PHONETIC RECOGNITION FOR SPOKEN DOCUMENT RETRIEVAL[†]

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 describes the development and application of a phonetic recognition system to the task of spoken document retrieval. The recognizer is used to generate phonetic transcriptions of the speech messages which are then processed to produce subword unit representations for indexing and retrieval. Subword units are used as an alternative to words units generated by either keyword spotting or word recognition. We first investigate the use of different acoustic and language models in the speech recognizer in an effort to improve phonetic recognition performance. Then we examine a variety of subword unit indexing terms and measure their ability to perform effective spoken document retrieval. Finally, we look at some simple robust indexing and retrieval methods that take into account the characteristics of the recognition errors in an attempt to improve retrieval performance.

1. INTRODUCTION

As the amount of accessible data continues to grow, 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 the growing collections and streams of data. Much research has been done on the problem of selecting "relevant" items from large collections of text documents given a request from a user [2]. Only recently has there been work addressing the retrieval of information from other media such as images, video, audio, and speech [3, 5, 6, 9, 10]. Given the growing amounts of spoken language data, such as recorded speech messages and radio and television broadcasts, the development of automatic methods to index, organize, and retrieve spoken documents will become more important.

In our previous work [6], we investigated the feasibility of using subword unit indexing terms for spoken document retrieval as an alternative to words generated by either keyword spotting or word recognition. The investigation was 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. 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 the detection of new user-specified query terms during retrieval. Indexing terms attempt to represent the content of a speech message in an efficient format much like the entries in the index of a textbook describe the book's contents.

[†]This research was supported by research contracts from DARPA (N66001-96-C-8526) and NYNEX Science and Technology.

We examined a range of subword units of varying complexity derived from phonetic transcriptions and found that subword unit indexing terms are able to capture enough information to perform effective retrieval. With the appropriate subword units it was possible to achieve performance comparable to that of text-based word units if the underlying phonetic units were recognized correctly. In the presence of phonetic recognition errors, retrieval performance degraded but many subword units were still able to achieve reasonable performance even without the use of robust methods such as approximate matching.

In the above work, the phonetic recognition errors introduced into the spoken document collection, although modeled after the behavior of a real phonetic recognizer, were *simulated*. In this work, we train a phonetic recognizer and run it on the entire spoken document collection to generate phonetic transcriptions which are then processed to produce subword unit representations for indexing and retrieval. In this way, we can measure retrieval performance with *real* phonetic recognition output.

In the following sections, we give a brief description of the subword unit indexing terms explored, describe the information retrieval model, the speech recognition system, and the speech corpus used. We then present experiments that try to improve phonetic recognition performance by examining different acoustic and language models. Next, we present retrieval experiments that examine the performance of various subword unit indexing terms derived from the phonetic recognition output and also look at some simple robust indexing and retrieval methods.

2. SUBWORD UNIT REPRESENTATIONS

A range of subword unit indexing terms of varying complexity derived from the phonetic recognition output is explored. For consistency, the units are the same as those examined in our previous work [6]. The basic underlying unit is the phone; more and less complex units are derived by varying the level of detail and the sequence length of these units. Labels range from specific phones to broad phonetic classes. Automatically derived fixed- and variable-length sequences ranging from one to six units long are examined. Also, sequences with and without overlap are explored. In generating the subword units, each message/query is treated as one long phone sequence with no word or sentence 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 in length; a phone inventory of 41 classes is used. These subword units are derived by successively concatenating 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.

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_$\cdots$
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: Examples of indexing terms for different subword units.

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 original phones based on acoustic similarity. Broad class sets of size c=20, 14, and 8 are examined. Examples of some broad class subword units (class c=20, length n=4) are given in Table 1.

2.3. Phone Multigrams

We also look at non-overlapping, variable-length, phonetic sequences (*mgram*) discovered automatically by applying an iterative unsupervised learning algorithm previously used in developing "multigram" language models for speech recognition. The algorithm finds the set of non-overlapping phonetic sequences of a specified maximum length, m, that most likely account for the observations in the document 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. Examples of some syllabic subword units are given in Table 1.

