SUBWORD UNIT REPRESENTATIONS FOR SPOKEN DOCUMENT RETRIEVAL[1]

Kenney Ng and Victor W. Zue

Spoken Language Systems Group
MIT Laboratory for Computer Science
545 Technology Square, Cambridge, MA 02139 USA
{kng, zue}@mit.edu

ABSTRACT

This paper investigates the feasibility of using subword unit representations for spoken document retrieval as an alternative to using words generated by either keyword spotting or word recognition. Our investigation is motivated by the observation that word-based retrieval approaches face the problem of either having to know the keywords to search for *a priori*, or requiring a very large recognition vocabulary in order to cover the contents of growing and diverse message collections. In this study, we examine a range of subword units of varying complexity derived from phonetic transcriptions. The basic underlying unit is the phone; more and less complex units are derived by varying the level of detail and the length of sequences of the phonetic units. We measure the ability of the different subword units to effectively index and retrieve a large collection of recorded speech messages. We also compare their performance when the underlying phonetic transcriptions are perfect and when they contain phonetic recognition errors.

1. INTRODUCTION

With the explosion of available information spurred on by advances in information technologies, the need for automatic methods to process, organize, and analyze this data and present it in human usable form has become increasingly important. Of particular interest is the problem of efficiently finding "interesting" pieces of information from among the rapidly growing streams and collections of data.

Much research has been done, under the headings of document and text retrieval on the problem of selecting "relevant" items from a large collection of *text* documents given a request from a user [9]. Only recently has there been work addressing the retrieval of information from other media such as images, audio, and speech [3,6,7,13]. Given that increasingly large portions of the available data are made up of spoken language information, such as recorded speech messages and radio/television broadcasts, the development of automatic methods to index and retrieve spoken documents will become more important.

One approach to this problem is to perform keyword spotting on the spoken documents to obtain a representation in terms of a set of keywords [3]. In order to do this the set of keywords needs to be chosen *a priori*. This requires advanced knowledge about the content of the speech messages or what the possible user queries may be. Alternatively, the keywords can be determined after the user specifies the query. In this case, however, the user needs to wait while the message collection is searched which may lead to unacceptable delays in the response time.

Another approach is to first transform the spoken documents into text using a large vocabulary speech recognizer and then use a full-text retrieval system [6]. Although straightforward, there are some problems with this approach. One major issue is the growth of the recognizer vocabulary needed to handle new words from growing and diverse message collections. With current technology, there is a practical computational limit on the size of the recognition vocabulary; there is also the issue of determining when, how, and what new words need to be added.

An alternative approach that has the potential of dealing with many of the above problems is to use subword unit representations for spoken document retrieval. The use of subword units in the recognizer constrains the size of the vocabulary needed to cover the language; and the use of subword unit indexing terms allows for new query term detection during retrieval. Although there is a tradeoff between the size of the unit and speech recognition accuracy and index term discrimination, some of this can be mitigated by the choice of the subword units and the modeling of their sequential constraints. The effectiveness of relatively simple text retrieval algorithms that essentially match word fragments gives us hope that subword approaches can work with spoken documents.

Several subword based approaches have been recently proposed in the literature. One makes use of special syllable-like units derived from text [11] and others are based on post-processing the output of a phonetic recognizer [7,13]. However, there has been no study to date that explores the space of possible subword unit representations to determine the complexity of the subword units needed to perform effective spoken document retrieval and to measure the sensitivity of the different units to speech recognition errors.

This paper investigates the feasibility of using subword unit representations for spoken document retrieval. We examine a range of subword units of varying complexity derived from phonetic transcriptions and measure their ability to effectively index and retrieve a large collection of recorded speech messages. In addition, we compare the performance of the different subword units when the underlying phonetic transcriptions are perfect and when they contain phonetic recognition errors.

[1] This research was supported by a research contract from NYNEX Science and Technology.

