Experiments in Spoken Document Retrieval using Phoneme N-grams

Corinna Ng† Ross Wilkinson‡ Justin Zobel†

†Department of Computer Science, RMIT
GPO Box 2476V, Melbourne 3001, Australia
{chienn,jz}@cs.rmit.edu.au

‡CSIRO, Division of Mathematical and Information Science
723 Swanston St, Carlton VIC 3053, Australia
Ross.Wilkinson@cmis.csiro.au

Abstract

In spoken document retrieval, speech recognition is applied to a collection to obtain either words or subword units, such as phonemes, that can be matched against queries. We have explored retrieval based on phoneme n-grams. The use of phonemes addresses the out-of-vocabulary problem, while use of n-grams allows approximate matching on inaccurate phoneme transcriptions. Our experiments explored the utility of word boundary information, stop word elimination, query expansion, varying the length of phoneme sequences to be matched and, various combinations of n-grams of different lengths.

Given word-based recognition, we can match queries to speech using a phoneme representation of the words, permitting us to test whether it was the recognition or the matching process that was most crucial to retrieval performance. Our experiments show that there is some deterioration in effectiveness, but the particular form of matching is less vital if the sequence of phonemes was correct. When phone sequences are recognised directly, with higher error rates than for words, it was more important to select a good matching approach. Varying gram length trades precision against recall; combination of n-grams of different lengths, in particular 3-grams and 4-grams, can improve retrieval. Overall, phoneme-based retrieval is not as effective as word-based retrieval, but is sufficient for situations in which word-based retrieval is either impractical or undesirable.

1 Introduction

Information retrieval techniques are widely used in text databases to identify documents that are likely to be relevant to free-text queries [17]. The aim of spoken document retrieval is to provide similar functionality for databases of spoken documents. Such documents, in the form of audio signals, are recorded from many different sources, such as news broadcasts on radio and television. It would be valuable to be able to retrieve such documents in response to textual or spoken queries.

In order to be able to interactively query such media, speech signals are converted into words, phonemes, or other subword units [10], using a speech recognition system. This work focuses on the phoneme-based approach, where spoken documents are recognised as phoneme sequences and retrieval is based on matching the n-grams (sequences of n symbols) of these transcriptions.

These experiments were conducted as part of the Text REtrieval Conference (TREC), sponsored by NIST and DARPA to encourage research in information retrieval [21]. In the TREC Spoken Document Retrieval (SDR) experiments, word-based approaches have consistently outperformed phoneme approaches. However, there are several reasons for using phonemes.

In word-based recognition (WBR), three assumptions can be made about the recognition process. First, the recogniser uses a large vocabulary [27] containing most of the words to be recognised. In TREC-7 SDR, a collection of 100 hours of spoken documents contain about 23,000 unique words. Therefore obtaining good recognition performance requires a vocabulary of at least that size. Second, the spoken documents must be consistent with the language model used. The model determines the recognition language structure, embodying certain assumptions about exactly what word can precede another word. If the model can predict the language well, higher recognition performance can be achieved. Finally, the computational resources required to perform WBR must be available.

We consider the situation where the above assumptions cannot be made. If the recognition process is on a small hand-held device where the computational resources are limited, a large vocabulary may not be possible. Furthermore, the resources required to build a language model may not be available and, even with a small language model, the spoken documents may not be consistent with it, thus limiting its usefulness. Therefore, we consider a basic recognition system, such as a phoneme recogniser, which can run on a small processor. Due to its simplicity, it suffers from poor recognition performance compared to the WBR approach. Second, we consider a spoken document collection containing a significantly high proportion of out-of-vocabulary (OOV) words which can adversely affect retrieval when mis-recognised. An example is a spoken collection of names and places. From a textual retrieval perspective, these words, due to their rarity in the document collection, usually assist in identification of relevant documents. The WBR system, not having seen sufficient examples of these words in the training process, may well incorrectly recognise them as some other words.

In our work, a basic phoneme recogniser is treated as a black box. Techniques are investigated here to improve retrieval and tested on standardised corpora such as the TREC-7 and TREC-6 SDR collections. These collections do not exhibit any of the properties outlined previously and hence they do not allow us to test our assumptions fully. Therefore, we do not expect our phoneme n-gram approach to be more effective than the word-based approach to retrieval. But using these available corpora, allows us to test some of the retrieval properties of phoneme n-grams.

The technique of using n-grams in retrieval [1] has shown reasonable performance in text-based collections. Previous experiments by Wechsler [23, 24], Ng and Zue [12, 13, 14] and Smeaton et al. [19] on n-gram retrieval from phoneme transcriptions obtained directly from a phoneme recogniser showed that phoneme n-gram retrieval can be effective in practice. Other experiments have shown that phoneme retrieval can be used to complement word retrieval [25] when word recognition has failed, especially in situations where names and unknown words are mis-recognised.

In this paper, we refine methods for using phoneme n-grams. We used two kinds of phoneme recognition: direct recognition as a series of phonemes, and transformation to phonemes from recognised words. In a typical word-based approach to retrieval, word boundary information and stopping (removal of common words) in the documents and queries usually aids retrieval performance [7]. We investigate whether the same properties hold for the phoneme n-gram approach to retrieval. The first set of experiments compare retrieval effectiveness for the two kinds of phoneme n-grams, either from direct phoneme recognition or transformation from words. Queries are varied to test the effects of stopping and word boundary information. We combined phoneme n-grams of varying lengths to investigate whether different evidence was being provided by various n-gram sizes. In addition, we examined an approximate technique of query expansion, adding neighbour terms to the query set. In another approach to query expansion, a list of the most frequent misrecognised query terms was obtained from a training collection and used to augment the queries.

Phoneme n-gram retrieval using phonemes transformed from recognised words showed better retrieval performance than from direct phoneme recognition but was not as good as word-based retrieval. Combination of phoneme n-grams showed some improvements in retrieval, in particular combination of 3-grams and 4-grams. Overall, the phoneme-based approach to retrieval is less effective than the word-based approach, but is nonetheless effective enough to be used in practice. This paper expands and consolidates work previously reported at TREC-7 [6] and TREC-6 [7].