3. INFORMATION RETRIEVAL MODEL

A standard vector space information retrieval (IR) model is used in the experiments [7]. In this model, the documents and queries are represented as vectors where each component in the vector is an indexing term. A term can be a word, word fragment, or in our case a subword unit. Each term has an associated weight based on the term's occurrence statistics both within and across documents. The weight of term i in the vector for document j is:

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

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

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

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 to score and rank the documents during retrieval:

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

4. PHONETIC RECOGNIZER

The MIT SUMMIT speech recognizer is used in this work [1]. It is a probabilistic segment-based approach that differs from conventional frame-based hidden Markov model (HMM) approaches. The recognizer uses context independent segment and context dependent boundary (segment transition) acoustic models. Acoustic feature vectors consisting of Mel-frequency cepstral coefficients (MFCCs), difference cepstra, energy, and duration are derived from the speech signal and used in the segment and boundary models. The distribution of the acoustic features are modeled using mixtures of diagonal Gaussians. A two pass search strategy is used during recognition. A forward Viterbi search is performed using a statistical bigram language model followed by a backwards *A** search using a higher order statistical n-gram language model.

5. SPEECH DATA CORPUS

The speech data used in this work consists of FM radio broadcasts of the NPR "Morning Edition" news show [8]. The data is recorded off the air, orthographically transcribed, and partitioned into separate news stories. The data is broken up into two sets, one for training and tuning the speech recognizer and another for use as the spoken document collection for the retrieval experiments.

The recognizer training set consists of 2.5 hours of clean speech from 5 shows and the development set consists of one hour of data from one show. In other experiments [8], it was found that training on speech from all noise conditions (noisy, telephone/field, music, etc.) does not significantly improve performance over training on only the clean speech. As a result, only clean speech is used for training in the following phonetic recognition experiments.

The spoken document collection consists of 12 hours of speech from 16 shows partitioned into 384 separate news stories. Each story averages 2 minutes in duration and typically contains speech from multiple noise conditions. A set of 50 natural language text queries and associated relevance judgments on the message collection are created to support retrieval experiments. The queries are created from the story "headlines" and are relatively short, each averaging 4.5 words. Each query has an average of 6.2 relevant documents. Although this data set is small in comparison to experimental text retrieval collections [2], it is comparable to data sets used in other speech retrieval experiments [3, 5, 9].

6. EXPERIMENTS AND RESULTS

6.1. Phonetic Recognition Experiments

A series of phonetic recognition experiments is performed exploring the effects of using different acoustic and language models to try to improve phonetic recognition performance.

6.1.1. Segment Acoustic Model

The most basic phonetic recognition system uses 61 context-independent acoustic models corresponding to the TIMIT phone labels. Performance, in terms of phonetic recognition error rate, is measured on a collapsed set of 39 classes typically used in reporting phonetic recognition results [8]. Results on the development set for speech from all noise conditions (*entire*) and from only the clean condition (*clean*) are shown in Table 2.

6.1.2. Boundary Acoustic Model

Boundary models are context-dependent acoustic models that try to model the transitions between two adjacent segments. They are used in conjunction with the segment models and provide more information to the recognizer. The use of boundary models significantly improves recognition performance as shown in Table 2.

Model	seg	+bnd	+agg	n=3	n=4	n=5
Dev (clean)	35.0	29.1	27.9	27.3	26.7	26.2
Dev (entire)	43.5	37.7	36.9	36.2	35.5	35.0

Table 2: Phonetic recognition error rate (%) using various models tested on the entire development set and on only the clean portions.

6.1.3. Aggregate Acoustic Models

Since the EM algorithm used to train the acoustic models makes use of random initializations of the parameter values and only guarantees convergence to a local optima, different models can result from different training runs using the same training data. An interesting question is then how to select the "best" model resulting from multiple training runs. It turns out that aggregating or combining the different models into a single larger model results in better performance than selecting just one of the models based on methods such as cross validation [4]. We aggregate five separate acoustic models trained using different random initializations and observe a performance improvement.

6.1.4. Language Model

The above recognizers use a statistical bigram language model to constrain the forward Viterbi search during decoding. More detailed knowledge sources, such as higher order n-gram language models, can be applied by running a second pass, backwards *A**, search. We examine n-grams of order n=3, 4, and 5 and observe that recognition performance improves as n increases.

The final phone error rate on the development set is 35.0%. To see if this performance is indicative of that on the speech document collection, three hours of the speech messages are phonetically transcribed, processed with the recognizer, and evaluated. A phone error rate of 36.5% is obtained, indicating a good match between the data in the development set and the message collection.

