RJ 10215 (95089) August 29, 2001
Computer Science

Research Report

ADVANCES IN PHONETIC WORD SPOTTING

Arnon Amir

IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

Alon Efrat

University of Arizona
Computer Science Dept.
Tucson, AZ 85721

Savitha Srinivasan

IBM Reasearch Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

Advances in Phonetic Word Spotting

Arnon Amir
IBM Research Division
650 Harry Road
San Jose CA 95120, USA
+(408) 927-1946
arnon@almaden.ibm.com

Alon Efrat
Computer Science Dept.
The University of Arizona
Tucson, AZ 85721, USA
+(520) 626-8047
alon@cs.arizona.edu

Savitha Srinivasan
IBM Research Division
650 Harry Road
San Jose CA 95120, USA
+(408) 927-1430
savitha@almaden.ibm.com

ABSTRACT

Phonetic speech retrieval is often used to augment word based retrieval in spoken document retrieval systems. However, a known problem with phonetic word spotting is the alarmingly high number of false positives as the collection scales up. In this paper, we address this issue with the use of a new indexing and ranking scheme using stochastic phonetic edit distance, and an improved thresholding algorithm. We conduct an extensive set of experiments using a large unbiased query set, and show the improved accuracy when phonetic retrieval is combined with speech recognition retrieval. Using a hundred hours of HUB4 data with ground truth transcript as a benchmark, and several thousands query words, we show improvement of up to 15% in precision compare to speech recognition alone.

1 INTRODUCTION

We address the problem of phonetic speech retrieval for its use in spoken document retrieval (SDR). The SDR task is to quickly find and retrieve all the audio documents which are relevant to a query text provided by the user. In the larger scope, acoustic word spotting is often used to find information in audio documents. In audio documents it is usually not enough to retrieve the relevant documents. It is unexpected that a user will retrieve, say, 20 5-minutes speech segments and would then listen to all of them for more than an hour in order to find which are the relevant ones. In such a case it is desired to find the exact relevant point/s within each of the retrieved documents, to allow for efficient browsing of these small segments rather than entire audio documents. Hence in user's point of view these two problems are complementary. Here we focus on word spotting (WS), the problem of finding and retrieving all the occurrences of a query word in all documents.

While these two problems are very related, there are several important differences between them. While SDR finds relevant documents, or relevant segments of documents, WS retrieves only single words within documents. People have used WS for SDR, although in most situations a spoken document segment would have a more compact representation than just a collection of independent word-time pairs. SDR is often based on statistical information, such as word frequencies (e.g., the commonly used TF/IDF model). WS cannot use such statistics, but rather requires complete information on all the occurrences of all words in all documents. When we consider these two problems, and look at the matching and ranking of retrieved matches, WS uses only a very small amount of local information, while SDR uses word frequencies. WS is typically at the level of identifying a single spoken word, sometimes with assistance from preceding and succeeding words using a language model. SDR uses the entire speech segment, or document, containing multiple words. Hence the performance expected from WS is not as high as expected from SDR, where some amount of word errors can be tolerated. Note also that

evaluation of SDR performance is more subjective than of WS performance, as it requires to rank documents relevant to a query. This involves human decision, and different people might have different opinion on the amount of relevancy of a document to certain queries. The performance of WS, on the other hand, may be measured more objectively, since given the perfect manual transcription of the audio document, there is no subjectivity associated with the relevance of a match. A word or phone sequence at a particular point in time either matches the word in the manual transcription or not.

The search terms can either be known a-priori, from a predefined speech vocabulary (in-vocabulary, or IV), or new to the system, denoted as out of vocabulary words (OOV). For IV words, the most common approach is to use a speech recognition system, trained on word models and language models corresponding to the predefined vocabulary words. Such models include an additional 'filler' component which attempts to match everything else, including all the OOV. This technique, while proved to be very effective for IV words, is clearly limited by the predefined (closed) vocabulary, as it cannot retrieve OOV query words. However, in many common situations the query words are completely unconstrained, and include OOV words such as names of people, locations, companies and products, acronyms etc. The OOV words would often make the best queries. These are rarely used words, and thus would better distinct the thought for documents from the rest. Yet, as they are not retrieved by the speech recognition, a different retrieval method need to be applied.

Word spotting of OOV terms is possible by phonetic retrieval, using phonetic transcriptions of the audio documents. In its simplest form, the phonetic transcript is a string of phonemes, using a phonetic alphabet of the spoken language. The query is converted from text to phonemes, using text pronunciation techniques, and a string matching algorithm is used to find similar strings of phonemes within the phonetic transcript. Hence the phonetic retrieval system is not limited in its retrieval vocabulary. However, the phonetic decoding of speech is not as reliable as speech recognition of words. Hence the process of phonetic WS suffers from a relatively high error rate compared to speech recognition. The work described in this paper combines both to achieve better performance than each method can perform alone.