2 Document and Query Collections

The experiments described in this paper were based on two sets of spoken document collections provided by the Linguistic Data Consortium (LDC). Most of the spoken documents consist of speech segments on a variety of topics. At times, there is simultaneous speech spoken by different speakers as well as background music at varying levels. Other resources used are the pronunciation dictionary used to translate the words to phonemes, and phoneme recognisers.

SDR collection from TREC-7 The TREC-7 SDR track test collection consists of 100 hours of news broadcast [4]. The full TREC-6 SDR track collection was used as the training corpus. There were approximately 2870 documents, containing about 23,000 unique words, with an average length of 275 words. For this collection, another 2580 words were added to the pronunciation dictionary to transform all words, including query words, to phoneme sequences. In practice, this is done via an automated process [23]. Here, four versions of the transcriptions were used for retrieval:

- Reference (REF-97), human transcribed reference transcripts, assumed to have little or minimum recognition error.
- Baseline 1 (WORD$_1$-97), recognised transcript from CMU (33.8% word error rate).
- Baseline 2 (WORD$_2$-97), recognised transcript from CMU (46.6% word error rate).
- Phoneme (PHN-97), recognised phoneme-based transcript of our own, built using the HTK system, described later (45% phoneme error rate).

The figures for the word error rate were obtained from the overview given at TREC-7 [8]. The first three sets of transcripts are word-based documents while the last set of transcripts is phoneme-based. The retrieval task here is the traditional ad-hoc relevance task, where topics and assessments were generated by human assessors. In this task, there was an average of about 17 relevant documents per query.

The query set consists of 23 queries, with an average length of 16 words. In these experiments, the original set of queries is denoted as (*full*); a stopped derivation is denoted as (*stopped*). A stoplist of approximately 370 words was used to derive the stopped queries. The average length of stopped queries was 9 words.

SDR collection from TREC-6. The TREC-6 SDR collection was taken from the LDC 1996 Broadcast News corpus, as used in the DARPA 1996 "Hub-4" experiments [2, 3]. There were approximately 1450 documents in 50 hours of broadcast news. The document collection consisted of 18,000 unique words, 2000 of which were not in the pronunciation dictionary.

Three versions of the transcriptions were used for experimentation:

- Reference (REF-96), human-transcribed reference transcripts.
- Baseline (WORD-96), recognised transcript from IBM, with a 50% word error rate [8].
- Speech (PHN-96), recognised phoneme-based transcript from ETH (Swiss Institute of Technology), with a 45% phoneme error rate.

Both the Reference (REF-96) and Baseline (WORD-96) collections are word-based while the Speech (PHN-96) collection is phoneme-based. The retrieval task here is a known-item search, where it is assumed that there is only one relevant document per query. For this TREC-6 collection, a set of 49 queries were used. The stoplist of about 370 words used in TREC-7 was applied here to stop the query set. The average length of the full queries was 12 words, while stopped queries averaged 6 words.

Phoneme recognisers. Each of the TREC-7 and TREC-6 collections had a phoneme transcript, but from different phoneme recognisers. The recognisers were developed using the HTK toolkit [28]. Both used a speaker-independent phoneme recogniser based on hidden Markov models [15].

For TREC-6, the phoneme recognition system used was developed at the Swiss Institute of Technology. Models used included acoustic models for 40 monophones, trained using the TIMIT speech corpus, and context-dependent biphone models, trained on the TREC-6 SDR training collection. The recognition process used a stochastic phone-bigram language model to eliminate generation of less probable phone sequences. These phone sequences were then post-processed by clustering acoustically similar monophones into 30 broader classes called phonemes [11]. A recognition accuracy of about 45% was obtained when evaluated on 7.5 hours of the TREC-6 SDR training collection.

For TREC-7, the models used were 39 monophones and around 1500 right-context biphones. These models were trained on about 19 hours of clean TREC-6 SDR data. The right-context models were chosen instead of the left-context models because informal experiments showed that the right-context models achieved better recognition performance. A backoff bigram language model was also used to increase recognition performance. Recognition accuracy of about 45% was obtained when evaluated on about 5 hours of the TREC-7 SDR training collection.

Translation of words to phonemes using a pronunciation dictionary. The CMU pronunciation dictionary [5] was used to translate words to phoneme sequences for all the documents and queries. This approach required a large pronunciation dictionary, which must include all the words in the spoken document collection. For TREC-6, the original pronunciation dictionary had about 110,000 entries and 2,000 words had to be added from the test collection. A year later, for TREC-7, an additional 2,580 words from the test set had to be added. The kinds of words that had to be added were mainly compounds, including names as well as inflections of words already in the dictionary. For compound words, pronunciation entries were concatenated; inflections of known words were added by modifying the phoneme sequences of the suffix to those of the required inflections. Unknown words such as names were added by first using a rule-based word-to-phoneme algorithm before converting to the same phoneme set as that used in the dictionary. These OOV words did not have a significant impact on the retrieval process because they were not used in the queries. None of the queries contained OOV words.

Experiments using documents translated to phonemes are shown with the extension (*-phn*). Queries, both (*full*) and (*stopped*), were translated to phonemes using the same dictionary.

Creation of n-grams from phonemes. To allow matching of phoneme strings we represented them as n-grams. Any sequence of symbols can be transformed into a sequence of n-grams; for example, transforming the sequence "ABCDEF" into a sequence of 3-grams yields "ABC BCD CDE DEF". We varied n-gram size between 2 and 5 inclusive; documents and queries in phoneme n-gram form are shown with an extension of (*-phn-N*) where N denotes the size of the n-gram.

Two types of n-gram sequence were formed for each type of query. The first form had n-grams created across word boundaries, so that the information as to where each word started or ended was removed. The second form disallowed creation of n-grams across word boundaries. Therefore, n-grams were only formed within each word's phoneme sequence. These are denoted as (*bound*). Bounded n-grams of stopped queries (*stopped,bound*) were similarly created.