6.2. Information Retrieval Experiments

We examine a range of subword unit indexing terms of varying complexity derived from the phonetic recognition output (as described in Section 2) and measure their ability to perform effective spoken document retrieval. We compare it to using perfect phonetic transcriptions obtained via a pronunciation dictionary and to using word-level text transcriptions. We also examine some simple robust indexing and retrieval methods that attempt to improve retrieval performance by taking into account and compensating for the phonetic recognition errors.

6.2.1. Baseline Text Retrieval

A baseline text retrieval run is performed using word-level text transcriptions (*word*) of the spoken documents and queries. This is equivalent to using a perfect word recognizer to transcribe the speech messages followed by a full-text retrieval system. Retrieval performance, measured in *non-interpolated average precision* (as used in TREC [2]), is p=0.87. This number is high compared to text retrieval performance using very large document collections [2] 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. The baseline text performance is plotted in the figures using a dotted line.

6.2.2. Perfect Phonetic Transcriptions

An upper bound on the performance of the different subword unit indexing terms is obtained by running retrieval experiments using phonetic expansions of the words in the messages and queries obtained via a pronunciation dictionary. This experiment was done in our previous work [6] and we found that many of the subword unit indexing terms are able to capture enough information to perform effective retrieval. 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. This performance is plotted using dashed lines.

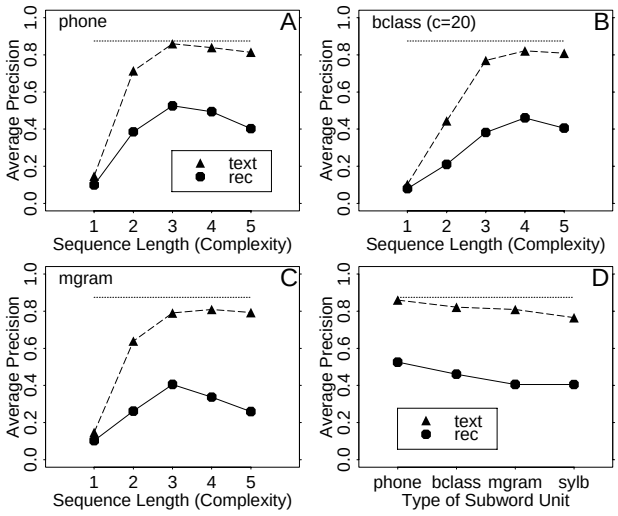

Figure 1: Performance of various subword unit indexing terms with perfect (text) and errorful (rec) phonetic transcriptions.

6.2.3. Errorful Phonetic Transcriptions

We next examine the retrieval performance of the subword unit indexing terms derived from errorful phonetic transcriptions created by running the phonetic recognizer on the entire spoken document collection. Figures 1A,B,C, show the retrieval performance, measured in average precision, of the phone, broad class, and multigram subword units with perfect (*text*) and errorful (*rec*) phonetic transcriptions. We can make several observations. First, as the length of the sequence is increased, performance improves, levels off, and then declines for all cases. As the sequence becomes longer the units begin to approximate words and short phrases, but after a certain length they become too specific. Second, as the sequences get longer, performance falls off faster in the errorful case than in the perfect case. This is because more errors are being included in the errorful case which leads to more term mismatches. Finally, in the errorful case, broad class units are slightly better than phone units for longer (n=4, 5) sequences. It turns out that 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% versus 36.5% for the original set of classes.

Retrieval performance for selected subword units (*phone*, n=3; *bclass*, c=20, n=4; *mgram*, m=4; and *sylb*) is shown in Figure 1D for perfect (*text*) and errorful (*rec*) phonetic transcriptions. We note 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, are able to find fewer consistent or regularized units when there are errors.

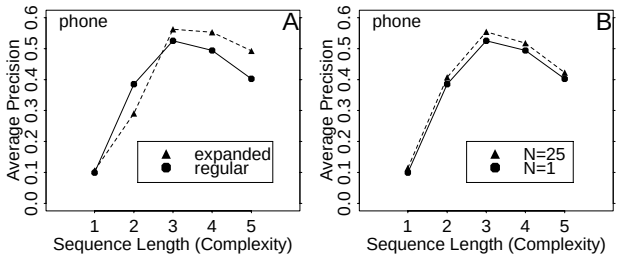

Figure 2: Performance of phone subword units of varying lengths with and without query (A) and document (B) expansion.