In this paper we suggest an efficient word spotting system, composed of speech recognition and phonetic indexing. Our approach is as follows: we generate a phonetic transcript of the audio documents, using novel index terms specifically targeted to address phone substitution/insertion/deletion errors. A Bayesian phonetic edit distance is computed, using a statistical model of phone confusions to rank the candidate results. The matching terms are found using both phonetic word spotting and speech recognition based word spotting, and combine these matches to a single unified list of results. We evaluate the system performance by comparing the results with the ground truth transcript of the speech, using 100 hours of HUB4 speech. The rest of the paper is organized as follows: In section 2, we reference related work and section 3 describes our indexing and ranking algorithms. We conclude with an explanation of our experimental evaluation and results in section 4.

2 RELATED WORK

Spoken Document Retrieval systems using large vocabulary speech recognition when combined with text retrieval methods have recently been deemed to be a tractable task in the TREC SDR tasks [Garofalo98, Garofalo99, Voorhees97]. Phoneme based retrieval techniques are generally considered to compliment word based retrieval and are typically used to address the retrieval of out-of-vocabulary words using techniques such

as phone lattice scanning, inverted index of phones, and phone confusion matrices [Spärck Jones96, Wechsler98, Witbrock97].

A phonetic string representation for spoken document retrieval was first introduced by Schäuble and Wechsler [Schäuble95] where five thousand phonetic strings of lengths between three and six phones instead of words were used. The phone recognition system achieved an accuracy of 60-70% on Swiss broadcast news. However, it has been shown that a large vocabulary speech recognition system can do better than existing phone recognition system. Combined word and phone representations to improve retrieval were first evaluated by James [James96] where a statically computed phone lattice was searched during retrieval. Such phone lattice scanning techniques combined with word recognition have been reported in the context of SDR of video mail application to yield 82-85% relative precision compared to perfect text retrieval, and were shown to be better than either method alone [Jones95, Jones96]. Combined word and phonetic retrieval has also been explored in the Informedia project [Witbrock97] where an inverted index for a phonetic transcript comprising of phonetic sub-strings of 3 to 6 phones in length have been used. Experiments on a corpus of about 500 ABC and CNN news stories using combined word and phone indexes resulted in an average precision of 67% with an overall SDR performance of 84.6% of that of a full-text retrieval system.

A variety of phone based subword indexing terms have been investigated by Ng and Zue [Ng98]. They report nominal improvements in precision on FM radio broadcasts of NPR news as a result of these retrieval techniques. Phone confusions have also been used in the probabilistic formulation of term weighting in a Bayesian framework on real world corporate training video collections [Srinivasan00]. In this case, phonetic retrieval has been reported to improve recall over text based retrieval for high word error rates.

Zobel has drawn parallels between phonetic string matching techniques and information retrieval techniques where several approximate string matching techniques and edit distances have been evaluated in the context of phonetic string matching on text [Zobel96]. Wechsler and Schäuble have used phone confusion statistics from the recognizor to compute similarity between phone sequences based on phone substitution, insertion and deletion probabilities [Wechsler98]. Van Leeuwen has presented a model to predict the false alarm rate on the basis of the phonetic content of a query word. However, the revised weighting of the retrieved documents based on this model did not yield improved precision [van Leeuwen99].

Our work builds upon the ideas of phone confusions and edit distance, and makes the following novel contributions: additional index terms based on an understanding of the type of errors made in the phonetic transcript, a ranking and thresholding scheme based on Bayesian phonetic edit distance and likelihood ratio thresholding. This is supported with efficient data structures and extensive experimental evaluation to validate our contributions.

3 PHONETIC WORD SPOTTING METHODOLOGY

We generate a phonetic transcription of the input audio, using the IBM speech recognition system [ViaVoice] with a broadcast news language model to create time aligned word transcripts, and automatically generate equivalent phonetic sequences using the US English phone set [Srinivasan00]. We refer to this as phonetic

transcript, or text, since it is convenient to be considered as text over an alphabet of 52 letters, namely the standard US English phones set.