3 Retrieval System

In a retrieval environment, queries and documents can be represented by vectors in a high-dimensional vector space. The similarity between a query q and document d can then be estimated by the cosine measure. One formulation that is effective for word-based retrieval [16] is:

$$sim(q,d) = \frac{\sum_{i \in q \wedge d}(w_{q,t} \cdot w_{d,t})}{\sqrt{\left(\sum_{t \in q}(w_{q,t})^2 \cdot \sum_{t \in d}(w_{d,t})^2\right)}}, \tag{1}$$

where $w_{x,t}$ is defined as the weight of term t in either document d or query q. The weight of an index term t in document d is:

$$w_{d,t} = \log_2(f_{d,t} + 1), \tag{2}$$

where the term frequency $f_{d,t}$ is the number of occurrences of term t in document d. Similarly, $w_{q,t}$ is defined as:

$$w_{q,t} = \log_2(\frac{N}{f_t} + 1), \tag{3}$$

where N is the total number of documents in the collection and f_t is the number of documents containing term t. This measure was used for the TREC-6 SDR experiments [7]. The equivalence of this cosine measure in the SMART system is approximated to lxc.btx [16, 18].

Another similarity measure is the Okapi formulation [22], defined as:

$$sim(q,d) = \sum_{t \in q \wedge d} \frac{(k_1+1) \cdot f_{d,t}}{k_1 \cdot [(1-b) + b \cdot \frac{W_d}{avr\_W_d}] + f_{d,t}} \cdot \frac{(k_3+1) \cdot f_{q,t}}{k_3 + f_{q,t}} \cdot \log \frac{N - f_t + 0.5}{f_t + 0.5}, \tag{4}$$

where k_1, k_3 and b are constants, set in our experiments to 1.2, 1000 and 0.75 respectively (as suggested by the City University group [22]). The value W_d is the length of document

d in bytes and avr_W_d is the average document length (in bytes) across the collection. The value N is the total number of documents in the collection, f_t is the number of documents containing term t, and $f_{x,t}$ is the frequency of term t in either document d or query q. This Okapi measure was used for the TREC-7 SDR experiments [6]. The MG [26] text engine was used for all retrieval experiments. The more complex Okapi formulation (4) was preferred to the simpler similarity formula (1) because previous experiments found that it led to improved retrieval performance.

The parameters used in the Okapi formulation were based on ad-hoc text collections, several times the size of the collections used here. The parameters were tailored for TREC ad-hoc collections and had not been tested on the TREC SDR collections. The parameters used in the current Okapi formula (4) may not be suitable on the SDR collections, especially in the phoneme n-gram forms, and in practice may need to be modified.

4 Experimental Questions

The primary objective of the experiments described in this paper were to investigate whether techniques that usually improve retrieval performance for text-based collections can improve retrieval performance for spoken document retrieval using phoneme n-grams.

A technique used for word-based retrieval is to remove stopwords, usually consisting of high-frequency function words such as conjugations, prepositions and pronouns. These words do not contribute much to the overall weight of the document or query. If a general stopword list is used in a non-typical document collection, then it is possible that aggressive stopword removal may degrade retrieval effectiveness. Tested on the TREC ad-hoc collection [7], we found that stopping improved retrieval effectiveness by about 12% (after casefolding). From a phoneme-based SDR prospective, it is not possible to apply stopping to the documents, but it can be applied to the text-based queries. In the experiments conducted here, a stopword list of about 370 words was used to stop the queries. The query sets, both stopped (*stopped*) and non-stopped (*full*), were translated to phonemes before being converted to n-grams.

Documents recognised directly as phonemes do not have word boundary information but the queries, which are in textual word form, do. The effect of word boundaries in queries can be tested by not permitting phoneme n-grams to be formed across query words. It is possible that phoneme n-grams formed across word boundaries are causing too many false matches. An example of how word boundary information should aid retrieval is the query phrase “olympic torch”, whose phoneme equivalent is “aWlimpik tOrJ”. Assuming no recognition error for the phrase and retrieval using 3-grams, the gram “ikt” could possibly match the word “predictably”, whose phoneme transcription is “pridiktabliw”.

The lengths of phoneme n-grams of the documents and queries were varied from 2 to 5. The effects of different sizes of n-grams on retrieval effectiveness were investigated, and the combination of phoneme n-grams of different lengths was also tested. For each document, n-grams of differing sizes were extracted, and the resulting sequences of n-grams were concatenated together to form one large document containing n-grams of different lengths. Various combinations were tried, varying from combination of two n-gram lengths to combining all n-gram lengths. By combining n-grams of different length, we can investigate whether different n-gram sizes provide different forms of evidence for retrieval [9]. Combination permitted us to investigate whether boundary information and stopping affected retrieval on larger collections.

When phoneme recognition was performed on the training collection of TREC-6 SDR, we found that, for each recognised word, similar recognition errors were occurring throughout the training collection. Therefore, it is likely that similar recognition errors should occur in the test collection. On the smaller TREC-6 test collection, we experimented with augmentation of the original query terms by a list of their incorrectly recognised forms, obtained from the training collection, selected based on their frequency of occurrence in the training collection. From the training documents in words, a list of documents containing the query words was found. By manually going through the same documents recognised as phonemes, we were able to detect the incorrectly recognised query sequences. For example, the query word "olympic", whose phoneme equivalent is (in one representation of the international phonetic alphabet) "aWlimpik", was incorrectly recognised as "ilimpik" and "awTawbik". This set of augmented queries is labelled as *err*. N-grams of the augmented queries were created. The rationale for this is that the incorrect transcription may be able to match relevant documents, which may also contain the incorrect transcription. This method is similar in concept to using a confusion matrix based approach by Wechsler [23] on the training collection, which can be used to determine which recognised phoneme is most likely to be recognised incorrectly as another. This technique, though not 100% accurate, is the only feasible approach for a larger collection.

