodic features (e.g., F0 features). Like all the prosodic features, these features can be automatically extracted from the speech data, using gender detection or speaker segmentation techniques. Turn-related features include whether or not there is a speaker change at a boundary, the time elapsed from the start of a turn, and the turn count within the current conversation.

3.3 The Gesture Model

Gestures are often decomposed into four phases: *preparation*, *stroke*, *hold* and *retraction*. Since gesture and speech in a coherent communication stem from the same mental process, one might expect a variety of useful gesture patterns that signal the beginning and ending of an SU (e.g., a hand retraction near the end of an SU, which marks the end of a thought or idea, as in Figure 1). In fact, using VisSTA, a multimodal signal visualization tool [26], we have observed, for example, that speakers often lift their hands from a rest position (e.g., on their lap) into the gesture space at the beginning of an SU and then retract their hands back to the area of rest near the SU end. There have been some past attempts to automatically segment gestures into phases [36], but these methods have not been found to be highly accurate for natural gestures. Human segmentation of gesture phase is also possible and potentially helpful; however, in this paper, we investigate whether, as in the prosody model, we can automatically extract and utilize vectors of gesture features to increase the accuracy of SU detection. Therefore, we obtain gesture features directly from measurements obtained from the video corpus without consideration of gesture phase.

To model gesture, we extract gestural features around each word boundary, based on forced alignments of a transcript to the audio, which is time synchronized to the video. These features can be modeled by a decision tree classifier that outputs posterior probability estimates $P(E_i|G_i)$, where E_i is the boundary event after W_i, and G_i is the corresponding gesture feature vector. Given our corpus, we use VCM to obtain the raw gesture features. Although this algorithm can provide $3D$ *hand position*, *Hold*, and *Effort*, we focus on the latter two features. Given the size of our corpus, the data sparsity of $3D$ *hand position* precludes its use in this initial study. In future work, we plan to investigate clustering techniques on $3D$ *hand position* to identify rest positions and gesture space positions. Hence, gestural features G_i at each inter-word boundary consists of a vector of numerical features reflecting *Effort* and *Hold* of the right and left hand of each communicant in our corpus.

- **Hold Features**

  Since audio pause duration has proven to be an effective feature in spoken SU detection [32], one might expect that the gestural correlate of a pause, namely a

Hold would be highly informative. Therefore, we define two features concerning the *Hold* duration around an inter-word boundary.

Hold overlap around an inter-word boundary: The ratio of *Hold* frames within the time interval marked by the beginning of the first word in an adjacent word pair up to the end of the second word.

Hold overlap with pauses: The ratio of *Hold* frames that overlap with an audio pause at an inter-word boundary is a potentially helpful combined prosody-gesture feature [12].

- **Effort Features**

  Using VisSTA, we have observed a common pattern of gesture phase changes around SU boundaries. Since *Effort*, the velocity of hand motion, is expected to change around transitions in gestural phase, one would expect that these features might be helpful for SU detection. Therefore, a series of gestural features are defined related with *Effort* around an inter-word boundary. Note that *Effort* values are normalized given the value range of the conversant so that the range is between 0.0 and 1.0.

  intra-window features: We calculate sets of features relative to a time window preceding or following a given inter-word boundary since gestures may slightly precede or follow a boundary. Two types of windows are used: the first is based on the duration of the preceding (or following) word, and the second on a fixed time window preceding or following the boundary. We set the time interval for the second type of window to 0.5 seconds in order to reflect the fact that gestures have a larger time granularity than a typical word's duration[1]. For each window type and perspective, we calculate the *minimum*, *maximum*, and *average* normalized *Effort*.

  inter-window features: We also calculate sets of features concerning the change in *Effort* between each pair of adjacent windows, i.e., between the values obtained for the window preceding and following an inter-word boundary. For example, *minmax* is the absolute difference between *minimum Effort* of the interval preceding the boundary and *maximum Effort* of the interval following it. Note that *minmin*, *minave*, *avemin*, *aveave*, *avemax*, *maxmin*, *maxave*, and *maxmax* are derived similarly.