Our approach to effective word spotting is to combine text-based word spotting with phonetic word spotting thereby leveraging the benefits of both. For the text-based word spotting we create an index of words from the text transcript of the audio. We use standard stop-words removal and word stemming techniques. The phonetic word spotting consists of two stages: Given a query word, we first find a set of candidates in the phonetic transcript which might contain the query word. This is done using several overlapping triphone index terms, with efficient set join operation. In the second stage, we check each candidate to see how similar it is to the query word using a Bayesian edit distance. This score is then normalized and used to rank the candidate results in response to the query. This approach has the following benefits:

1. It allows to compare different ranking methods for estimating the similarity between the candidate and the query word .
2. It allows to gradually widening the search during the first stage to generate a candidates set of reasonable size.
3. The model extends to multi-word queries.
4. Most importantly, the Bayesian edit distance rank highest the most-likely candidates, based on certain assumptions in our model.

3.1 Index Terms

For a given phonetic alphabet, we define the *phone confusion matrix* to model the probability of a phone to be mistakenly recognized by a phone recognition system as a different phone. For the US English phone set of 52 phones, the confusion matrix C is given by:

$$C = \begin{matrix} P(AA/AA) & P(AA/AE) & \dots & \dots & P(AA|Z) & P(AA|ZH) \\ P(AE|AA) & P(AE|AE) & \dots & \dots & P(AE|Z) & P(AE|ZH) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ P(Z|AA) & P(Z|AE) & \dots & \dots & P(Z|Z) & P(ZH|ZH) \\ P(ZH|AA) & P(ZH|AE) & \dots & \dots & P(ZH|Z) & P(ZH|ZH) \end{matrix} \quad \dots(1)$$

Where $AA, AE \dots Z, ZH$ are the phones in the US English phonetic set, and each element in the matrix, C_{ij}, represents the probability of recognizing phone q_j as phone q_i, that is $C_{ij} = P(q_i|q_j)$

We use the phone confusion matrix as our model for phone recognition errors. These probabilities mainly depend on the phonetic recognition system in use and on the audio quality. The phone confusion matrix we use is derived from a theoretical speech performance model, with parameters that model the performance of the speech recognizor for "good speech" (i.e. non-spontaneous speech, native speakers, etc.). These can be adapted to the data in hand in a training phase.

From this confusion matrix, we can also compute the phone insertion/deletion probability. We define a *metaphone group* as a set of phones that are most likely to be confused with each other. Table 1 lists the set of metaphones we use for the US English phone set. These sets were derived from the matrix C, by searching for a

nearly block-diagonal row-permutation, that would place similarly sound phones in consecutive rows. We consider each metaphone group as a new generic phone, and add those metaphones to the confusion matrix. We used seven metaphones, as shown in Table 1.

Table 1: Metaphone Groups

Metaphone	**Metaphone Group**
Atl	AA, AE, AH, AO, AW, AX, AXR, AY, EH, ER
Ctl	CH, JH, SH
Etl	EI, IH, IX,IY
Gtl	B, BD, DD, GD
Th	TH, F
N	N, NG
G	G, D

The confusion probability of each metaphone group with a phone out of the group is the sum of the individual confusion probabilities of the phones in the group.

We use triphone sequences as our indexing unit. In choosing the indexing unit, there is a tradeoff between making it too small, hence not enough discriminative, and making it too long, increasing the chance to miss it if one or more of its phones is miss-recognized. Those a shorter index unit increases the recall as well as the computational effort. Here we mainly focus on the precision-recall, less on the computation speed, although it remains. Each triphone is indexed several times. Each index term is also represented by metaphone representation. Each triphone indexing unit is used as a key to a record that contains necessary information to answer a query. Thus both the process of creating the index, as well as the search process, are preceded by replacing each phone in the key and the query by their metaphone representative. We also generate another type of index term referred to as "don't care" terms. A don't-care index term is a triphone sequence created by skipping one phone out of four consecutive phones. This type of index term addresses phone substitution and phone deletion errors beyond the errors addressed by the metaphone grouping. To address the issue of short word retrieval we generate biphone sequences also. Since boundary word errors such as word joins and word splits are frequent, we add the first phone of the successive word to the biphone sequence.
The record pointed to by each triphone index term includes a segment of the phonetic transcript in the neighborhood of the triphone indexing unit, the document index, the time offset within the document, and a confidence level associated with the recognized word (as provided by the speech recognition engine). All records pointed to by a given key are stored in a linked list of arrays, and we maintain a table of pointers, where each entry of the table corresponds to a possible key (as said, a key is always a triphone). It is easy to maintain this list when records are added or removed, and it is easy to add additional records to obtain efficient searching structure for multi-words searching, when prerequisite knowledge exists to narrow the search. Note the possible numbers of triphones is 180000, so this arrays is of reasonable size, and this indexing method turns out to be rather efficient.