2. SUBWORD UNITS

A range of subword units of varying complexity derived from phonetic transcriptions is explored. The basic underlying unit of representation is the phone. More and less complex subword units are derived by varying the complexity of these phonetic units in terms of their level of detail and sequence length. For level of detail, we look at labels ranging from specific phone classes to broad phonetic classes. For sequence length, we look at automatically derived fixed- and variable-length sequences ranging from one to six units long; in addition, sequences with and without overlapping units are also examined. Since it is difficult to obtain word and sentence boundary information from phonetic transcriptions, all subword units are generated by treating each message/query as one long phone sequence with no boundary information.

2.1. Phone Sequences

The most straightforward subword units that we examine are overlapping, fixed-length, phonetic sequences (*phone*) ranging from n=1 to n=5 phones long; a phone inventory of 41 classes is used. These subword units are derived by successively concatenating together the appropriate number of phones from the phonetic transcriptions. Examples of n=1 and n=3 phone sequence subword units for the phrase "weather forecast" are given in Table 1. For large enough n, we see that cross-word constraints can be captured by these units (e.g., dh_er_f, er_f_ow).

2.2. Broad Phonetic Class Sequences

In addition to the original phone classes, we also explore more general groupings of the phones into broad phonetic classes (*bclass*) to investigate how the specificity of the phone labels (level of detail) impacts performance. The broad classes are derived via unsupervised hierarchical clustering of the 41 original phones using acoustic measurements derived from the TIMIT corpus [4]. The goal of the clustering is to group acoustically similar phones into the same class. Figure 1 shows the resulting cluster tree and the "cuts" used to derive three different sets of broad classes with 20, 14, and 8 distinct classes. Examples of some broad class subword units (class c=20, length n=4) are given in Table 1.

Subword Unit	Indexing Terms
word	weather forecast
phone (n=1)	w eh dh er f ow r k ae s t
phone (n=3)	w_eh_dh eh_dh_er dh_er_f er_f_ow f_ow_r ow_r_k r_k_ae k_ae_s ae_s_t
bclass (c=20, n=4)	liquid_frntvowel_voicefric_retroflex frntvowel_voicefric_retroflex_weakfric voicefric_retroflex_weakfric_···
mgram (m=4)	w_eh_dh_er f_ow_r k_ae_s_t
sylb	w_eh dh_er f_ow_r k_ae_s_t

Table 1: Indexing terms for different subword units.

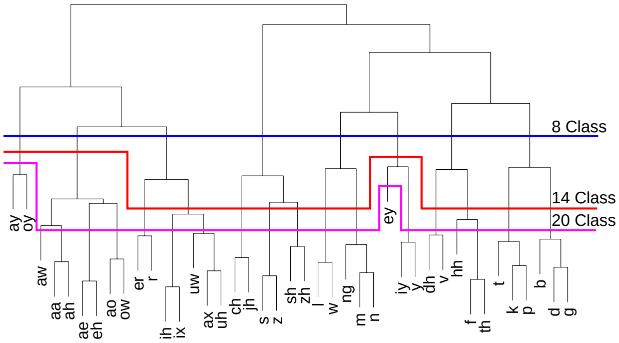

Figure 1: Clustering tree for phonetic broad classes.

2.3. Phone Multigrams

We also examine non-overlapping, variable-length, phonetic sequences (*mgram*) discovered automatically by applying an iterative unsupervised learning algorithm previously used only in developing "multigram" language models for speech recognition [1]. The multigram model assumes that a phone sequence is composed of a concatenation of independent, non-overlapping, variable-length, phone subsequences (with some maximum length m). The likelihood of the entire sequence is computed as the product of the likelihoods of the individual subsequences. In this model, the parameters are the set of subsequences and their likelihood values. Both of these can be trained automatically via maximum likelihood (ML) estimation using the estimate-maximize (EM) algorithm. Given a set of trained model parameters, a Viterbi-type search can then be used to generate a ML segmentation of an input phone sequence to give the most likely set of non-overlapping, variable-length, phone subsequences. The multigram model, with m=1,...,5, was trained on and then used to process the speech message collection. Examples of some multigram (m=4) subword units are given in Table 1.

