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 used to augment word based retrieval in spoken document retrieval systems, for in and out of vocabulary words. In this paper, we present a new indexing and ranking scheme using metaphones and a Bayesian phonetic edit distance. We conduct an extensive set of experiments using a hundred hours of HUB4 data with ground truth transcript and twenty-four thousands query words. We show improvement of up to 15% in precision compare to results obtained speech recognition alone, at a processing time of 0.5 Sec per query.

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 used to find information in audio documents. In audio documents it is usually not enough to retrieve the relevant documents. The exact relevant point/s within each of the retrieved documents has to be retrieved to allow for efficient browsing. 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 requires complete information of all the occurrences of all words in all documents. Hence the performance expected from WS is not as high as expected from SDR, where some amount of word errors can be tolerated.

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). The set of IV words is defined by the automatic speech recognition (ASR) system in hand. 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 less commonly used words, and thus would better distinct the thought for documents from the rest. Yet, as they are not retrieved by ASR, 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 ASR based word spotting of the IV words.

The goals of the work described in this paper is twofold: develop advanced phonetic retrieval to improve WS performance for IV words, and to provide efficient WS of OOV. Towards these goals we suggest an improved phonetic indexing scheme that accommodates commonly confused phones, and a Bayesian phonetic edit distance to better rank the retrieved candidates.

2 RELATED WORK

Retrieval of spoken documents has been an active research area for around ten years. Systems using large vocabulary speech recognition, combined with text retrieval methods, have recently been deemed to be a tractable task in the TREC SDR tasks [2,12]. Phoneme based retrieval techniques are generally considered to compliment word based retrieval and are typically used to address the retrieval of OOV words using techniques such as phone lattice scanning, inverted index of phones, and phone confusion matrices [12, 13]. This is considered as one of the unsolved issues in the SDR filed.

Combined word and phone representations to improve retrieval were used by James [4] where a statically computed phone lattice was searched during retrieval. Such phone lattice scanning techniques combined with word recognition were shown to be better than either method alone [6]. Combined word and phonetic retrieval has also been explored in the Informedia project [13]. A variety of phone based subword indexing terms have been investigated by Ng and Zue [8]. Phone confusions have also been used in the probabilistic formulation of term weighting in a Bayesian framework on real world corporate training video collections [9], and 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 [14]. Wechsler and Schäuble have used phone confusion statistics from ASR to compute similarity between phone sequences based on phone substitution, insertion and deletion probabilities [12]. 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 [10].

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

3 PHONETIC WS METHODOLOGY

We generate a phonetic transcription of the input audio, using the IBM speech recognition system [11] with a broadcast news language model to create time aligned word transcripts, and automatically generate equivalent phonetic sequences using the US English phone set [9]. 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.

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. Each element in the matrix, C_{ij}, represents the probability of miss-recognizing phone q_j as phone q_i, that is $C_{ij} = P(q_i|q_j)$. We add to this matrix several more rows and columns to model addition/deletion errors and pauses.

Based on the content of the confusion matrix we have identified seven groups of phones that are more likely to be confused with each other, denoted as *metaphones* groups. Note that without special care, a single confused phone in the phonetic transcript would prevent us from finding all three three-phones keys which contain it. To address this problem we use seven metaphone groups, each contains between two and ten similar phones. For example, the phones B, BD, DD, GD form one metaphone group. We consider each metaphone group as a new generic phone. Each key that contains one or more phones from metaphone groups is also indexed using its metaphones representation, where all the phones are replaced with their metaphones. All records pointed to by a given key are stored in a linked list, pointed from a table where each entry of the table corresponds to a three-phones key. This indexing method turns out to be rather efficient.

3.2 Bayesian Edit Distance

For a given query term $q = q_1...q_n$ and an observed phone sequence in the transcript $o = o_1...o_m$, we would like to evaluate how similar (sound) are they. For a single phone-to-phone comparison we use Bayes rule to compute the probability that the observed phone o_i origin from a (possibly confused) q_j:

$$P(q_j|o_i) = P(q_j)P(o_i|q_j)/P(o_i)$$

where $P(o_i|q_j) = C_{o_i,q_j}$, $P(o_i)$ are derived from a statistics over a large corpus of data, and $P(q)$ drops later when we normalize the score. We cannot assume, however, a sequential one-to-one correspondence between the two phonetic strings. Deletions and additions are common errors in the phonetic transcripts. Therefor we model the string similarity using the edit distance (see, e.g., [3]). A sequence of edits is a transformation that works on *o* and converts it to *q*, using a sequence of allowed single phone operations which consist of substitution, addition and deletion, as modeled by the confusion matrix. The likelihood of an editing sequence is the joint likelihood of its editing steps, assumed here to be independent. The edit distance is the maximum likelihood among all possible edit sequences that convert *o* to *q*. It can be efficiently computed using dynamic programming. Note that at this point it is not a distance measure, as even a substitution of a phone with itself is scored with a non-zero likelihood. To derive the final score of the match, we normalize the result with *q*,'s self edit distance.

3.3 Answering a Query

The phonetic word spotting consists of two stages: Given a query word, we generate the phonetic representation of it and create multiple three-phones keys, each composed of three consecutive phones. We then retrieve a list of (Document,Offset) candidate records, found in the phonetic transcript, which might contain the query word. Then we compare each candidate with query word and rank it using a Bayesian phonetic edit distance.