An extension of these ideas on index terms would be to create additional index terms based on the *signature* of the phonetic transcript. The signature is, for example a collection of phones that appear in the segment, and are relatively reliable.

3.2 Bayesian Edit Distance

For a given query term q_i and an observed phone sequence in the transcript $o_{i1}o_{i2}...o_{im}$, Bayes rule can be used to compute the probability that the observed phone sequence $o_{i1}o_{i2}...o_{im}$ is relevant to q_i:

$$P(q_i|o_{i1}o_{i2}...o_{im}) = \frac{P(q_i)P(o_{i1},...o_{im}|q_i)}{P(o_{i1},...o_{im})} \quad \ldots (2)$$

Given the frequency of occurrence of each phone (from the analysis of a large corpus of US English text), and the individual phone confusion probability (from the confusion matrix), we can compute the probabilty that the observed phone o_i is relevant to the actual phone a_i in the query term q_i.

$$P(a_i|o_i) = \frac{P(a_i)P(o_i|a_i)}{P(o_i)} \quad \ldots (3)$$

Thus, using ideas from the standard edit distance, we can compute the probabilty that the phonetic transcript contains the individual phones n the query term. A typical step in the Bayesian edit-distance computation has the following general form: We assume that an array representing the text and an array representing a query word are given. The dynamic programing scheme follows from compute the following expression for every *i*, and *j*.

```
TBL[i][j] = min(
                TBL[i-1 , j-1] + log C[p1,p2] ,
                TBL[i ,j-1] + log P_Delation,
                TBL[i-1, j] + log P_Insertion
```

Figure 1: Edit distance computation

TBL is a two dimensional array where TBL[i,j] stores the probability of the most-likely sequence of substitute, insert and delete operations that match the query with the corresponding phonetic transcript. Here, C[p1,p2], P_Delation and P_Insertion represent the substitution, deletion and insertion probabilities respectively of the *i*'th phone in the query. The resulting number is the minimum probability that a sequence of substitute/insert/delete operations over all possible sequences transformed the phones in the audio to the recognized phonetic transcript. We use this computation of edit distance to rank the candidate terms retrieved by our index terms.

An extension of this idea is to apply Bayes rule to triphone sequences. This requires the collection of statistics regarding triphone sequences on a large corpus of US English text, and an edit distance scheme that computes triphone replacement probabilities. The intuition for computing edit distance on triphones rather than individual phones is that the errors made by the speech recognition system are on sounds created by phone sequences rather than individual phones. The probability that a specific triphone sequence is replaced by a different triphone sequence is generally not the product of the individual phone replacement probabilities in the triphone sequence.

3.3 Answering a Query

Given a query term q_i, we perform the text-based word spotting by searching the index of words recognized by the speech recognition system. The speech recognition confidence for the recognition of each occurrence is used as its text retrieval score.

Next, we generate the phonetic representation of the query and create multiple triphone queries, each is composed of three consecutive phones. The triphone queries overlap, for best coverage of all potential points under high

phonetic error rate. For each triphone query, we retrieve a list of all its occurrences from the in-memory data. These individual lists are typically several hundreds to several thousands potential (Document,TimeOffset) candidates, corresponding to the entire audio collection. We then merge the lists retrieved by different triphones into one list of unique candidates. For each candidate we compute the Bayesian edit distance, as shown in figure 1, and use it as the phonetic score of that candidate.

The last step is to combine the text retrieval candidates with the phonetic candidates. One could suggest different ways to combine these two scores. We choose a very simple method, which gives high preference to speech recognition over phonetic. This eliminates the otherwise very possible case in which phonetic retrieval adds a lot of false positives to the combined list and thus affecting the performance of the text retrieval. We first place all the candidates found by text retrieval, if any, sorted by decreasing text retrieval score. Then we report all the phonetic-only candidates, sorted by decreasing phonetic score. While being a conservative approach, it is proven not to hurt the speech recognition retrieval performance for in-vocabulary words, a concern which is often mentioned in this context. As we found in the experimental results it also helps to improve these results. For out-of-vocabulary words the phonetic candidates are the only candidates, and thus they are naturally listed according to their phonetic score.

3.4 Thresholding

The issue of thresholding is an important one in phonetic retrieval since the number of false positives is dependent on the threshold value in the ranked results. In acoustic word spotting, thresholds are set on the acoustic match score such that words with a score above the threshold are considered matches, and words with a score below the threshold are considered false alarms. In phonetic word spotting an arbitrary threshold value does not yield the optimal results.