2.4. Syllable Units

We also consider linguistically motivated syllable units (*sylb*) composed of non-overlapping, variable-length, phone sequences generated automatically by rule. The rules take into account English syllable structure constraints and allow for ambisyllabicity [2,8]. Syllabic units were generated for the speech messages and queries using these rules. Examples of some syllabic subword units are given in Table 1.

3. RETRIEVAL MODEL

A standard information retrieval (IR) model based on the SMART system [9, 10] is used. The documents and queries are represented as vectors with simple term and collection frequency weightings. The weight of term i in the vector for document j is:

$$\mathbf{d}_j[i] = \log(f_j[i] + 1)$$

and the weight of term i in the vector for query k is:

$$\mathbf{q}_k[i] = \log(f_k[i] + 1)\log(N/n_i)$$

where $f_j[i]$ is the frequency of term i in document or query j, n_i is the number of documents containing term i, and N is the total number of documents in the collection. The second term is the inverse document frequency (idf) for term i. A normalized inner product similarity measure between document $\mathbf{d}_j$ and query $\mathbf{q}_k$ is used for retrieval:

$$S(\mathbf{d}_j, \mathbf{q}_k) = \frac{\mathbf{d}_j \cdot \mathbf{q}_k}{||\mathbf{d}_j||\ ||\mathbf{q}_k||}$$

4. EXPERIMENTS AND RESULTS

4.1. Data Set

The data set used in this work is composed of approximately ten hours of speech recorded from National Public Radio (NPR) radio news broadcasts [12]. The speech is transcribed at the word level and partitioned into 384 separate news stories of varying lengths according to topic with an average of 325 words per message. In addition, a set of 50 natural language queries and associated relevance judgments on the message collection were created to support retrieval experiments. Each query has an average of 4.5 words and 6.2 relevant messages. Although this data set is small in comparison to experimental text retrieval collections [5], it is comparable to data sets used in other speech retrieval experiments [6, 7, 13].

4.2. Baseline Text Retrieval

To provide a baseline for comparison, retrieval using word-level text transcriptions (*word*) of the spoken documents and queries is performed. This baseline is equivalent to using a perfect word recognition system to transcribe the speech messages followed by application of a standard text retrieval system. The retrieval performance of this configuration, measured in *non-interpolated average precision* (as used in TREC [5]), is p=0.87. Compared to results on large text collections, this performance number is very high and indicates that this task is relatively straightforward. This is due, in part, to the relatively small number and concise nature of the speech messages.

4.3. Performance of Subword Units

Experiments that examine the feasibility of using the different subword units for indexing and retrieval are performed using phonetic expansions of the words in the messages and queries obtained via a pronunciation dictionary. This experiment provides an *upper bound* on the performance of the different subword units since it assumes that the underlying phonetic recognition is perfect. It can also be used to eliminate poor subword units from further consideration.

Retrieval performance for the different subword units, measured in average precision, is shown in Figure 2A. We can make several observations. First, as the length of the sequence is increased, performance improves, levels off, and then slowly declines. This is because as the sequence becomes longer the units begin to approximate words and short phrases, but after a certain length they become too specific and more difficult to match. Second, overlapping subword units (*phone*, n=3, p=0.86) perform better than non-overlapping units (*mgram*, m=4, p=0.81). Units with overlap provide more chances for partial matches and, as a result, are more robust to variations in the phonetic realization of the words. Third, between the two non-overlapping subword units (*mgram* and *sylb*), the automatically derived multigram units (m=4, p=0.81) perform better than the rule-based syllable units (p=0.76) when no word boundary information is used. If the word boundaries are given, then improved syllabic units are generated and retrieval performance improves from p=0.76 to 0.82 (not plotted). Fourth, even after collapsing the number of phones from 41 down to 20 broad classes, enough information is preserved to perform reasonable retrieval (*bclass*, c=20, n=4, p=0.82). From this experiment, we see that it is possible, with the appropriate choice of subword units, to achieve performance approaching that of text-based word units if the underlying phonetic units are recognized correctly.