4. SU DETECTION EXPERIMENT

4.1 Setup

As the KDI data set is of limited size, for the non-visual component models, our models may be better trained by utilizing the larger dataset that is available for the spoken SU boundary detection problem, namely the conversational telephone speech (CTS) dataset drawn from the Switchboard corpus that was annotated by LDC according to DARPA Rich Transcription for use in the Fall 2003 evaluation. This set contains about 40 hours of conversations that can be used to train the language model, as well as a conversational prosody model. There is some similarity between the CTS corpus and KDI multimodal corpus. First, they both involve conversational speech. Also these corpora have a similar SU percentage; 14% of the word boundaries are SU boundaries in CTS compared to 15% in the multimodal set. However, there are also some differences since the KDI multimodal corpus involves two people speaking face-to-face solving a problem; whereas, the CTS set involves telephone conversations between two people who are unknown to each other discussing a particular topic.

The small size of the KDI set precludes its use for LM training; hence, only the CTS data is used to train the hidden event language model (denoted LM_{CTS}). The word, POS, and automatically induced class based LMs are all trained using the NIST 2003 Rich Transcription (RT03) CTS training data set. We also utilized a word-based hidden event LM that was trained using the sentences and SU boundaries in the Switchboard Penn Treebank (which is marked with metadata according to Meteer's specification [20]). This second word-based model was added because the RT03 CTS set is somewhat small. The LMs are combined via linear interpolation, and the weights are optimized using the development set from the RT-03F MDE evaluation. We also train two different prosody models, one on the CTS data (denoted P_{CTS}) and one on the KDI dataset described in Table 1 (denoted P_{KDI}); however, due to the need for visual features, we train the gesture model only on the KDI set (denoted G_{KDI}). The weight for combining the LM with the CTS prosody model was set using the development set from the RT-03F MDE evaluation. The same weight is used to combine the LM with the KDI prosody and gesture models, as well as their combinations.

A standard CART style decision tree is used to train P_{CTS} since there is sufficient data to use this method. In constructing the gesture and prosody models using the KDI data, we chose to use Bayesian option style trees [7]. Option trees extend Bayes trees by growing many different trees and storing them in a compact form. They are a generalization of the standard tree where options are included at each point. At each interior node, instead of there being a single test and subtrees for its outcomes, there are several optional tests with respective subtrees. The resultant structure, which looks like an and-or tree, is a compact way of representing many different trees that share common prefixes. Option trees are grown using an N-ply lookahead such that the best few tests are grown as optional tests at the node. Although this method is more time and memory intensive, the improvement in prediction accuracy for a very small data set can be quite significant [7].

To evaluate our model, we use 10-fold cross validation in training and testing. For the prosody and gesture models, in order to address the imbalanced data problem (since there are many fewer SU events than nonevents at interword boundaries), we use a downsampled training set in which SU and non-SU have equal prior probabilities. Additionally, we employ ensemble bagging to reduce the variance of the prosodic classifier for the non-SU class. In this method several random downsampled training sets are generated, and each is resampled multiple times and corresponding classifiers are combined via bagging [19]. This has been found to improve the performance of the prosody model on the CTS SU task [19], so we also use the method for our prosody and gesture models. Note that the test portion of each fold represents the true distribution of SU and non-SU boundaries[2].

To evaluate the performance of our models, we use the Error Rate metric defined by NIST for the DARPA EARs

[1]This window size contains around fifteen gesture values. A smaller window would contain fewer values, which could lead to poorer estimates of the maximum and minimum. Note this 0.5 sec window is slightly longer than the window used in prosodic modeling (0.2 sec). Future research will consider the impact of window size.

[2]The posterior probabilities provided by the decision tree models are normalized to reflect the distribution of SU/non-SU boundaries in the training set.

metadata evaluation for comparison with the literature. To calculate the SU Error Rate, the estimated SU string is compared with the gold standard SU string to determine the number of misclassified boundaries per SU. Since SU boundaries may be incorrectly deleted or inserted, we also provide the Insertion Rate and Deletion Rate to determine whether there are different patterns of insertions and deletions among the different models. The Insertion Rate is the number of incorrect insertions of an SU in the estimated SU string that does not appear in the gold standard per SU boundary; whereas, the Deletion Rate is the number of incorrect deletions of an SU that appears in the gold standard string per SU boundary. The three metrics appear below:

1. *Error Rate* = (# *Deletion* + # *Insertion*)/ # SUs in the reference

2. *Insertion Rate* = # *Insertion*/ # SUs in the reference

3. *Deletion Rate* = # *Deletion*/ # SUs in the reference

4.2 Results and Discussion

The performance of each of the individual knowledge source models and their combination appears in Table 2. The first four lines in the table show the performance using each knowledge source individually, that is the hidden event LM trained on the CTS data (LM_{CTS}), the prosodic model trained on the CTS data (P_{CTS}), the prosodic model trained on the KDI data (P_{KDI}), and the gesture model trained on the KDI model (G_{KDI}). For comparison, the last line shows a baseline indicating the error that would be obtained by always selecting the majority class (i.e., non-SU). Note that the models that are trained using CTS data essentially ignore the KDI training data in each fold and simply generalize what they have learned from the CTS set to the KDI set. Each individual knowledge source performs better than the baseline in both deletion and in overall error rate. The prosody model trained from KDI data has a lower deletion rate and overall error rate compared to the prosody model trained from CTS data. This suggests that the use of prosody in the KDI dataset may be slightly different than in the CTS set. It should be noted that the prosody models for this task have a higher error rate than has been found in related research on SU detection in speech. This is most likely due to the fact that the KDI corpus was recorded in a fairly noisy environment with boom microphones that pick up the speech of both subjects. The language model has the highest overall error rate; however, as we will see this information adds significantly to the model combination performance. The gesture model trained on the KDI set has a slightly greater deletion and insertion error rate than the prosody model trained on the same set; however, its deletion rate is much lower than the prosody model trained on CTS. The gesture error rate does not differ significantly from either prosody model using the sign test.

The next three lines (five through seven) in the table show the performance of several combinations of the prosody and gesture models. We combine the separately trained prosody and gesture models using linear interpolation with a weight of 0.5 for each source[3]. We also trained a single decision tree that more tightly integrates the prosodic and gesture features on the KDI data. All of the interpolated models have a lower error rate than the respective prosody and gesture models alone, largely due to a reduction in the deletion rate. $P_{KDI} + G_{KDI}$ has a significantly lower error rate than G_{KDI}, suggesting that the prosody model adds information that is complementary to the gesture model. One might expect that the tight integration of the prosody and gesture features in a single decision tree (i.e., P,G_{KDI}) would obtain the best performance; however, the error rate is far worse than $P_{KDI} + G_{KDI}$. This result may indicate that the gesture and prosody features are redundant or too highly correlated for tight integration; however, it is far more likely that the different time granularities of the prosody and gesture features prevent them from combining well in a single model.

The eighth through the tenth lines in the table shows the impact of combining the LM with each of the individual prosody and gesture models. In all cases, the deletion rate drops, the insertion rate increases, and the overall error rate decreases significantly. As in prior studies investigating SU, the LM adds an important information source for detecting a sentence boundary. What is quite striking in this study is that, despite the fact that the LM was trained on a different corpus and performs more poorly than the other single knowledge source models, it is an important knowledge source in the combination models. Lines 11-13 in the table show the performance when each of the combination prosody/gesture models is combined with the LM. Overall, the three-way model combinations give a lower overall error rate than the single and pair-wise combinations. Note that adding LM_{CTS} to each of the combined prosody/gesture models also significantly lowers error rate. $LM_{CTS} + P_{KDI} + G_{KDI}$ achieves a lower overall error rate than the pairwise combination of the gesture or prosody model with the LM (i.e., $LM_{CTS} + P_{KDI}$ and $LM_{CTS} + G_{KDI}$); however, the differences are not statistically significant.

Model	*Deletion*	*Insertion*	Error
LM_{CTS}	49.16	17.51	66.67
P_{CTS}	47.31	9.76	57.07
P_{KDI}	38.89	14.31	53.20
G_{KDI}	40.40	17.68	58.08
$P_{CTS} + G_{KDI}$	44.61	10.61	55.22
$P_{KDI} + G_{KDI}$	37.21	14.31	51.52
P,G_{KDI}	40.24	14.78	54.71
$LM_{CTS} + P_{CTS}$	24.75	21.21	45.96
$LM_{CTS} + P_{KDI}$	21.21	22.90	44.11
$LM_{CTS} + G_{KDI}$	24.24	21.38	45.62
$LM_{CTS} + P_{CTS} + G_{KDI}$	23.03	21.55	44.61
$LM_{CTS} + P_{KDI} + G_{KDI}$	20.03	22.05	42.09
$LM_{CTS} + P,G_{KDI}$	20.88	22.22	43.10
Baseline	100	0	100

Table 2: Results for SU detection using gesture, prosody, and language models, alone or in combination. An error baseline obtained by always assuming a boundary is a member of the majority class appears in the last row of the table.

