

Speech Retrieval using Phonemes with Error Correction

Corinna Ng Justin Zobel

Department of Computer Science, RMIT, GPO Box 2476V, Melbourne 3001, Australia

www.cs.rmit.edu.au/~chienn,~jz

Abstract *In speech retrieval, text queries are matched to spoken-word recordings. We undertook a series of retrieval experiments on such speech documents, using both phonemes and words for matching text to speech. We found that, given a reliable automatic word-based process for speech recognition, phoneme-based retrieval can be as effective as word-based retrieval. However, retrieval is worse with phoneme-based speech recognition. Error correction techniques, such as manual error correction of phoneme sequences and use of a standard string edit distance, did not improve the effectiveness of phoneme-based retrieval.*

In speech retrieval, textual queries are used to retrieve speech documents from a corpus. An approach to speech retrieval is to preprocess the speech documents with a speech recognition system, thus converting the audio signals into indexable tokens such as words or phonemes.

The word-based approach to speech recognition has two key problems. First, it requires a dictionary. Unknown words, those not present in the dictionary, will not be recognised, a problem that is likely to be acute for names. Second, the detection of word boundaries can be difficult because there is no explicit delimiting of words in continuous speech. These problems are likely to decrease retrieval effectiveness, particularly because more discriminating query terms are rare and therefore less likely to be recognised.

We investigated a phoneme-based approach to both the recognition and retrieval processes. Prior to retrieval, a recogniser is used to extract phonemes from the audio documents. A phoneme is the smallest context-free unit of speech that has meaning; the physical sound of a phoneme is called a phone. A small, fixed set of phones is used for phoneme recognition. The two difficulties with word-based recognition do not arise with phonemes. First, words are represented as a sequence of phonemes, and words known or unknown are represented by their component phones. Second, recognition of word boundaries is unnecessary. However, phoneme recognition is more error-prone than word recognition. Phoneme boundaries are difficult to detect, and, in word recognition, semantic context can be used to aid the recognition process; no such context is available for phoneme recognition.

Once the recognition is complete, index features must be extracted from the documents, to allow efficient retrieval on a large corpus. In text documents, the features ex-

tracted are usually words. In speech documents, individual phonemes are not sufficiently discriminating; we chose to index all phoneme strings of length n , or n -grams, where individual n -grams were allowed to overlap. An example of 4-grams, or quadgrams, is shown in Figure 1 for the phrase "a question about", where word boundary information has been discarded. N -gram indexing has been shown to be language independent [1].

As for word-based retrieval, n -gram-based retrieval has potential limitations. An n -gram created from a query term can match any word containing that n -gram, and in the document collection the n -gram may be formed across a word boundary, that is, by composing the end of one word with the start of another. These false matches can degrade effectiveness.

The document collection used in our experiments consists of 1500 speech documents and 49 queries, provided by Text Retrieval Conference (TREC) sponsored by NIST and DARPA to encourage research in information retrieval. For our experiments, three transcriptions of the speech documents were used. These are a manual transcription in words, an automatic transcription in words, and an automatic transcription in phonemes. The queries and documents were neither stopped nor stemmed. There is also a training set of about 1500 documents. This set of documents was used to train the phoneme recognition system.

The Carnegie-Mellon Pronouncing Dictionary (CMU) was used to translate words to phonemes. In this speech document collection, there are approximately 18,000 unique words. Of these, 2000 are not in the dictionary. Most of these words are names and plurals. We manually added these pronunciations, thus mimicking the behaviour of a word-based speech recogniser that outputs phonemes.

The text retrieval engine MG [4] was used for the retrieval experiments. For each query, 1000 documents are retrieved. A standard cosine formulation was used to estimate similarity.

Average precision was used to measure retrieval performance. For most of the queries for this collection there is only one relevant document per query. Hence, precision can be computed as the mean reciprocal of the rank. Average precision of the set of queries Q is calculated as:

$$\text{Average precision} = \frac{\sum_{q \in Q} (\text{rank}_q)^{-1}}{|Q|}$$

where rank_q is the rank at which the relevant document for query q is found.

To establish baselines we used the word-based manual and automatic transcriptions of the speech documents; results are shown in Table 1. These transcriptions were also translated to phonemes, to simulate the best possible retrieval results given minimum transcription errors. For these experiments, trigrams and quadgrams are used.

The first two columns (unbounded) of Table 2 show effectiveness using queries where n -grams were created from

Permission to make digital/hard copy of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copying is by permission of ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or fee. SIGIR'98, Melbourne, Australia © 1998 ACM 1-58113-015-5 8/98 \$5.00.