Another method of query expansion, which in contrast does not make use of the training collection, is to find all n-grams, or neighbour terms, that differs in at most one character from the query n-gram. This approach used the technique of string edit distance based on phoneme distances to find neighbour terms. It had been shown to be useful for name matching on text [29]. Although this technique is likely to increase the false alarm rate by exhaustively finding all possible neighbour n-grams, it allowed us to preliminarily test the feasibility of retrieval using approximate string matching techniques without prior information of the recognition process. Idiosyncrasy in the recognition process will not unduly affect retrieval, in contrast to the previous expansion technique. These neighbour terms are added to the original set of terms. This new set of queries is labelled *nbr*. This approach does not require recognition-dependent information such as confusion matrices. This tentative experiment was conducted only on the smaller TREC-6 SDR collection.

5 Results

Our experiments use the document sets described in Section 2 with the query variations based on stopping (*full* or *stopped*), word boundaries (*bound* or otherwise), n-grams of different sizes, n-gram combination of different sizes, and expansion (*err* or *nbr*). Table 1 summaries the types of documents used in the experiments. To be consistent with TREC [20] evaluation methods, we retrieve up to 1,000 documents per query. The retrieved documents are ranked according to their likely relevance based on the similarity measures described in Section 3.

Retrieval effectiveness is most often compared in terms of precision (the proportion of retrieved documents that are relevant) at different and fixed levels of recall (the proportion of relevant documents that have been retrieved). We use average precision (AP) as one measure of effectiveness, which is precision across all the queries. This is appropriate in the TREC-7 environment. Given that the task in TREC-6 SDR was a known-item search, the queries were assumed to have only one relevant document; effectiveness can therefore be determined as the reciprocal rank of the relevant document. To compare the different parameters of each retrieval experiment, the mean reciprocal rank (MRR) is computed. Another performance measure is to calculate the total number of relevant documents retrieved in the top 5 or top 10 returned documents across all queries.

5.1 Results on TREC-7 data

Results for word-based retrieval using both full and stopped queries are shown in Table 2. Documents were neither stemmed nor stopped by default. When documents were stemmed, queries were similarly stemmed prior to retrieval, to prevent mismatches between words in the documents and queries. The individual and combined effects of stopping and stemming on retrieval were investigated. Among the different transcripts, similar trends in retrieval performance were observed when documents were stopped, stemmed, or both. From Table 2, it can be seen that retrieval degraded as the word error rate increased, for both full and stopped queries. When the documents were not stopped and stemmed, a small improvement was observed using stopped queries. It was also better to use stopped queries when documents were stopped and stemmed.

Average precision results using the phoneme n-gram representations of the documents and queries are shown in Table 3. Phoneme n-grams of the word-based documents were allowed to cross word boundaries. The documents were not stopped or stemmed prior to translation to phonemes. N-grams of queries, either stopped or full, were formed with and without crossing word boundaries. Comparing the average precision figures of Table 2 with Table 3, we can see that phoneme n-gram retrieval is much less effective than word-based retrieval. For direct phoneme recognised documents (PHN-97), retrieval was significantly lower than for word-based recognised documents. Phoneme n-gram matching of word-based documents increased the number of false matches, because a word sequence in phonemes had been broken into several n-grams and there had been partial matching. Experiments using phoneme 2-grams have consistently shown them to be much poorer and we do not comment on 2-grams further.

In Table 3, the effects of using stopped query terms in phoneme n-gram retrieval were different to those observed for word-based retrieval. In most of the cases, stopped queries did not perform as well as the full queries. In additional experiments we investigated how stopping word-based documents prior to converting to phoneme n-grams would affect retrieval; average precision results are shown in Table 4. Retrieval effectiveness, comparing different query strategies, was slightly improved from that shown in Table 3 for phoneme 3-grams for all types of transcripts, and for most cases with 4-grams. As in the word-based retrieval case, the effect of using stopped queries on retrieval effectiveness is relatively small compared to other factors, such as stopping documents, varying n-gram sizes, or using boundary information. Stopping documents is more effective, but circumstances do not always permit it. Stopping of queries on documents recognised directly as phonemes had little effect.

Table 3 shows results with phoneme n-grams of documents created across word boundaries. The effect of word boundary information in queries was investigated using both the full (*full,bound*) and stopped (*stopped,bound*) queries. Retrieval degraded using these queries except, marginally, in some cases of phoneme 3-grams, on the first baseline (WORD_1-97) and on the direct phoneme recognised (PHN-97) documents. Additional experiments were used to investigate the effect of boundary information in documents for phoneme n-gram retrieval. Results are shown in Table 5. In almost all cases of using bounded queries, both (*full,bound*) and (*stopped,bound*) retrieval was more effective on bounded documents. Again, we see that boundary information in the documents for phoneme n-gram retrieval has a greater influence than in the queries. For documents recognised as phonemes (PHN-97), retrieval was improved using bounded queries when phoneme 4-grams were used. However, word boundary information is not available in documents recognised directly as phonemes.

Experiments on the combined effect of word boundary and stopping in documents yielded the average precision figures shown in Table 6. By comparing the results in

Table 6 to Tables 3, 4, and 5, we can observe that word boundaries had a stronger effect than stopping in the documents.

The effect of varying phoneme n-gram sizes was also observed. No particular n-gram size performed well across the different types of queries on all versions of the document collections. For word-based documents translated to phoneme n-grams, the effects of word boundary information and stopping on different n-gram sizes were difficult to evaluate. Depending on the retrieval strategies, whether the queries were stopped or bounded, retrieval performance varied between 3-grams and 4-grams. Overall, 3-grams and 4-grams gave better retrieval results than 2-grams and 5-grams.

With the TREC-7 SDR collection, we combined evidence from phoneme n-grams of different lengths [6]. For each document, phoneme n-grams of different sizes were, then all n-grams were concatenated to give one large document, which was then indexed in the usual way. Phoneme n-grams of queries were combined in a similar manner. Extensive experiments were used to test the combination of phoneme n-grams ranging from 2-grams to 5-grams. We varied the combination from combining two types of n-gram sizes to combining four types of n-gram sizes (that is, adding 2-, 3-, 4-, and 5-grams together). The best combination we found in our experiments was the combination of 3-grams and 4-grams, reported in Table 7. For the case of combination, bounded queries do not in general improve retrieval although there is a slight improvement for phoneme-recognised documents. The effects of stopping and word boundary information in documents had a greater effect than in the queries. Retrieval was more effective on bounded documents using bounded queries and stopped queries were more effective on stopped documents.

