FACTORS AFFECTING SPEECH RETRIEVAL

Corinna Ng† *Ross Wilkinson*‡ *Justin Zobel*†

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

Email: {chienn,jz}@cs.rmit.edu.au

‡CSIRO, Division of Mathematical and Information Science
723 Swanston St, Carlton VIC 3053
Australia

Email: Ross.Wilkinson@cmis.csiro.au

Abstract

Collections of speech documents can be searched using speech retrieval, in which the documents are processed by a speech recogniser to give text that can be searched by standard text retrieval techniques. Recognition is the translation of speech signals into either words or subword units such as phonemes. We investigated the use of a phoneme-based recogniser to obtain phoneme sequences. We found that phoneme recognition is worse than word recognition, because of lack of context and difficulty in phoneme boundary detection. Comparing the transcriptions of two different phoneme-based recogniser, we found that the effects of training using well-defined phoneme data, the lack of a language model, and lack of a context-dependent model affected recognition performance. Retrieval was based on n-grams. We found that trigrams performed better than quadgrams because the longer n-gram features contained too many transcription errors. Comparing the phonetic transcriptions from a word recogniser to transcriptions from a phoneme recogniser, we found that using 61 phones modelled with an algorithmic approach were better than using 40 phones modelled with a dictionary approach.

1 INTRODUCTION

Speech retrieval is a process where speech documents are retrieved in response to a textual query. A typical strategy for retrieval requires the speech documents to be transcribed into some intermediate textual form prior to applying standard textual information retrieval techniques. The transcriptions of speech documents can be in the form of words or subword units such as phonemes or syllables.

There are several problems originating from word-based speech recognition, including non-recognition of unknown words and requiring word boundary information. Phoneme-based speech recognition, on the other hand, is noiser and less accurate.

In the experiments conducted here, we proposed a phoneme-based approach to both the recognition and retrieval processes. Retrieval is conducted using n-grams, which have the advantage of being language independent [2].

We compared the retrieval performance of our phoneme-based recognised transcriptions to two other word-based transcriptions, which were translated to phoneme sequences using either an algorithmic or dictionary approach. Recognition results indicate that phonemes are worse than words. The existence of a well-defined language model as well as phone models that modelled monophones

and context dependent phones were important in the recognition process. The initial training of the phone models also affected recognition. The class of phone sets used also influenced the recognition and retrieval tasks.

Retrieval used either trigrams or quadgrams. Our results show that trigrams were more effective at identifying relevant documents. They also indicated that retrieval using transcriptions obtained from word-based approaches performed better than the phoneme-based approach although results are better when the word-based transcriptions are translated to phoneme sequences.

This work was completed as part of the TREC experiments organised by NIST to encourage retrieval into information retrieval.

2 SPEECH RECOGNITION

Automatic speech recognition is the transcription of an acoustic signal into its textual form [4]. Two stages are involved in an pattern recognition approach. The signal processing stage parameterises the acoustic speech signal into its digital form. The acoustic modelling stage determines the textual form of the digitised speech.

The acoustic modelling stage models speech signals as words or subword units. Word-based recognition perform better using subword models as base models. It has two inherent problems. Unknown words cannot be recognised and word boundary information is necessary for word recognition to be accurate. Other factors that influence recognition accuracy of a word-based speech recogniser include the amount of training; the class of phones used; the existence and type of language models used; and whether the phone models used include context dependency models or monophones only.

A language model constrains the recognition process. It estimates the probability of occurrence of each word sequence [6]. The existence of a language model for a word-based recogniser limits the search space of potential words that can be formed from a speech segment. Without it, recognition can become highly inaccurate. Phone models should be initially trained with phone data with explicit phone boundaries known. Subsequent training can use noisy training data similar to the test data. Better recognition performance is possible with more accurate models of the speech signal. If both monophones and context dependent phone models are used, higher accuracy can also be gained.