This investigation highlights the importance of using multiple knowledge sources to detect sentence boundaries in human dialogs. It also highlights the value of a direct modeling approach for constructing a model of conversational gesture. The gesture model improves upon the baseline and does not perform significantly better or worse than any of the other single knowledge source models. Furthermore, we find that gesture is complemented by the prosodic and lexical knowledge sources for this task. By using all three knowledge sources, the model is able to achieve the lowest overall error rate. This result has implications for the annotation of SU boundaries in multimodal corpora. Although we annotated the SU boundaries using only the transcriptions and audio files to prevent any positive bias toward gesture for this study, it is likely that annotators would effectively utilize

[3]We chose to simply average here rather than tune the weight since there is insufficient data to use as a heldout set.

the visual information provided by gesture as an additional cue for the presence of an SU boundary.

In future work, we will investigate a wider variety of gesture features. For example, since it is common for speakers to retract their hands to their laps when they complete an idea unit, we plan to incorporate *3D hand position* into the model. To use this cue effectively, we will investigate automatic methods of identifying the location of the lap. We also plan to increase the size of our multimodal corpus. One way to do this most easily with existing data is to evaluate the impact of using completely unsanitized visual features (currently, the tracked hand position is lightly human corrected which limits the amount of available data). We are also beginning to collect a meeting room corpus using Vicon data capture, which will make extraction of gesture features both fast and reliable. Finally, we will expand our work to other multimodal human communication tasks. A strong motivation for using the direct modeling approach is that in many cases we can utilize the same extracted features to attack a number of interesting problems such as topic boundary detection.

5. ACKNOWLEDGEMENTS

This work was supported in part by NSF under award number 9980054, ARDA under contract number MDA904-03-C-1788, DARPA under contract MDA972-02-C-0038, and Purdue Research Foundation. Part of this work was carried out while the third author was on leave at NSF. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of NSF, ARDA, or DARPA. We would like to express thanks to Francis Quek for providing us with the video data features, and David McNeill and Susan Duncan for their helpful discussions on the use of gesture in human communication.

6. REFERENCES

[1] http://vislab.cs.wright.edu/kdi/.

[2] http://www.darpa.mil/ipto/programs/ears/.

[3] P. Boersma and D. Weeninck. Praat, a system for doing phonetics by computer. Technical Report 132, University of Amsterdam, Inst. of Phonetic Sc., 1996.

[4] T. Brants. TnT-a statistical part-of-speech tagger. In *ANLP*, pages 224–231, 2000.

[5] P. F. Brown, V. J. D. Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, pages 467–479, 1992.

[6] R. Bryll, F. Quek, and A. Esposito. Automatic hand hold detection in natural conversation. In *IEEE Workshop on Cues in Communication*, Kauai,Hawaii, Dec 2001.

[7] W. Buntine. Learning classification trees. *Statistics and Computing*, 2:63–73, 1992.

[8] J. Cassell and M. Stone. Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems. In *AAAI*, 1999.

[9] J. Chai, P. Hong, and M. Zhou. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proc. of the 2004 International Conference on Intelligent User Interfaces (IUI-04)*, Madeira, Portugal, January 2004.

[10] L. Chen, M. P. Harper, and F. Quek. Gesture patterns during speech repairs. In *Proc. of the Fourth IEEE International Conference on Multimodal Interfaces(ICMI'02)*, Pittsburg, PA, Oct. 2002.

[11] S. Coquoz. Broadcast news segmentation using mde and stt information to improve s peech recognition. Technical report, International Computer Science Institute, 2004.

[12] A. Esposito, K. E. McCullough, and F. Quek. Disfluencies in gesture: Gestural correlates to speech silent and filled pauses. In *Proceeding of IEEE Workshop on Cues in Communication*, Kauai,Hawaii, 2001.

[13] S. Fels and G. Hinton. Glove-talk II - A neural-network interface which maps gestures to parallel formant speech synthesizer controls. *IEEE Transactions on Neural Networks*, 8:977–984, Sept. 1997.