In Section 3 we discussed better modelling of document length by changing the b parameter in the Okapi measure. Tentative experiments were conducted, varying the parameter between 0.5 and 1.0. We were unable to identify a suitable value, except that values between 0.75 and 0.83 were giving comparable retrieval performance to those reported here.

The experiments on the TREC-7 SDR data showed that retrieval was ineffective for direct phoneme recognised documents, compared to word-based documents. The effects of stopping and word boundary information in word-based documents were more significant than those in the queries. Bounding phoneme n-grams in the query set slightly reduced retrieval effectiveness. In terms of n-gram sizes, we found that phoneme n-grams of 3 and 4-grams retrieved better than 2-grams and 5-grams. TREC-7 SDR consists of only about 2800 documents and 23 queries.

The collection is too small to draw any firm conclusions from the experiments, but the results are indicative. We used the Wilcoxon significance test to explore the results further. These showed that stemming was consistently significantly superior to the combination of stemming and stopping, but that most of the other differences were not statistically significant.

5.2 Results on TREC-6 data

Experiments using TREC-6 SDR data are shown in Tables 8 to 12. Documents represented as words retrieved using both full and stopped queries for both automatic (that is, WORD-96) and manual (that is, REF-96) recognised spoken documents are shown in Table 8. MRR results with using phoneme n-gram representations to query the REF-96 and WORD-96 versions, and the phoneme-recognised documents (PHN-96), are shown in Table 9, while the number of documents retrieved in the top 5 and 10 ranks are shown in Table 10.

Comparing the MRR of Table 8 to Table 9, once again word-based retrieval was shown to be more effective than phoneme n-gram retrieval. An interesting result was observed by comparing effectiveness of word-based and phoneme n-gram retrieval on the automatically word-recognised version (that is, WORD-96). Comparing the results for WORD-96 in Table 8 to Table 9, the effectiveness of phoneme 3-grams was greater than that of word-based retrieval. Similar improvements were also observed using phoneme 3-grams and 4-grams based on stopped queries, as shown in Table 9. Although the number of times the relevant documents retrieved in the top 10 was reduced, the use of phoneme n-grams did improve the ranks of relevant documents retrieved. Analysis showed that by, translating words to phonemes, incorrectly recognised words were translated to phoneme sequences that were similar to the phoneme sequence of the correct word. This result supports those of by Witbrock and Hauptmann [25] in the Informedia project, where they combined word-based transcript and phoneme transcripts to improve retrieval performance. That is, phoneme n-gram retrieval has the capacity to correct some of the word-based recognised documents. Therefore, by taking the n-grams of the phonemes, relevant documents were retrieved at higher ranks.

Stopping did not lead to retrieval of more relevant documents or improve the ranks of relevant documents. We found that some relevant documents were being retrieved using phrasal linkage between stopwords and keywords.

Given that phoneme recognition has a high error rate, a large proportion of incorrectly recognised phonemes is inherent to spoken document retrieval based on a phoneme recogniser. As discussed before, following our belief that transcription errors of the same word in the training collection would also occur in the test collection, the original query was augmented by a set of potential match terms obtained from the training collection. These terms were chosen based on their frequency of occurrence. Table 11 shows the retrieval performance using a combination of the initial query with the mistranscribed terms. Overall effectiveness significantly degrades with the use of these additional terms. This technique of query expansion is similar in concept to using confusion matrices and calculating the probability of recognition error between phonemes to determine the phoneme sequences to use. Experiments by Wechsler [23] and Ng [12] using dynamic programming techniques have shown that the confusion matrix approach can be effective.

Experiments with using neighbour terms to expand the query set, shown in Table 12, showed that the technique was a complete failure. It did not use the phoneme similarity (or otherwise) of the substituted characters. By using all the terms that differed in one character from the original query term, many incorrect or irrelevant terms were included. A more considered approach to selecting additional query terms—that takes into account the similarity of the phonemes involved—appears to be required.

Although these results consistently show word-based retrieval to be more effective than phoneme-based retrieval for SDR, limitations of the document and query collections may have affected these results. These collections are not an environment in which phoneme-based retrieval is likely to be favoured. In comparison to SDR in practice, recognition processes will be less reliable. Furthermore, it is possible that the collection is too small, so that the results may not generalise for larger collections. Translating words to phonemes using a pronunciation dictionary required unknown words from the collections to be manually added. This was a time-consuming task and would not occur in a practical system. A rule-based algorithm for translating words to phonemes would have been less costly, and does not require additional work for unknown words.

6 Conclusions and Future Work

We have explored methods of spoken document retrieval based on phoneme n-grams. We investigated the effects of word boundary information, stopping of queries, varying n-gram sizes, combination of phoneme n-grams and query expansion using incorrectly transcribed phoneme sequences and similar query terms. Two sets of spoken documents were used to compare retrieval performance. For each set, representations of the documents were generated from manual transcription, automatic word-based recognition, and automatic phoneme-based recognition; the word-based documents were translated to phonemes using a pronunciation dictionary. Queries were similarly translated, using the pronunciation dictionary. These experiments confirm that phoneme-based retrieval is less effective than word-based retrieval, but is nonetheless reliable enough to be used in practice.

When translating queries into phoneme n-grams, we tested the effect of using or ignoring word boundaries. We found that word boundary information in the queries did not have much impact. Stopping the queries did not improve retrieval, in contrast to retrieval experiments on text-based collections. Word boundary and stopping the word-based documents prior to conversion to phoneme n-grams had a greater impact on retrieval effectiveness. Such processing cannot, of course, be applied to documents recognised directly as phonemes. Phrasal retrieval may be useful for retrieval when a phrase is formed by a stopword and keyword, and this was only observed in the TREC-6 SDR collection.