4 EXPERIMENTAL EVALUATION

Experimental results for SDR (i.e., not word spotting) were reported in detail in the last several years in the SDR track of the TREC series of conferences [Garofalo97, Garofalo98, Johnson98, Siegler98, Singhal98, Voorhees97]. The first SDR track introduced in TREC-6 was based on a 50 hour test set of radio and broadcast news recordings which were hand-annotated with temporal story boundaries and unique story ID tags to facilitate retrieval. The final test set consisted of 1451 stories where more than half the stories contained less than 100 words (approximately 1 minute). This test set was evaluated against forty-seven 10-15 word test queries by various participants. Subsequent TREC SDR evaluations have been using progressively larger corpuses (100 hours with an average of 7 words per query) and are migrating towards ad-hoc retrieval as opposed to known-item retrieval in order to simulate realistic retrieval tasks. Experiments on SDR data from the TREC-6 conference using about 50 hours of broadcast news data with a phone error rate of 55% has been reported by Wechsler and Schauble to result in 40% of the documents being retrieved at rank 1 [Wechsler98]. An extensive set of experiments using combined retrieval methods including large vocabulary speech recognition and phone lattice scanning have been used on an experimental corpus consisting of 300 messages and an average query length of 12 words. In general, the combination methods gave the best performance reaching 80-85% of full text retrieval [Jones96].

4.1 Evaluation Criteria

Our task and evaluation criteria are different from those mentioned above. A common theme from all the previous SDR experimental work is that *multi-words queries* have been used in the test query set, *whole-story*

segments were retrieved, and that *relevance judgments* were made by humans in order to identify the relevant documents for each query. Our task differs from this in that our query set consists of single-word queries, our word spotting task aims at retrieving all occurrences of each query in all documents, and our evaluation measure is at word time-of-occurrence level, compared to ground truth time-aligned manual transcription of the speech.

The evaluation criteria is based on the occurrence of each query word in the manually transcribed perfect transcript with no human relevance judgments involved. There is no subjectivity associated with what constitutes a false positive. A match is correct if the exact word was said within a window of two seconds around the retrieved point in time. This time window tolerates for imprecise word-times in our ground truth, which only provides times at sentence granularity. An inexact match or a larger time difference is resulted in a false positive. This evaluation method is similar to the one reported in [sdr00].

4.2 Evaluation Data Set

The test collection is based on 100 hours of HUB4 data [LDC98] where word error rate is about 35%. It consists of about 12Gb of 22.05Khz WAV audio files. The speech contains 24,018 different words, of which 17,955 are in-vocabulary words and 6,063 are out-of-vocabulary words (after stop words removal). The query words are retrieved from the 1.04 million spoken words composing the 100 hours of speech. Although 25% of the words are OOV, they only occure about 3% of the time. Still, they are of special significance for speech retrieval tasks, as explained before. Table 2. Evaluation data and query set

Total Duration of audio in test collection	100 Hrs
Number of spoken words in test collection	1,040,456
Total Number of in-vocabulary queries	17,955
Total Number of out-of-vocabulary queries	6,063
Total number of queries	24,018

Our exhaustive queries test set consist of all the words which are listed in the ground turth transcription, namely 24,018 different queries (stop words excluded). These queries are divided into IV/OOVgroups, and are further divided by the query length in phones. The number of phones was identified to be an important factor from previous work, e.g. [sdr00]. In general, the longer the phonetic query is, the more accurate the result.

4.3 Experimental Results

We have processed each of the queries independently, and then combined the results to generate graphs according to 32 lists of words, namely IV/OOV for length 3 to 18 phones. Here we show only the odd-numbered lengths, to reduce the load on the figures. The even-numbered lengths behave similarly. Figure 3a,b, and c show three representative examples from the sixteen graphs of in-vocabulary sets of queries. The graphs present the precision as function of recall. The recall is normalized to the number of ground truth word occurrences, such that recall of 0.6 means the recall of 60% of all the occurrences of the query words. These figures show an improvement of 5-15% in the precision of combined phonetic retrieval compared to speech recognition alone. The largest improvement is achieved for words length in the midrange. We believe that the reason for that is the improvement of both speech recognition retrieval and phonetic retrieval with the increase of word length. Hence while speech recognition results for very short words (Figure 3.a) is relatively low, the

phonetic index has also hard time with lots of false positives. For very long words, on the other hand, the speech recognition alone reaches precision of over 90%, hence there is only very little room for improvement left for the phonetic retrieval. The highest improvement is achieved in between, over a range of lengths between 7 and 10.

