TOWARDS ROBUST METHODS FOR SPOKEN DOCUMENT RETRIEVAL[†]

Kenney Ng

Spoken Language Systems Group
MIT Laboratory for Computer Science
545 Technology Square, Cambridge, MA 02139 USA

ABSTRACT

In this paper, we investigate a number of robust indexing and retrieval methods in an effort to improve spoken document retrieval performance in the presence of speech recognition errors. In particular, we examine expanding the original query representation to include confusible terms; developing a new document-query retrieval measure based on approximate matching that is less sensitive to recognition errors; expanding the document representation to include multiple recognition hypotheses; modifying the original query using automatic relevance feedback to include new terms found in the top ranked documents; and combining information from multiple subword unit representations. We study the different methods individually and then explore the effects of combining them. Experiments on radio broadcast news data show that using a combination of these methods can improve retrieval performance by over 20%.

1. INTRODUCTION

With the continuing growth in the amount of accessible data, the need for automatic methods to process, organize, and analyze this data has become increasingly important. Of particular interest is the problem of efficiently finding "interesting" pieces of information from the growing collections and streams of data. Much research has been done on the problem of selecting "relevant" items from large collections of text documents given a user query [2, 7]. Only recently has there been work addressing the retrieval of information from other media such as images, video, audio, and speech [3, 6, 8]. The development of automatic methods to index, organize, and retrieve spoken documents will become more important as the amount of spoken language data continues to grow. In addition, these methods will have a significant impact on the use of speech as a data type because speech is currently a difficult medium to browse and search efficiently.

The overall goal of this work is to investigate the feasibility of using subword unit representations for spoken document retrieval as an alternative to words generated by either keyword spotting or continuous speech recognition. The investigation is motivated by the observation that word-based retrieval approaches face the problem of either having to know the keywords to search for *a priori*, or requiring a very large recognition vocabulary in order to cover the contents of growing and diverse message collections. The use of subword units in the recognizer constrains the size of the vocabulary needed to cover the language; and the use of subword units as indexing terms allows for the detection of new user-specified query terms during retrieval. Subword-based approaches can also be used to complement word-based methods in situations where it is difficult to train a large vocabulary recognizer or when out-of-vocabulary words occur in the user query.

[†] This research was supported by DARPA under contract N66001-96-C-8526, monitored through Naval Command, Control and Ocean Surveillance Center.

There are three research issues that need to be addressed. First, what are suitable subword units and how well can they perform? Second, how can these units be extracted from the speech signal in a reliable and efficient manner? And third, how can the indexing and retrieval methods be modified to take into account the fact that the speech recognition output will be errorful? The first issue is addressed in our Eurospeech'97 paper [5] where we explore a range of subword units of varying complexity derived from phonetic transcriptions. We find that subword units are able to capture enough information to perform effective retrieval. With the appropriate subword units it is possible to achieve performance comparable to that of text-based word units if the underlying phonetic units are recognized correctly. In our ICASSP'98 paper [6], we explore the second issue by developing a phonetic speech recognizer, running it on the spoken documents, processing the recognition output to create subword units for indexing and retrieval, and then examining the effects of recognition errors on retrieval performance. We find that in the presence of phonetic recognition errors, retrieval performance degrades but many subword units are still able to achieve reasonable performance even without the use of any error compensation techniques.

In this paper, we focus on the third issue by investigating robust indexing and retrieval methods in an effort to improve retrieval performance when there are speech recognition errors. Although there has been some work in trying to compensate for optical character recognition (OCR) errors introduced into scanned text documents [4], the area of robust methods for dealing with speech recognition errors in the context of spoken document retrieval is still largely unexplored. We examine a number of methods that take into account the characteristics of the recognition errors and try to compensate for them. In the first approach, the original query representation is modified to include similar or confusible terms that could match erroneously recognized speech; these terms are determined using information from the phonetic recognizer's error confusion matrix. The second approach is a generalization of the first method and involves developing a new document-query retrieval measure using approximate term matching designed to be less sensitive to speech recognition errors. In the third method, the document representation is expanded to include multiple recognition candidates (e.g., N-best) to increase the chance of capturing the correct hypothesis. The fourth method modifies the original query using automatic relevance feedback [7] to include new terms found in the top ranked documents. The last method involves the "fusion" or combination of information from multiple subword unit representations.

In the following sections, we first briefly describe the speech corpus, the phonetic speech recognizer, the subword unit indexing terms, and the information retrieval model used. We then describe the different robust indexing and retrieval methods and present information retrieval experiments using these methods to examine their behavior and ability to improve retrieval performance in the presence of phonetic recognition errors. Finally we close with some conclusions and possible directions for future work.

Subword Unit	Indexing Terms
word	weather forecast
1phn (n=1)	w eh dh er f ow r k ae s t
2phn (n=2)	w_eh eh_dh dh_er er_f f_ow ow_r r_k k_ae ae_s s_t
3phn (n=3)	w_eh_dh eh_dh_er dh_er_f er_f_ow f_ow_r ow_r_k r_k_ae k_ae_s ae_s_t

Table 1: Examples of nphone subword unit indexing terms.

2. SPEECH DATA CORPUS

The speech data used in this work consists of FM radio broadcasts of the NPR "Morning Edition" news show. The data is recorded off the air, orthographically transcribed, and partitioned into separate news stories. The data is broken up into two sets, one for training and tuning the speech recognizer and another for use as the spoken document collection for the information retrieval experiments [5]. The speech recognition training set consists of 2.5 hours of clean speech from 5 shows and the development set consists of one hour of data from one show. The spoken document collection consists of 12 hours of speech from 16 shows partitioned into 384 separate news stories. Each story averages 2 minutes in duration and typically contains speech from multiple noise conditions. A set of 50 natural language text queries and associated relevance judgments on the message collection are created to support retrieval experiments. The queries are created from the story "headlines" and are relatively short, each averaging 4.5 words. Each query has an average of 6.2 relevant documents.