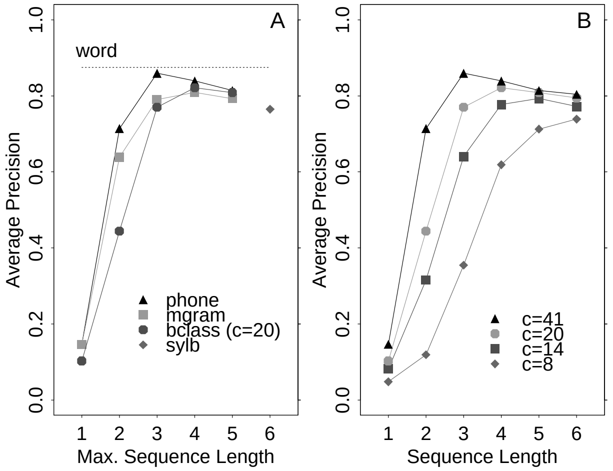

Figure 2: (A) Retrieval performance of the different subword units with perfect underlying phonetic transcriptions. (B) Performance of broad phonetic class subword units with varying number of classes and sequence lengths.

Subword Unit	Top N=5 Stop Terms
phone (n=3)	ae_n_d ax_n_t sh_ax_n dh_ae_t f_ao_r
mgram (m=4)	dh_ax ae_n_d ix_nx t_uw ax_n
sylb	dh_ax t_uw ax t_ax t_iy

Table 2: Examples of automatically derived stop terms.

Figure 2B shows the retrieval performance of broad class subword units for a varying number of classes. We see that there is a tradeoff between the number of broad classes and the sequence length required to achieve good performance. It is interesting to note that with 8 broad classes, only sequences of length 5 or 6 are needed to obtain reasonable retrieval performance. This result indicates that there is a lot of phonological constraint in the English language that can be captured at the broad class level.

4.4. Removal of Subword "Stop" Units

A standard IR technique that has been shown to improve performance is to remove frequently occurring, non-content words ("stop words") from the set of indexing terms [9]. We briefly explored several methods to automatically discover and remove "stop terms" from the different subword unit representations to try to improve performance. We used both term and inverse document frequencies to rank order and select the top $N = 5, \ldots, 200$ indexing terms as the "stop terms" to remove. A list of the top five terms for the phone (n=3), multigram (m=4), and syllable subword units is shown in Table 2. We see that they mainly consist of short function words and common prefixes and suffixes. We found that although removing these terms does improve retrieval performance, the gain was very small (less than 1%). So for simplicity, "stop term" removal was not done in any of the subword unit experiments.

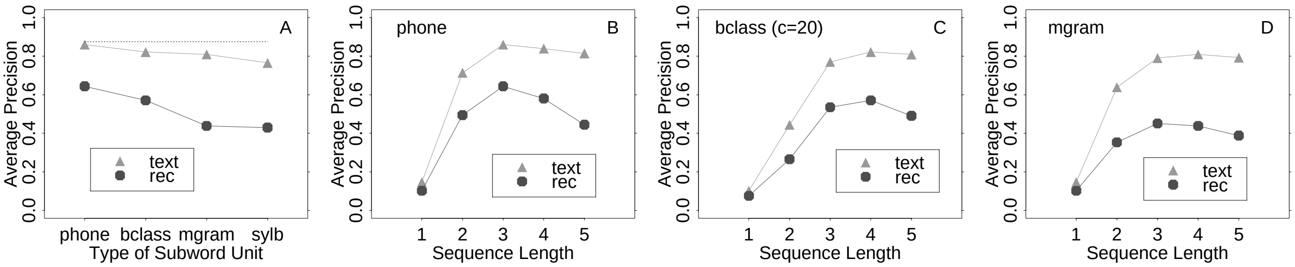

Figure 3: Performance of different subword units with perfect (text) and errorful (rec) phonetic transcriptions.

4.5. Robustness of Subword Units

Since automatic phonetic recognition of the speech messages will not be perfect, we next examine the sensitivity of the subword units to errorful phonetic transcriptions. In this experiment, recognition errors are generated by simulation based on phonetic recognition error statistics (23.3% substitution, 4.5% insertion, and 7.8% deletion) derived by running the SUMMIT recognizer on speech from the TIMIT corpus [4]. Errors are introduced only into the message collection; the queries are not corrupted. The performance reported is the average of five independent random simulation runs.