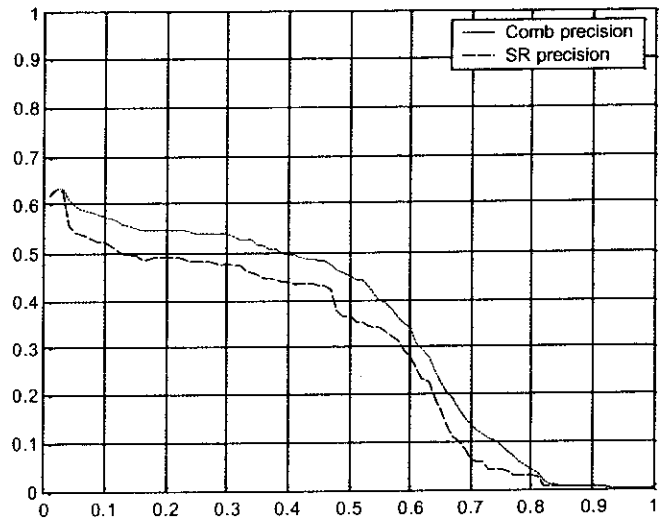

Figure 3a. Retrieval results for in-vocabulary words three phones long. The graph represents speech recognition results (dashed line) and combined SR-phonetic (solid line) retrieval of 13,221 word occurrences in 100 hours of speech. Short words make the hardest case for phonetic retrieval, which improves the precision by 5-8%.

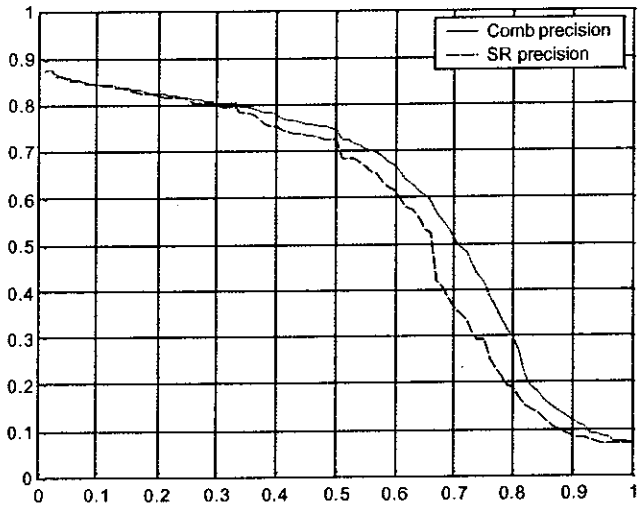

Figure 3b. Same as Figure 3a, for in-vocabulary words nine phones long. Retrieval precision of 18,739 word occurrences in 100 Hours of speech is improved by 5-15%.

The dependency of precision on the query word length for speech recognition and for combined phonetic retrieval are best captured in Figure 4 and Figure 5, respectively. It shows how precision improves with word length. The

improvement is very significant for shorter words, starting from three phones long at the lowest curve and going up to 15 phones at the highest curve (only odd lengths are drawn). Precision increases for longer words, but in smaller and smaller ratios.

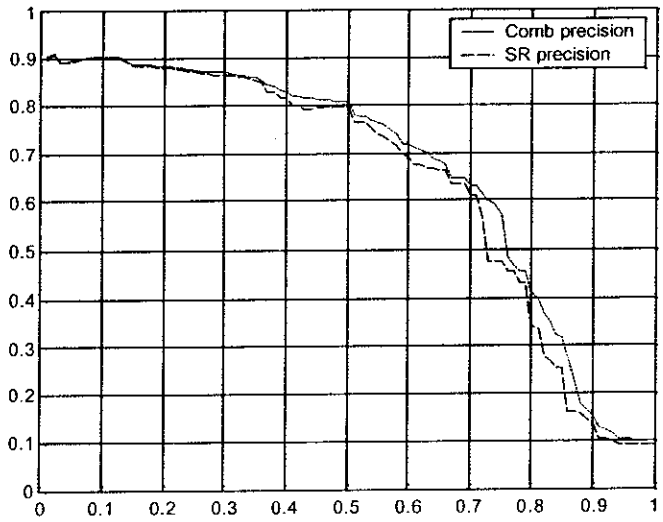

Figure 3c. Same as Figure 3a, for in-vocabulary words thirteen phones long. The already high precision of speech recognition retrieval of 1,812 word occurrences is slightly improved by the phonetic retrieval.