Varying phoneme n-gram of 3-grams and 4-grams led to some improvement in effectiveness in comparison to other n-gram sizes. The combination of 3-grams and 4-grams was also more effective than other combinations in the TREC-7 collections. Other combinations did not perform as well, apparently because shorter n-grams were retrieving more irrelevant documents and the longer n-grams were not finding some relevant documents.

It would be expected that transforming word sequences into phoneme sequences would lose information and thus result in deterioration of retrieval. This expectation was confirmed in experiments with the TREC-7 data, but there was no difference for the TREC-6 data. One explanation is that the task of finding a single item is much easier than that of finding all relevant documents, so that the TREC-7 experiment is more able to separate techniques based on different hypotheses. The TREC-6 experiments indicated that occasionally word errors are remedied if the word and its inaccurate replacement both map to a similar phoneme sequence.

Query expansion was tested using two different approaches. The first used information about recognition errors derived from a training collection: the most frequent incorrect phoneme sequences generated for the query words was used to augment the original queries. The second approach was to use neighbour n-grams obtained using string edit distance. Each query term was augmented by a list of potential substitutes from a list of unique terms from the test collection. Both methods of query expansion degraded effectiveness, with the—admittedly simplistic—neighbour method in particular giving very poor results. Although the first approach performed better, in some individual cases improving effectiveness, it required information from the training collection that would not always be available.

The high quality recognition environment characterized by the TREC experiments is not always available. Degraded recognition due to noise and uncertainty as well as the lack of key words in the dictionary can lead to poor recognition. Proper names are often the most important terms in a query and are generally not available in pronunciation dictionaries. Each of these conditions needs further experimentation, and particularly in experimental environments that reflect the difficult conditions that so often characterise the speech retrieval environment. The experiments here on phoneme n-grams investigated some of the retrieval techniques which might be useful in such conditions.

Acknowledgments

Thanks to Peter Schäuble, Eugene Munteanu and Martin Wechsler of the Swiss Institute of Technology (ETH) for providing phoneme transcriptions of the spoken document collections. Thanks also to Michael Fuller and Victor Poznański. We are particularly grateful to the anonymous referees for their detailed, constructive comments. This work was supported by the Australian Research Council.

References

[1] W. B. Cavnar. Using an n-gram-based document representation with a vector processing retrieval model. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 269 – 277, 1994.

[2] Linguistic Data Consortium. Continuous speech recognition corpus-V: 1996 broadcast news speech (CSR-V hub-4), CD-ROM. ldc@ldc.upenn.edu, Philadelphia, PA 19104-2608, USA, 1996.

[3] Linguistic Data Consortium. DARPA continuous speech recognition corpus-IV: Radio broadcast news(CSRIV hub-4), CD-ROM. ldc@ldc.upenn.edu, Philadelphia, PA 19104-2608, USA, 1996.

[4] Linguistic Data Consortium. Continuous speech recognition corpus-VI: 1997 broadcast news speech (CSR-VI hub-4), CD-ROM. ldc@ldc.upenn.edu, Philadelphia, PA 19104-2608, USA, 1997.

[5] Carnegie Mellon University Pronouncing Dictionary. Cmudict.0.4. Available from http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 1995.

[6] M. Fuller, M. Kaszkiel, D. Kim, C. L. Ng, J. Robertson, R. Wilkinson, M. Wu, and J. Zobel. TREC 7 Ad Hoc, Speech, and Interactive tracks at MDS/CSIRO. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 465 – 474, 1998.

[7] M. Fuller, M. Kaszkiel, C. L. Ng, P. Vines, R. Wilkinson, and J. Zobel. MDS TREC-6 report. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 241 – 258, Nov 1997.

[8] J. Garofolo, E. Voorhees, C. Auzanne, V. Stanford, and B. Lund. 1998 TREC-7 spoken document retrieval track overview and results. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 79 – 90, Nov 1998.

[9] G. J. F. Jones, J. T. Foote, K. Sparck Jones, and S. J. Young. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the Nineteenth ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 30 – 38, 1996.

[10] K. F. Lee. *Automatic Speech Recognition, The Development of the SPHINX System*. Kluwer Academic Publishers, 1989.

[11] B. Mateev, E. Munteanu, P. Sheridan, M. Wechsler, and P. Schäuble. ETH TREC-6: Routing, chinese, cross-language and spoken document retrieval. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 623 – 636, Nov 1997.

[12] K. Ng. Towards robust methods for spoken document retrieval. In *Proceedings of International Conference on Spoken Language Processing*, volume 3, pages 939 – 942, Dec 1998.

[13] K. Ng and V. Zue. Phonetic recognition for spoken document retrieval. In *Proceedings of International Conference on Acoustic Speech and Signal Processing ICASSP*, pages 325 – 328, 1998.

[14] K. Ng and V. W. Zue. Subword unit representations for spoken document retrieval. In *Proceedings of the European Conference on Speech Communications and Technology, EUROSPEECH*, pages 1607 – 1610, Rhodes, Greece, 1997.

[15] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[16] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513 – 523, 1988.

[17] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, 1983.

[18] A. Singhal. AT&T at TREC-6. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, NIST Special Publication 500-240, pages 215–226, November 1997.

[19] A. F. Smeaton, M. Morony, G. Quinn, and R. Scaife. Taiscéalaí: Information retrieval from an archive of spoken radio news. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (EuroDL)*, Sep 1998.

[20] E. Voorhees and D. Harman. Overview of the Sixth Text REtrieval Conference. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 1 – 24, Nov 1997.

[21] E. Voorhees and D. Harman. Overview of the Seventh Text REtrieval Conference. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 1 – 24, Nov 1998.

[22] S. Walker, S. E. Robertson, M. Boughanem, G. J. F. Jones, and K. Sparck Jones. Okapi at TREC-6 Automatic ad hoc, VLC, routing, filtering and QSDR. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 125 – 136, Nov 1997.

[23] M. Wechsler, E. Munteanu, and P. Schäuble. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the Twenty First ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 20 – 27, Aug 1998.

[24] M. Wechsler and P. Schäuble. Speech retrieval based on automatic indexing. In *Workshop in Computing Science-MIRO*. Springer Verlag, Oct 1995.