[14] F.Quek, D. McNeill, R. Ansari, X. Ma, R. Bryll, S. Duncan, K.-E. McCullough, and C. Kirbas. Gestures cues for conversational interaction in monocular video. In *ICCV'99 Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pages 119–126, Corfu, Greece, 1999.

[15] D. Gibbon, B. Hell, K. Looks, and T. Trippel. Formal syntax of gesture : Cogest1.1. Technical report, Univ. of Bielefield, 2003.

[16] A. Kendon. Some relationships between body motion and speech. In A. W. Siegman and B. Pope, editors, *Studies in Dynamic Communication*. Pergamon, New York, 1972.

[17] S. Kettebekov, M. Yeasin, and R. Sharma. Prosody based co-analysis for continuopus recognition of coverbal gestures. In *International Conference on Multimodal Interfaces (ICMI'02)*, Pittsburgh USA, 2002.

[18] Y. Liu, A. Stolcke, E. Shriberg, and M. P. Harper. Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech. In *Proceedings of the Empirical Methods in Natural Language Processing*, 2004.

[19] Y. Liu, A. Stolcke, E. Shriberg, and M. P. Harper. Using machine learning to cope with imbalanced classes in natural speech: Evidence from sentence boundary and disfluency detection. In *Proceedings of the International Conference on Spoken Language Processing*, 2004.

[20] M. Mateer and A. Taylor. Disfluency annotation stylebook for the Switchboard corpus. Technical report, Department of Computer and Information Science, University of Pennsylvania, 1995.

[21] D. McNeil. *Hand and Mind: What Gestures Reveal about Thought*. Univ. Chicago Press, 1992.

[22] D. McNeill and S. Duncan. *Growth points in thinking-for-speaking*, chapter 7, pages 141–161. Cambridge Univ. Press, 2000.

[23] F. Quek, R. Bryll, and X. Ma. A parallel algorighm for dynamic gesture tracking. In *ICCV Workshop on RATFG-RTS*, Gorfu,Greece, 1999.

[24] F. Quek, M. P. Harper, Y. Haciahmetoglu, L. Chen, and L. Ramig. Speech pauses and gestural holds in Parkinson's disease. In *Seventh International Conference on Spoken Language Processing, ICSLP*, Denver,CO, Sept. 2002.

[25] F. Quek, D. McNeill, R. Bryll, S. Duncan, X. Ma, C. Kirbas, K. E. Mccullough, and R. Ansari. Multimodal human discourse: Gesture and speech. *ACM Transactions on Computer-Human Interaction*, 9(3), Sept. 2002.

[26] F. Quek, Y. Shi, C. Kirbas, and S. Wu. Vissta: A tool for analyzing multimodal discourse data. In *Seventh International Conference on Spoken Language Processing*, Denver,CO, Sept. 2002.

[27] F. Quek and Y. Xiong. Oscillatory gestures and discourse. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP2003*, Hong Kong, April 2003.

[28] F. Quek, Y. Xiong, and D. McNeill. Gestural trajectory symmetries and discourse segmentation. In *7th ICSLP*, Denver, CO, Sept. 2002.

[29] L. Rabiner and B. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[30] R. Sharma, J. Cai, S. Chakravarthy, I. Poddar, and Y. Sethi. Exploiting speech/gesture co-occurrence for improving continuous gesture recognition in weather narration. In *Proceedings of Automatic Face and Gesture Recognition*, 2000.

[31] E. Shriberg and A. Stolcke. Direct modeling of prosody: An overview of applications in automatic speech processing. In *International Conference on Speech Prosody*, 2004.

[32] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communications*, pages 127–154, 2000.

[33] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. In *Proceedings of the International Conference on Spoken Language Processing*, pages 3189–3192, 1998.

[34] A. Stolcke and E. Shriberg. Statistical language modeling for speech disfluencies. In *ICASSP*, 1996.

[35] S. Strassel. *Simple Metadata Annotation Specification*. Linguistic Data Consortium, 5.0 edition, 2003.

[36] A. Wilson, A. Bobick, and J. Cassell. Recovering the temporal structure of natural gesture. In *Automatic Face and Gesture Recognition*, pages 66–71, Oct. 1996.