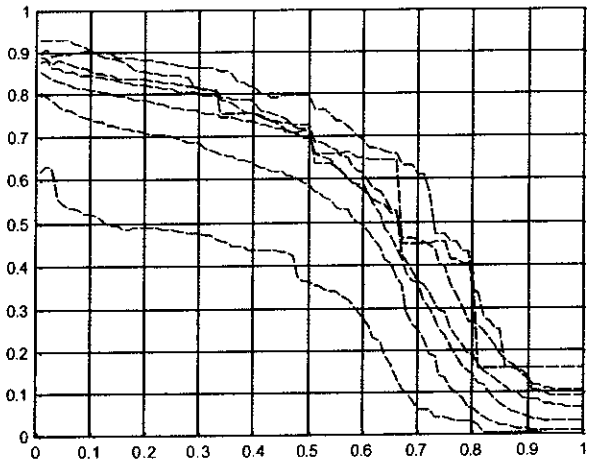

Figure 4. Retrieval precision-recall for speech recognition alone, as it changes according to the query word length in phones from three (bottom curve) to 15 (upper curve).

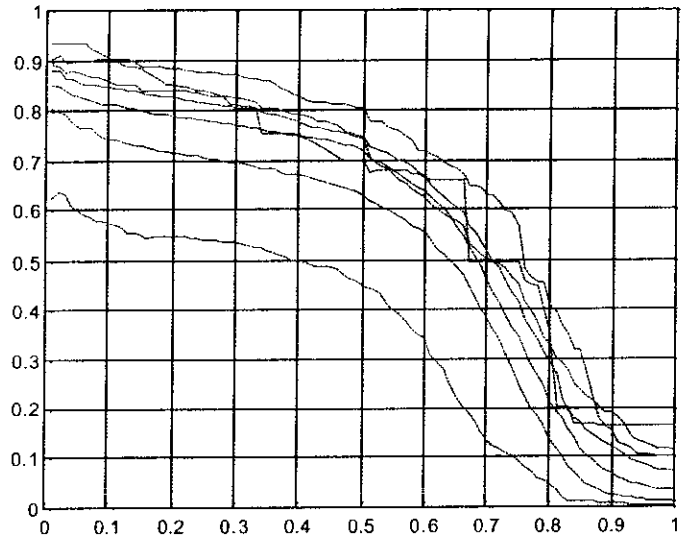

Figure 5. Retrieval precision-recall for combined phonetic retrieval, as it changes according to the query word length in phones from three (bottom curve) to 15 (upper curve).

The results for out-of-vocabulary words is captured in a single graph, shown in Figure 6. Here the speech recognition is of no use, as it does not recognize any of these words. These are all retrieved by the phonetic search alone. As most of these words occur in the data set only once, the task in hand can be refereed to as finding a single word occurrence in a heap of one million words, under imprecise phonetic spelling. An average precision of 0.2 means that the single true result is, on average, likely to be in the first five matches. This is very practical in many practical situations.

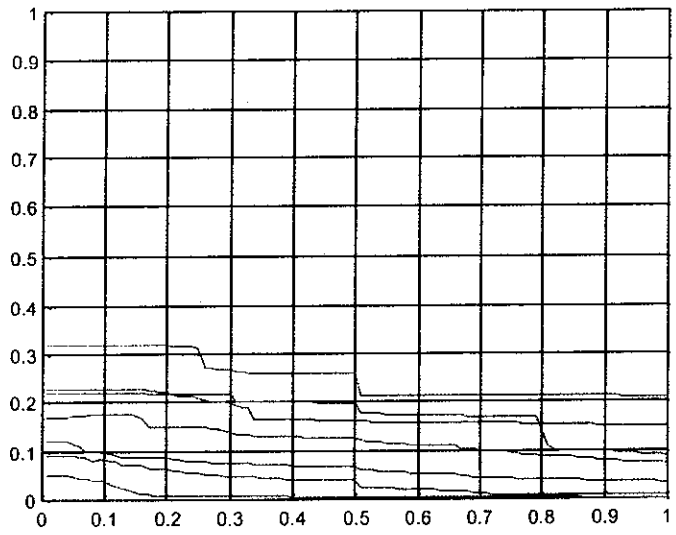

Figure 6. Retrieval precision-recall for out-of-vocabulary words. These query words are found only by the phonetic retrieval. The precision increases as the query word length increases, from three phones (bottom curve) to fifteen phones (upper curve).

5 CONCLUSION AND FUTURE WORK

This work provides a method for phonetic speech retrieval. The main contribution is in the introducing of metaphones indexing, that overcome phonetic errors, and in the Bayesian edit distance, which model the imperfect matching between phonetic strings, using a phonetic confusion matrix, additions and deletions model. The results of extensive experimentation of 24,000 queries in 100 hours of speech show improvement of 5-15% over speech recognition alone.