Another approach is to perform phoneme-based recognition. Recognition is used to produce a sequence of phonemes. Unknown words are recognised as component phonemes and word boundaries are no longer necessary. The factors that affect word-based recognition also apply to phoneme-based recognition, but their effect on recognition accuracy are not as problematic as the problem of phoneme boundary detection. This is the main cause of recognition error in the phoneme-based approach.

Word-based recognition requires a large dictionary of about 20,000 words to perform well. For phoneme-based recognition only a finite phone set is required. In terms of performance, phoneme-based recognition is more inaccurate than word-based due to higher noise.

We compare the performance of two phoneme-based approaches varying several of the factors, testing how each affects recognition. In one approach, phoneme models are bootstrapped using TIMIT data, trained extensively, and include context dependency models. In another, phoneme models used were monophones and limited training was provided using noisy Speech TREC training data. The recognition and training of the first task was accomplished by the Swiss Institute of Technology.

For the recognition tasks, we used the HTK [10] speech toolkit, developed at Cambridge.

3 INFORMATION RETRIEVAL

Information retrieval (IR) has been defined as the process of obtaining a set of relevant documents using a set of natural language textual queries [7]. The documents and queries are usually text-based. A document is deemed relevant when it satisfies the user's query requirement. In a textual IR context, documents and queries are both stopped and stemmed prior to retrieval.

Sentence : this is an Ngram Phoneme equivalent : dIS IZ aN ENGRaM trigrams created: dIS ISI SIZ IZa ZaN aNE NEN ENG NGR GRa RaM

Figure 1: An example of a trigram

The matching between a document and query is usually accomplished by calculating the similarity between them. The cosine similarity function is such an example. This calculation predicts the relevance of the document to the query depending on the existence of the query terms within the document collection. The similarity between a document d and a query q can be calculated as follows,

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

where w_{d_t} and w_{q_t} are defined as the weights of the word t in either the document d or query q. The weights of an index word t in document d is defined as

$$w_{d,t} = \log_2(\mathrm{F}_{d,t} + 1) \quad (2)$$

where the term frequency $\mathrm{F}_{d,t}$ is defined as the frequency of occurrences of t in document d. The weights of a query word t in query q is defined as $w_{q,t} = w_t$ where w_t is the inverse document frequency, which is calculated as

$$w_t = \log_2(\frac{N_d}{f_t} + 1) \quad (3)$$

where N_d is the total number of documents in the collection and f_t is defined as the number of documents that contain t. Hence, the more frequent a word occurs in a document, the smaller the value of its weight. The similarity value calculated between a document and query can be used to rank the retrieved document.

For the retrieval experiments, most queries have only one relevant document. To evaluate the performance of the retrieval, two measures are used. They are

- Mean Rank
  The average rank at which the relevant document is retrieved across the 49 queries. Hence, the lower the value, the better the retrieval performance.
- Mean Reciprocal
  Mean of the reciprocal of the rank. This measurement penalises the non-retrieval of a relevant document, whose reciprocal will be 0 instead of $\frac{1}{2000}$. Therefore, the larger the value, the better the performance.

Precision figures are useful for comparing retrieval performance but can be easily skewed when some queries performed badly.

The retrieval system used here is MG [9]. The S-stemmer and a stop list of about 400 words are used in this set of experiments.

4 SPEECH RETRIEVAL

Speech retrieval is when speech documents are returned in an information retrieval environment given some free text queries.

Retrieval using phoneme is difficult. It is equivalent to retrieval using letters instead of words. To overcome this, we propose the use of phonetic n-grams to perform retrieval. An example of trigam retrieval is illustrated in the Figure 1 of the phrase "this is an Ngram".

N-gram retrieval removes the need for boundary information., and also removes some contextual information provided from boundaries which may be useful for retrieval. The likelihood of occurrence

of an n-gram is smaller than that of a phone. Also, the longer the n-gram sequence, the less likely it will occur within a document collection. Hence, a potential relevant document is found when a query n-gram feature is found within a document. It is also possible that substantial noise within a document causes many n-grams to be matched. Previous experiments found that the longer the n-gram used, the better the retrieval performance but at the expense of transcription errors. Hence the size of the n-gram is a tradeoff between recognition error and retrieval effectiveness.