[25] M. J. Witbrock and A. G. Hauptmann. Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents. In *Proceedings of the Digital Library Conference*, pages 30 – 35, Philadelphia, PA, USA, July 1997.

[26] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and indexing documents and images*. Van Nostrand Reinhold, 1994.

[27] P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchev, and S. J. Young. The 1994 HTK large vocabulary speech recognition system. In *Proceedings of International Conference on Acoustic Speech and Signal Processing ICASSP*, 1995.

[28] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK Book*. Entropic Cambridge Research Laboratory, 1995.

[29] J. Zobel and P. Dart. Phonetics string matching: Lessons from information retrieval. In *Proceedings of the Nineteenth ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 166 – 172, 1996.

REF-97	human transcribed reference transcripts (TREC-7)
WORD_1-97	recognised transcript of 33.8% word error rate (TREC-7)
WORD_2-97	recognised transcript of 46.6% word error rate (TREC-7)
PHN-97	recognised phoneme-based transcript of 45% phoneme error rate (TREC-7)
REF-96	human transcribed reference transcripts (TREC-6)
WORD-96	recognised transcript of 50% word error rate (TREC-6)
PHN-96	recognised phoneme-based transcript of 45% phoneme error rate (TREC-6)

Table 1: *Types of documents used in experiments.*

Document set	Query set	
	full	stopped
REF-97	0.389	0.395
REF-97 *+stopping*	0.396	0.396
REF-97 *+stemming*	0.446	0.454
REF-97 *+stopping +stemming*	0.412	0.455
WORD_1-97	0.310	0.319
WORD_1-97 *+stopping*	0.318	0.318
WORD_1-97 *+stemming*	0.405	0.417
WORD_1-97 *+stopping +stemming*	0.364	0.420
WORD_2-97	0.245	0.253
WORD_2-97 *+stopping*	0.255	0.255
WORD_2-97 *+stemming*	0.340	0.326
WORD_2-97 *+stopping +stemming*	0.274	0.326

Table 2: *Average precision, for word-level matching with different query types. When documents are stemmed, the queries are similarly stemmed, to prevent mismatches of words. TREC-7 data.*

Document set (unbounded phoneme n-grams)	Query set full	full,bound	stopped	stopped,bound
Experiments with phoneme 2-grams				
REF-97	0.104	0.156	0.117	0.148
WORD_1-97	0.071	0.134	0.106	0.127
WORD_2-97	0.061	0.108	0.095	0.104
PHN-97	0.017	0.028	0.020	0.027
Experiments with phoneme 3-grams				
REF-97	0.307	0.303	0.305	0.302
WORD_1-97	0.260	0.276	0.256	0.261
WORD_2-97	0.205	0.202	0.213	0.207
PHN-97	0.042	0.056	0.050	0.059
Experiments with phoneme 4-grams				
REF-97	0.323	0.301	0.300	0.299
WORD_1-97	0.294	0.253	0.256	0.251
WORD_2-97	0.227	0.187	0.198	0.187
PHN-97	0.101	0.098	0.085	0.098
Experiments with phoneme 5-grams				
REF-97	0.293	0.258	0.265	0.257
WORD_1-97	0.251	0.213	0.217	0.213
WORD_2-97	0.209	0.153	0.168	0.153
PHN-97	0.076	0.075	0.070	0.076

Table 3: *Average precision, for different speech recognition processes, gram lengths, and query types. Documents were not stopped nor stemmed prior to conversion to phoneme n-grams. TREC-7 data.*

Document set *+stopped* (unbounded phoneme n-grams)	Query set full	full,bound	stopped	stopped,bound
Experiments with phoneme 2-grams				
REF-97	0.094	0.132	0.163	0.169
WORD_1-97	0.074	0.098	0.132	0.143
WORD_2-97	0.070	0.082	0.118	0.121
Experiments with phoneme 3-grams				
REF-97	0.308	0.310	0.319	0.308
WORD_1-97	0.271	0.280	0.272	0.275
WORD_2-97	0.218	0.214	0.226	0.219
Experiments with phoneme 4-grams				
REF-97	0.316	0.310	0.313	0.307
WORD_1-97	0.266	0.258	0.255	0.257
WORD_2-97	0.211	0.196	0.207	0.195
Experiments with phoneme 5-grams				
REF-97	0.274	0.256	0.262	0.253
WORD_1-97	0.215	0.212	0.207	0.212
WORD_2-97	0.174	0.154	0.170	0.154

Table 4: *Average precision, for word-based documents which were stopped prior to conversion to phoneme n-grams, varying gram lengths, and query types. TREC-7 data.*

Document set (bounded phoneme n-grams)	Query set			
	full	full,bound	stopped	stopped,bound
Experiments with phoneme 2-grams				
REF-97	0.107	0.177	0.130	0.173
WORD_1-97	0.091	0.146	0.102	0.138
WORD_2-97	0.076	0.104	0.086	0.098
Experiments with phoneme 3-grams				
REF-97	0.293	0.327	0.302	0.320
WORD_1-97	0.253	0.284	0.267	0.282
WORD_2-97	0.198	0.226	0.208	0.223
Experiments with phoneme 4-grams				
REF-97	0.298	0.335	0.304	0.322
WORD_1-97	0.260	0.276	0.259	0.270
WORD_2-97	0.195	0.210	0.194	0.206
Experiments with phoneme 5-grams				
REF-97	0.259	0.301	0.254	0.295
WORD_1-97	0.213	0.247	0.210	0.237
WORD_2-97	0.157	0.193	0.157	0.185

Table 5: *Average precision, for word-based documents converted to bounded phoneme n-grams where n-grams were not formed across word boundaries, varying gram lengths, and query types. TREC-7 data.*