The results from this experiment closely match those obtained in our previous work [6] where the errorful phonetic transcriptions were generated by simulation. From this experiment, we see that although performance is worse for all units when there are phonetic recognition errors, some subword units can still give reasonable performance even before the use of any error compensation techniques such as approximate term matching.

6.2.4. Robust Indexing and Retrieval

In our next set of experiments, we try to see if we can improve retrieval performance by using "robust" indexing and retrieval approaches which take into account and try to compensate for the speech recognition errors introduced into the spoken document collection. We look at two simple approaches to try to compensate for the recognition errors. One involves modifying the query representation and the other the document representation.

In the first approach, we modify the query representation to include additional approximate match terms. The main idea is to include terms that are likely to be confused with the original query terms. The approximate terms are determined using information from the phone error confusion matrix derived from running the phone recognizer on the development data set. The terms are weighted based on how confusible they are to the original terms; the closer they are to the original term, the higher the weight. The hope is that some of the newly added terms will match the corrupted terms in the document collection.

Another approach is to modify the speech document representation by expanding them to include multiple phone recognition candidates. Basically, use the top N instead of just the top one recognition hypothesis. The main idea here is to include high scoring recognition alternatives in the document representation. This should increase the chance of the correct hypothesis being included in the representation. The terms are weighted according to the number of times they occur in the top N hypotheses. The more often a term appears in the top N hypotheses, the more confident we are that it actually occurred, and the higher the weight.

Figure 2A shows retrieval performance in average precision for phone subword units of varying sequence length with and without query expansion. Regular queries are compared with modified queries that include approximate match terms. We note that performance with the shorter sequences (n=1,2) is worse when using the expanded queries. This is likely due to spurious matches from the extra query terms generated during the expansion. However, the longer phone sequences (n=3,4,5), which dropped off in performance due to more accumulated errors and fewer matches in the original case, are much improved using the expanded queries. More approximate terms are being matched while the longer sequence length makes it more difficult to get spurious matches.

Retrieval performance in average precision for phone subword units of varying sequence length for the case of modifying the spoken document representation is shown in Figure 2B. We compare using just the top one phone recognition hypothesis (the original approach) to using the top N=25 hypotheses. We see that using multiple recognition hypotheses in the document representation consistently improves performance, albeit by a small amount. The performance difference at length n=3 is statistically significant at the 0.002 level using the significance test specified in [7].

7. CONCLUSION

In this paper, we train and tune a phonetic recognizer to operate on radio broadcast news data and use it to process the entire spoken document collection to generate phonetic transcriptions. We then explore a range of subword unit indexing terms of varying complexity derived from the phonetic recognition output and measure their ability to perform effective spoken document retrieval. We find that in the presence of phonetic recognition errors, retrieval performance degrades, as expected, compared to using perfect phonetic transcriptions or word-level text units. However, many subword unit indexing terms can still give reasonable performance even before using any error compensation techniques such as approximate term matching. We also examine some simple robust indexing and retrieval methods that take into account the characteristics of the recognition errors and saw that they can help improve retrieval performance.

These results indicate that subword-based approaches to spoken document retrieval are feasible and merit further research. In terms of current and future work, we are expanding the corpus to include more speech for both recognizer training and the speech message collection; exploring ways to improve the performance of the phonetic recognizer; and investigating more sophisticated robust indexing and retrieval methods in an effort to improve retrieval performance when there are recognition errors.

8. REFERENCES

[1] J. Glass, *et. al*, "A Probabilistic Framework for Feature-Based Speech Recognition," ICSLP 1996, pp. 2277-2280.

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

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

[4] T.J. Hazen and A.K. Halberstadt, "Using Aggregation to Improve the Performance of Mixture Gaussian Acoustic Models," Submitted to ICASSP 1998.

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

[6] K. Ng and V. Zue, "Subword Unit Representations for Spoken Document Retrieval," Eurospeech 1997, pp. 1607-1610.

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

[8] M. Spina and V. Zue, "Automanic Transcription of General Audio Data: Effect of Environment Segmentation on Phonetic Recognition," Eurospeech 1997, pp. 1547-1550.

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

[10] S.J. Young, *et. al.*, "Acoustic indexing for multimedia retrieval and browsing," ICASSP 1997, pp. 199-202.