Although phone is defined as the smallest unit of speech, there are several ways of defining a phone. The breakdown of a word into its component phones can be accomplished using several different methods each yielding different phone sets and phoneme sequences.

Retrieval-based factors include the length of the document and query in the experiments as well as the type of matching function. The longer the document, the more likely a query feature is to be found in the document even if the document is irrelevant. Hence, the retrieval system must take these into consideration when similarities are calculated. Another potential area for errors is when the document collection consists of many repetitions. This may be of importance when comparing the recognition performance but, in terms of retrieval, it could cause the weighting of indexing features to be wrong.

5 EXPERIMENTS

We compare the retrieval performance of the different recognition transcriptions and investigate the effects of both recognition and retrieval processes.

The experiments were conducted using Speech TREC [3] data obtained from the Linguistic Data Consortium. The collection consists of about 1500 documents, which is about 50 hours of news broadcasts. There is one relevant document for each of the 49 queries used in this set of experiments. Four versions of the transcriptions were used for the retrieval experiments:

- Manual transcriptions (REF)
- Automatic Word-based recognised transcriptions (SRT)
  The word-based transcription collection contained less documents than the manual transcriptions because some documents were not recognised.
- Automatic phoneme-based recognised transcriptions with about 50% accuracy (SP1)
- Automatic phoneme-based recognised transcriptions with about 10% accuracy (SP2)

The word-based recognised transcripts were provided by IBM. They had an error rate of approximately 35%. There were two phoneme-based transcriptions. SP1 was provided by the Swiss Institute of Technology. Their recogniser is a speaker-independent phoneme recogniser which uses 40 monophones and a set of context-dependent biphone models. These phone models were bootstrapped using the TIMIT speech corpus [8]. and trained using 50 hours of news broadcasts from the training collection. The second set of phoneme transcriptions was obtained from our own speech recogniser which uses 61 monophones. These phone models were not bootstrapped using TIMIT data. These two phonetic transcriptions uses different phone sets, SP1 uses those from the CMU dictionary while SP2 were those based on the Ainsworth [1] algorithm.

There are also three versions of the queries, in words and in phonemes. Since the phonetic transcriptions were based on different phone sets, two versions of the queries were produced, using the CMU dictionary and Ainsworth's algorithm.

Our baseline retrieval experiments used the manual transcriptions. Phoneme-based retrieval used trigrams. Previous experiments have found that trigram retrieval is optimal [5]. We first compared the recognition transcriptions and then investigated the type of errors that were generated. The retrieval results were compared between the phoneme transcriptions to investigate how the transcriptions errors actually affected performance.

Changing the size of the indexing features also changed the retrieval performance. This was compared between the different phonetic transcriptions as well as to the word-based transcriptions.

	Mean Rank	Mean Reciprocal
REF	5.48	0.7036
SRT	10.11	0.5207

Table 1: Baseline results using word-based transcriptions

	Mean Rank	Mean Reciprocal
REF-trigram	11.47	0.7340
SRT-trigram	23.49	0.5472
SP2-trigram	229.20	0.0046

Table 2: Phoneme-based retrieval results using 60 phones based on the Ainsworth algorithm

6 RESULTS

The phoneme-based transcription, SP1, had a recognition error rate of about 55%. SP2 had a recognition error rate of about 90%. Several factors contributed to this high error rate. The lack of extensive training and a well-defined language model prevented the parameters of the HMM models from trained accurately. SP2 was trained using only 10 hours of speech and was not bootstrapped using data containing well-defined phoneme boundaries. Furthermore, only monophones were used in the recognition process and that causes many phonemes to be mis-recognised due to their similarity to other phonemes.