Sentence: a question about quadgrams created: aque ques uest esti stio tion iona onab nabo about

Figure 1: An example of quadgrams.

Document collection	Average precision
Manual	0.702
Automatic	0.558

Table 1: Baseline results for word-based retrieval, using the manual and automatic transcriptions against textual queries.

Av. precision	unbounded		bounded		augmented-manual		augmented-edit	
	trigram	quadgram	trigram	quadgram	trigram	quadgram	trigram	quadgram
manual	0.698	0.687	0.616	0.569	0.423	0.509	0.014	0.136
automatic	0.582	0.538	0.579	0.486	0.463	0.479	0.011	0.117
phoneme	0.204	0.236	0.230	0.205	0.153	0.197	0.014	0.056

Table 2: Results of experiments with phoneme-based speech retrieval. Each column is a different type of query. Each row is a different form of the document collection.

query terms across word endings. The effect of including boundary information for queries, that is, of not creating n-grams across word boundaries in queries, is shown in the next two columns (bounded). Within the document collections, n-grams are created across word boundaries in both cases, as it is assumed that word-boundary information is not available within the corpus. As can be seen, use of word boundaries generally degrades effectiveness; we hypothesize that the across-boundary n-grams provide a limited form of phrase matching. For the automatic transcription, retrieval effectiveness is slightly, but not significantly, improved.

As can also be seen, retrieval based on phoneme recognition is poor, due to a high rate of recognition error. We investigated several techniques for managing such errors. First, we tested manual error correction. It is plausible that transcription errors in a word in the training collection would also occur in the test collection. From the training collection we obtained a set of query terms that had been incorrectly transcribed and included the most frequent erroneous transcriptions in the query. Only key query words that had occurred in the training collection were augmented in this way, and the phonetic n-grams created did not cross word endings. Retrieval results using these augmented queries are shown as augmented-manual in Table 2.

Second, we investigated the use of approximate string matching techniques to obtain additional phonetic strings for the retrieval process. This technique was successful in name-matching experiments [5]. A standard string edit distance was used on the phonetic n-grams. The effectiveness of queries augmented by n-grams found in the edit neighborhood of query n-grams is shown as augmented-edit in Table 2.

However, manual error correction did not improve retrieval. Analysis of the results per query found that the ranking of the relevant documents was rather different from that in previous retrieval experiments. Error correction using a standard edit distance was extremely poor, as too many noisy n-grams were added.

Overall, our experimental results show that retrieval based on word recognition is more effective than retrieval based on phoneme recognition, which was degraded by a high error rate. Our experimental results also showed that trigrams are more effective than quadgrams, largely be-

cause the high recognition error rate meant that longer sequences of phonemes included significantly more errors [2]. (Results from other experiments with n-gram indexing on text collections also indicated that retrieval using trigrams is optimal.) Comparing the results between unbounded and bounded queries showed that the lack of boundary information aided in retrieving relevant documents.

Phoneme recognition solves the problem of recognition of unknown terms and removes the need for word boundary information. Our experimental results show that n-gram retrieval [3] with phonemes is as effective as word-based retrieval, given a reliable automatic speech recognition process. The results also show that word boundary information is irrelevant when phonemes and n-grams are used for retrieval purposes. We are conducting ongoing experiments to investigate techniques for error correction involving approximate string matching.

Acknowledgements We are grateful to Ross Wilkinson for his contributions to this project, and to Peter Schäuble and Eugene Munteanu for providing their phoneme transcriptions of the speech document collections.

References

- [1] W. B. Cavnar. Using an n-gram-based document representation with a vector processing retrieval model. In *Proc. Third Text REtrieval Conference (TREC-3)*, pages 269–277, Gaithersburg, MD, 1994. National Institute of Standards and Technology Special Publication 500-225.
- [2] K. Ng and V. W. Zue. Subword unit representations for spoken document retrieval. In *Proc. ESCA Eurospeech Conference*, pages 1607–1610, Rhodes, Greece, 1997.
- [3] M. Wechsler and P. Schauble. Speech retrieval based on automatic indexing. In *Workshop in Computing Science-MIRO*. Springer Verlag, 1995.
- [4] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and indexing documents and images*. Van Nostrand Reinhold, 1994.
- [5] J. Zobel and P. Dart. Phonetic string matching: Lessons from information retrieval. In *Proc ACM-SIGIR*, pages 166–172, Zurich, Switzerland, 1996.