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 higher 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. 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. For OOV words the phonetic candidates are the only candidates, and thus they are naturally listed according to their phonetic score.

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 [2,5,12]. 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 the objective ground truth time-aligned manual and accurate transcription of the speech. 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 [2].

The test collection is based on 100 hours (1.04 million spoken words) of HUB4 data [7] where SR word error rate is about 35%. This data, accompanied with ground truth timed manual transcript, is traditionally used to assess speech recognition quality. We found it most suitable for our WS task, more than standard SDR data. 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). Although 25% of the different words are OOV, they only occur about 3% of the time. Still, they are of special significance for speech retrieval tasks, as explained before. Our exhaustive queries test set consist of all the words which are listed in the ground truth transcription, namely 24,018 different queries. These queries are divided into IV/OOV groups, and are further divided by the number of phones they contain. The number of phones was identified to be an important factor from previous work, e.g. [10]. In general, the longer the phonetic query is, the more accurate the result.

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. Due to space limitation we only show here the total average of IV words over the entire corpus. The graph shown in Figure 1 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. This figure shows an improvement of 5-10% in the precision of combined phonetic retrieval (solid line) compared to speech recognition alone (dashed line). The largest improvement, between 10-15%, is achieved for words length in the midrange, about 9 phones long (not shown). The average processing time per query retrieval on 100 hours of speech is 0.5 Sec. More technical details can be found in [1].

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 the speech recognition transcript.

ACKNOWLEDGMENT

We thank John Kececioglu for helpful discussions, and the anonymous referees for their most helpful comments

REFERENCES

[1] Amir, A., Efrat, A., and Srinivasan, S., "Advances in Phonetic Word Spotting", IBM Research Report RJ 10215, August 2001.

[2] Garofolo, J., et al. (1997). The TREC-7 Spoken Document Retrieval Track Overview and Results. In Proc. of the seventh Text Retrieval Conference (TREC-7), pp. 79. NIST Special Publication 500-242.

[3] Gusfield, D., *Algorithms on Strings, Trees, and Sequences*, Cambridge, 1997.

[4] James, D. System for Unrestricted Topic Retrieval from Radio News Broadcasts, In Proc. of ICASSP-96, Atlanta, GA, 1996, pp. 279-282.

[5] Johnson, S.E. et al., Spoken Document Retrieval for TREC-7 at Cambridge University. In Proceedings of the Seventh Text Retrieval Conference (TREC-7), 1998

[6] Jones, G. J. F. et al., Retrieving Spoken Documents by Combining Multiple Index Sources. In Proc. of SIGIR 96, pp. 30-38, Zurich, Switzerland.

[7] 1998 HUB-4 Broadcast News Evaluation English Test Material, LDC catalog no.: LDC2000S86, ISBN: 1-58563-172-8

[8] Ng, K. and Zue, V. Phonetic Recognition for Spoken Document Retrieval. In Proc. of ICASSP 98, pp. 325-328.

[9] Srinivasan, S. and Petkovic, D., Phonetic Confusion Matrix Based Spoken Document Retrieval. In Proceedings of SIGIR-2000. July 2000, Greece.

[10] Van Leeuwen, D.A. et al., Prediction of keyword spotting performance based on phonemic contents. Proc. of ESCA ETRW workshop: Accessing Information in spoken audio, U. of Cambridge, 1999.

[11] See URL at http://www-4.ibm.com/software/speech/

[12] Wechsler, M., Munteanu, E., and Schäuble, P. New techniques for open vocabulary spoken document retrieval. In Proc. of SIGIR'98, pp. 20-27, Melbourne, Australia, 1998

[13] Witbrock, M. and Hauptmann, A. Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents. In Proc. of ACM DL97, 2nd ACM Int. Conf. on Digital Libraries, Philadelphia, PA.

[14] Zobel, J. and Dart, P. Phonetic String Matching: Lessons from Information Retrieval. In Proceedings of SIGIR-96, Zurich, Switzerland.

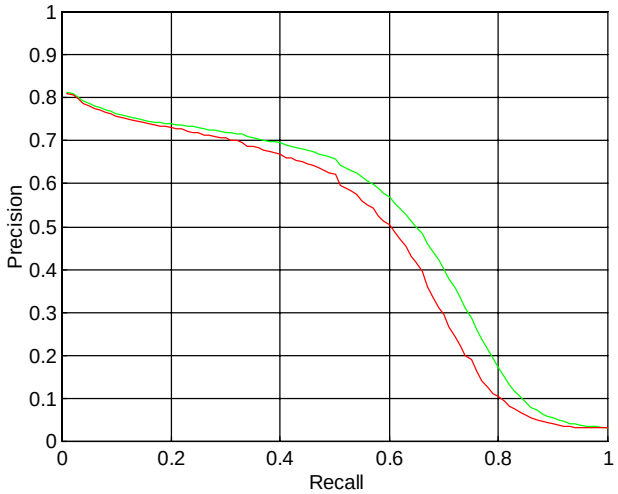

Figure 1. Precision recall results for 18,000 queries on 100 hours of speech, compare to ground truth. The combined phonetic and speech recognition result (solid line) is 5-10% higher than the speech recognition alone (dashed line).