Retrieval performance for the various subword units (*phone*, n=3; *bclass*, c=20, n=4; *mgram*, m=4; and *sylb*) is shown in Figure 3A for perfect (text) and errorful (rec) phonetic transcriptions. We observe that the overlapping subword units (*phone*, *bclass*) are less sensitive to the errors than the non-overlapping units (*mgram*, *sylb*). There are two contributing factors. One is the robustness of the overlapping units to variations because they allow for more partial matching. The other is that the multigram and syllable algorithms, which discover their units from the phone stream, have a more difficult time finding regularized units when there are errors.

Figures 3B,C,D show, in more detail, the performance of the phone, broad class, and multigram subword units with perfect (text) and errorful (rec) phonetic transcriptions. In all cases, we see that as the sequences get longer, performance falls off faster in the errorful case. This is because more errors are being included which leads to more term mismatches. Another observation is that the broad class units out perform the phone units for longer (n=4, 5) sequences. This is because there are fewer broad class errors than phone errors due to the collapsed number of classes. In fact, the broad class (c=20) error rate is 29.0% compared to 35.6% for the original 41 classes. From this experiment, we see that although retrieval performance is worse for all units when there are recognition errors, some subword units can still give reasonable performance even before using any error compensation techniques such as approximate term matching.

5. CONCLUSION

In this paper, we explore a range of subword units of varying complexity and measure their ability to effectively index and retrieve speech messages. We find that with the appropriate subword units it is possible to achieve performance comparable to that of text-based word units if the underlying phonetic units are recognized correctly. In the presence of phonetic recognition errors, performance degrades but many subword units can still achieve reasonable performance. These results indicate that subword-based approaches to spoken document retrieval are feasible and merit further research. We are currently expanding the message collection and processing the speech messages with our phonetic recognizer to generate more realistic phonetic transcriptions. In addition, we are investigating robust indexing and retrieval methods in an effort to improve retrieval performance when there are phonetic recognition errors. These include approximate term matching approaches which allow for substitution, deletion, and insertion errors and approaches that capture multiple recognition hypotheses which allow for extended matching.

6. REFERENCES

[1] S. Deligne and F. Bimbot, "Language Modeling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams", ICASSP 1995, pp. 169-175.

[2] W. Fisher, Automatic syllabification program based on algorithm in [8]. Available at ftp://jaguar.ncsl.nist.gov/pub/tsylb2-1.1.tar.Z.

[3] J.T. Foote, *et. al*, "Talker Independent Keyword Spotting for Information Retrieval", Eurospeech 1995, pp. 2145-2148.

[4] J. Glass, *et. al*, "A Probabilistic Framework for Feature-Based Speech Recognition," ICSLP 1996.

[5] D. Harman, "Overview of the Fourth Text REtrieval Conference (TREC-4)" NIST Special Publication 500-236, Gaithersburg, MD.

[6] A.G. Hauptmann and H.D. Wactlar "Indexing and Search of Multimodel Information", ICASSP 1997, Vol. 1, pp. 195-198.

[7] D.A. James, "The Application of Classical Information Retrieval Techniques to Spoken Documents", Ph.D. Thesis, University of Cambridge, UK, 1995.

[8] Daniel Kahn, "Syllable-based Generalizations in English Phonology", Ph.D. Thesis, M.I.T., 1976.

[9] G. Salton and M. McGill, "Introduction to Modern Information Retrieval", McGraw-Hill, NY, 1983.

[10] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval", Info. Processing and Management, 24(5), pp. 513-523, 1988.

[11] P. Schauble and U. Glavitsch, "Assessing the Retrieval Effectiveness of a Speech Retrieval System by Simulating Recognition Errors", Proc. of ARPA HLT Conference, pp. 370-372. 1994.

[12] M. Spina and V. Zue, "Automatic transcription of general audio data: preliminary analysis," ICSLP 1996.

[13] M. Wechsler and P. Schauble, "Indexing Methods for a Speech Retrieval System", MIRO Workshop 1995.