ACKNOWLEDGMENT

We thank John Kececioglu for helpful discussions.

REFERENCES

1. [Garofalo98] Garofolo, J.,Voorhees, E., Auzanne, C., Stanford, V. and Lund, B. (1997). The TREC-7 Spoken Document Retrieval Track Overview and Results. In Proceedings of the seventh Text Retrieval Conference (TREC-7), pp. 79. NIST Special Publication 500-242.
2. [Garofalo99] Garofolo, J., Auzanne, C. and Voorhees, E. The TREC Spoken Document Retrieval Track: A Success Story, page 107 National Institute of Standards and Technology.
3. [James96] James, D. System for Unrestricted Topic Retrieval from Radio News Broadcasts, In Proceedings of ICASSP-96, Atlanta, GA, May196, pp. 279-282.
4. [Johnson98] Johnson, S.E., Jourlin, P., Moore, G.L., Spärck Jones, K and Woodland, P.C. Spoken Document Retrieval for TREC-7 at Cambridge University. In *Proceedings of the Seventh Text Retrieval Conference (TREC-7), (NIST Special Publication)* 1998
5. [Jones95] Jones, G. J. F., Foote, J. T., Spärck Jones, K. , and Young, S. J.. Video Mail Retrieval: the effect of word spotting accuracy on precision. In Proceedings of *ICASSP 95, volume 1, pp. 309-312*, Detroit, MI.
6. [Jones96] Jones, G. J. F., Foote, J. T., Spärck Jones, K., and Young, S. J. Retrieving Spoken Documents by Combining Multiple Index Sources. In *Proceedings of SIGIR 96, pp. 30-38*, Zurich, Switzerland.
7. [LDC98]1998 HUB-4 Broadcast News Evaluation English Test Material, LDC catalog no.: LDC2000S86, ISBN: 1-58563-172-8
8. [Ng98] Ng, K. and Zue, V. Phonetic Recognition for Spoken Document Retrieval. In Proceedings of ICASSP 98, pp. 325-328.
9. William Pugh. Skip List A probablistic alternative to balanced trees. Communication of the ACM 33(6) 668--676 1990.
10. [Schäuble95]Schäuble, P. and Wechsler, M. First Experiences with a System for Content Based Retrieval of Information from Speech Recordings. In *IJCAI-95, Workshop on Intelligent Multimedia Information Retrieval*, Maybury, M.T.
11. [Siegler98] Siegler, M.A., Witbrock, M.J., Slattery, S.T., Seymore, K., Jones, R.E. and Hauptmann, A.G. Experiments in Spoken Document Retrieval at CMU. In *Proceedings of the Seventh Text Retrieval Conference (TREC-7), (NIST Special Publication)* 1998.

12. [Singhal98] Singhal, A., Coi, J., Hindle, D., Lewis, D. and Pereira, F. AT&T at TREC-7. In *Proceedings of the Seventh Text Retrieval Conference TREC-7, (NIST Special Publication)* 1998.
13. [Srinivasan00] Srinivasan, S. and Petkovic, D., Phonetic Confusion Matrix Based Spoken Document Retrieval. In Proceedings of SIGIR-2000. July 2000, Greece.
14. [van Leeuwen99] van Leeuwen, D.A., Kraaij, W. and Ekkelenkamp, R. Prediction of keyword spotting performance based on phonemic contents. In Proceedings of the ESCA ETRW workshop: Accessing Information in spoken audio, University of Cambridge, 1999.
15. [ViaVoice] See URL at http://www-4.ibm.com/software/speech/
16. [Voorhees97] Voorhees, E., Garofolo, J. and Spärck Jones, K. (1997). The TREC-6 Spoken Document Retrieval Track Overview and Results. In Proceedings of the sixth Text Retrieval Conference (TREC-6), pp. 83. NIST Special Publication 500-240.
17. [Wechsler98] Wechsler, M., Munteanu, E., and Schäuble, P. New techniques for open vocabulary spoken document retrieval. In *Proceedings of SIGIR '98*, pp. 20-27, Melbourne, Australia, 1998
18. [Witbrock97] Witbrock, M. and Hauptmann, A. Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents. In *Proceedings of DL97, The Second ACM International Conference on Digital Libraries,* Philadelphia, PA.
19. [Zobel96] Zobel, J. and Dart, P. Phonetic String Matching: Lessons from Information Retrieval. In Proceedings of SIGIR-96, Zurich, Switzerland.