Baseline retrieval results using the word-based transcriptions are shown in Table 1. It showed that there are substantial transcription errors at the word level that affect retrieval performance. Furthermore, there were several noisy queries where the relevant documents did not contain any of the query words, and hence were not retrieved.

The word-based reference and automatic transcriptions were translated to phoneme sequences. Two versions of these transcriptions were obtained using the CMU dictionary and the Ainsworth algorithm. Retrieval results of word-based transcriptions translated to phoneme sequences, using the Ainsworth algorithm and the phoneme-based transcriptions, are shown in Table 2. Compared to the word-based results, many relevant documents were not retrieved, while retrieved relevant documents were at higher ranks, as shown by the higher mean reciprocal values.

Retrieval results of trigram transcriptions obtained using the CMU dictionary are shown in Table 3. Comparing the results of REF-trigram and SRT-trigram, we could see the influence of the phone set used. The use of 61 phones modelled the documents better than using the 40 phones defined here. The impact of recognition error is clear from the retrieval results of the phoneme-based transcriptions. Relevant documents are being retrieved from the phoneme-based recognised transcriptions, although the results are significantly lower than those obtained from the word-based approach. Quad-grams performed worse, indicating that longer phoneme n-grams have a higher likelihood of errors.

7 CONCLUSION

Recognition performance between two different forms of phoneme recogniser were compared. The effects of recognition on retrieval was investigated as well as the effects of the size of the n-gram indexing features.

Phoneme recognition is more inaccurate than word recognition. The existence of a suitable language model and the types of phone models used also influenced the performance of the recognition tasks. The lack of phoneme boundary information affected the recognition performance, especially when the phone models were not explicitly initialised by well-defined phoneme data. Errors at the recognition level were propagated to the retrieval process, thus causing irrelevant documents to be retrieved.

By keeping the parameters of the retrieval processes constant, we were able to investigate the effects of the recognition tasks. We found that the algorithmic approach to translating the word-based transcripts using 60 phones performed better than those using the dictionary approach using

	Mean Rank	Mean Reciprocal
REF-trigram	28.42	0.6286
SRT-trigram	48.24	0.5475
REF-quadgram	46.84	0.6115
SRT-quadgram	68.88	0.5363
SP1-trigram	158.10	0.198
SP1-quadgram	210.37	0.178

Table 3: Phoneme-based retrieval results using 40 CMU phones

40 phones. We were unable to make the same comparison for the phoneme-based transcriptions because of the large difference in recognition error.

We compared retrieval performance of trigrams and quadgrams. We found that quadgrams did not perform as well as trigrams across all the transcriptions. This supported our argument that longer n-grams caused more noise to be matched, degrading retrieval.

The experiments conducted here is a limited study of various factors influencing recognition and retrieval of speech documents. Extensive comparison could not be accomplished between the phoneme-based recognisers and their transcriptions because of large differences in recognition accuracy and because of different phone sets used. The benefits of the phoneme-based approach could not be completely investigated because the queries consist entirely of words that exist in the pronunciation dictionary used.

8 ACKNOWLEDGMENTS

Thanks to Peter Schäuble and Eugene Munteanu and Martin Wechsler of the Swiss Institute of Technology for providing their phoneme transcriptions of the speech document collections.

References

[1] W. A. Ainsworth. A systems for converting english text into speech. *IEEE Transactions on Audio and Electroacoustics*, AU-21(3):288–290, Jun 1973.

[2] W. B. Cavnar. Using an n-gram-based document representation with a vector processing retrieval model. In *Proceedings of the Third TExt Retrieval Conference*, pages 269–277, 1994.

[3] D. K. Harman. *The Sixth Text REtrieval Conference*. Department of Commerce, National Institute of Standards and Technology, 1997. To be published.

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

[5] K. Ng and V. W. Zue. Subword unit representations for spoken document retrieval. In *Proc. ESCA Eurospeech Conference*, pages 1607–1610, Rhodes, Greece, 1997.

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

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

[8] M. Wechsler, E. Munteanu, and P. Schauble. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval*, August 1998. To appear.

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

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