Document set *+stopping* (bounded phoneme n-grams)	Query set			
	full	full,bound	stopped	stopped,bound
Experiments with phoneme 2-grams				
REF-97	0.105	0.159	0.147	0.181
WORD_1-97	0.089	0.117	0.121	0.155
WORD_2-97	0.078	0.101	0.097	0.125
Experiments with phoneme 3-grams				
REF-97	0.305	0.329	0.306	0.329
WORD_1-97	0.261	0.297	0.274	0.292
WORD_2-97	0.206	0.229	0.218	0.230
Experiments with phoneme 4-grams				
REF-97	0.299	0.328	0.300	0.328
WORD_1-97	0.259	0.274	0.262	0.272
WORD_2-97	0.195	0.209	0.192	0.207
Experiments with phoneme 5-grams				
REF-97	0.260	0.296	0.253	0.296
WORD_1-97	0.210	0.237	0.207	0.235
WORD_2-97	0.158	0.184	0.158	0.186

Table 6: *Average precision, for word-based documents which were stopped prior to conversion to bounded phoneme n-grams where n-grams were not formed across word boundaries, varying gram lengths, and query types. TREC-7 data.*

	Document set	Query set full	full,bound	stopped	stopped,bound
(unbounded)	REF-97	0.341	0.329	0.316	0.320
	WORD$_1$-97	0.307	0.289	0.273	0.275
	WORD$_2$-97	0.231	0.205	0.211	0.208
	PHN-97	0.082	0.105	0.085	0.101
(unbounded, +stopping)	REF-97	0.322	0.326	0.329	0.329
	WORD$_1$-97	0.284	0.285	0.269	0.280
	WORD$_2$-97	0.222	0.216	0.225	0.219
(bounded)	REF-97	0.315	0.338	0.312	0.332
	WORD$_1$-97	0.273	0.293	0.269	0.285
	WORD$_2$-97	0.209	0.227	0.207	0.220
(bounded, +stopping)	REF-97	0.314	0.334	0.312	0.335
	WORD$_1$-97	0.275	0.297	0.270	0.291
	WORD$_2$-97	0.217	0.225	0.211	0.225

Table 7: *Average precision using combination of 3-grams and 4-grams, for different speech recognition processes and query types. TREC-7 data.*

Document set	Query set	Mean Reciprocal Rank	No. relevant in the top: 5	10
REF-96	full,word	0.702	41	44
REF-96	stopped,word	0.700	38	43
WORD-96	full,word	0.558	36	40
WORD-96	stopped,word	0.521	33	38

Table 8: *Retrieval effectiveness using documents represented as sequences of words, for different speech recognition processes and query types. TREC-6 data.*

Document set	Query set full	full,bound	stopped	stopped,bound
Experiments with phoneme 3-grams				
REF-96	0.698	0.616	0.629	0.564
WORD-96	0.582	0.579	0.548	0.531
PHN-96	0.204	0.230	0.198	0.222
Experiments with phoneme 4-grams				
REF-96	0.687	0.569	0.581	0.552
WORD-96	0.538	0.486	0.536	0.488
PHN-96	0.236	0.205	0.178	0.187
Experiments with phoneme 5-grams				
REF-96	0.661	0.505	0.598	0.486
WORD-96	0.542	0.453	0.504	0.452
PHN-96	0.172	0.167	0.183	0.172

Table 9: *Mean Reciprocal Rank, for different speech recognition processes, gram lengths, and query types. TREC-6 data.*

Document set	Query set: full	full,bound	stopped	stopped,bound
Experiments with phoneme 3-grams				
REF-96	(38,43)	(35,42)	(38,42)	(34,40)
WORD-96	(35,38)	(35,37)	(32,33)	(33,35)
PHN-96	(13,19)	(13,17)	(14,18)	(15,18)
Experiments with phoneme 4-grams				
REF-96	(37,43)	(32,38)	(36,40)	(31,37)
WORD-96	(32,38)	(29,35)	(30,32)	(31,32)
PHN-96	(15,18)	(12,17)	(13,17)	(12,15)
Experiments with phoneme 5-grams				
REF-96	(35,41)	(30,33)	(32,38)	(29,32)
WORD-96	(31,35)	(27,31)	(28,34)	(26,31)
PHN-96	(12,16)	(8,15)	(8,15)	(9,16)

Table 10: *Number of relevant documents retrieved in the top 5 and 10 for different speech recognition processes, gram lengths, and query types. Figures are in the form of (top 5, top 10). TREC-6 data.*

Document set	Query set	Mean Reciprocal Rank	No. relevant in the top: 5	10
REF-96-phn-3	stopped,phn-3,bound,err	0.423	29	36
REF-96-phn-4	stopped,phn-4,bound,err	0.509	31	33
REF-96-phn-5	stopped,phn-5,bound,err	0.456	28	32
WORD-96-phn-3	stopped,phn-3,bound,err	0.463	28	31
WORD-96-phn-4	stopped,phn-4,bound,err	0.479	28	29
WORD-96-phn-5	stopped,phn-5,bound,err	0.442	25	29
PHN-96-phn-3	stopped,phn-3,bound,err	0.153	10	13
PHN-96-phn-4	stopped,phn-4,bound,err	0.197	11	13
PHN-96-phn-5	stopped,phn-5,bound,err	0.148	7	10

Table 11: *Retrieval effectiveness for phoneme n-grams (-phn-n) without crossing word boundaries (bound) and with query expansion using incorrect phoneme sequences of query terms (err) for different speech recognition processes and gram lengths. TREC-6 data.*

Document set	Query set	Mean Reciprocal Rank	No. relevant in the top: 5	10
REF-96-phn-3	stopped,phn-3,bound,nbr	0.014	1	3
REF-96-phn-4	stopped,phn-4,bound,nbr	0.136	10	13
WORD-96-phn-3	stopped,phn-3,bound,nbr	0.011	0	3
WORD-96-phn-4	stopped,phn-4,bound,nbr	0.117	9	12
PHN-96-phn-3	stopped,phn-3,bound,nbr	0.014	1	1
PHN-96-phn-4	stopped,phn-4,bound,nbr	0.056	3	4

Table 12: *Retrieval effectiveness for phoneme n-grams (-phn-n) without crossing word boundaries (bound) and with query expansion using string edit distance of 1 to obtain neighbour terms, for different speech recognition processes and gram lengths. TREC-6 data.*