3. PHONETIC SPEECH RECOGNIZER

As described in our previous work [6], a phonetic speech recognizer based on the probabilistic segment-based SUMMIT speech recognizer [1] is trained and tuned to operate on the radio broadcast news domain. The recognizer uses context independent segment and context dependent boundary (segment transition) acoustic models. Acoustic feature vectors consisting of Mel-frequency cepstral coefficients (MFCCs), difference cepstra, energy, and duration are derived from the speech signal and used in the acoustic models. The distribution of the acoustic features are modeled using mixtures of diagonal Gaussians. A two pass search strategy is used during recognition. A forward Viterbi search is performed using a statistical bigram language model followed by a backwards A^* search using a higher order statistical n-gram language model. The phonetic recognizer achieves a phone error rate of 35.0% on the development set and 36.5% on a portion (3 hours) of the spoken document collection [6].

4. SUBWORD UNIT REPRESENTATIONS

A range of subword unit indexing terms of varying complexity derived from the phonetic recognition output was explored in our previous work [6]. In this paper, we use one of the better performing sets of subword units: overlapping, fixed-length, phone sequences ranging from n=2 to n=6 in length with a phone inventory of 41 classes. These subword units are derived by successively concatenating the appropriate number of phones from the phonetic transcriptions. Examples of n=1,2 and 3 *nphone* subword units for the phrase "weather forecast" are shown in Table 1.

5. INFORMATION RETRIEVAL MODEL

A standard vector space information retrieval (IR) model [7] is used where the documents and queries are represented as vectors and each component in the vector is an indexing term. A term can be a word, word fragment, or in our case a subword unit.

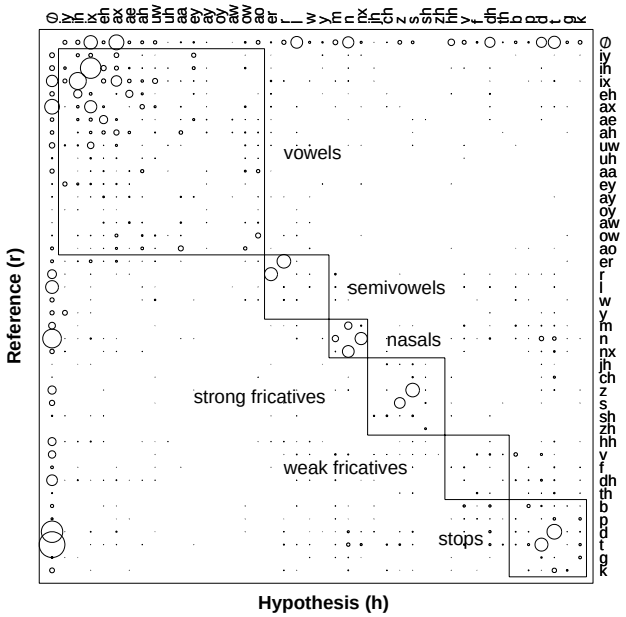

Figure 1: Phonetic recognition error confusion matrix C.

Each term has an associated weight based on the term's occurrence statistics both within and across documents. The weight of term i in the vector for document d is:

$$d[i] = 1 + \log(f_d[i])$$

and the weight of term i in the vector for query q is:

$$q[i] = [1 + \log(f_q[i])] \ \log(N_D/N_{D_i})$$

where $f_d[i]$ is the frequency of term i in document d, $f_q[i]$ is the frequency of term i in query q, N_{D_i} is the number of documents containing term i, and N_D is the total number of documents in the collection. The second term is the inverse document frequency (idf) for term i. A normalized inner product between the document and query is used to score each document during retrieval:

$$S_e(\mathbf{q}, \mathbf{d}) = (\mathbf{q} \cdot \mathbf{d}) \ / \ (||\mathbf{q}|| \ ||\mathbf{d}||) \quad (1)$$

6. ROBUST INDEXING AND RETRIEVAL METHODS

6.1. Expanding the Query Representation

Phonetic recognition errors in the spoken messages result in corrupted indexing terms in the document representation. One way to address this is to modify the query representation to include errorful variants of the original terms to improve the chance of matching the corrupted document terms [4]. These "approximate match" terms are determined using information from the phonetic recognition error confusion matrix (Figure 1) obtained by running the recognizer on the development data set. Each confusion matrix entry, $C(r, h)$, corresponds to a recognition error confusing reference phone r with hypothesis phone h. The bubble radius shown is linearly proportional to the error. The first row ($r = \emptyset$) and column ($h = \emptyset$) correspond to insertion and deletion errors, respectively. We note that many of the confusion errors occur within broad phonetic classes and that many of the insertion and deletion errors happen with short phones.

By thresholding the error, we obtain a set of phone confusion pairs which can then be used to generate approximate match terms via substitution into each original query term. The frequency of a new term j is weighted by its similarity to the original term i:

$$f_q[j] = \frac{\sum_{m=1}^{n} C(i[m], j[m])}{\sum_{m=1}^{n} C(i[m], i[m])}\ f_q[i]$$

where $i[m]$ is the m^{th} phone in subword unit term i with length n; exact term matches have unity weight. In this approach we are using the confusion matrix C as a *similarity* matrix with the error values as indicators of phone similarity.

6.2. Approximate Match Retrieval Measure

Instead of explicitly adding query terms to improve the chance of matching corrupted document terms, we can implicitly consider *all* possible matches between the "clean" query terms and the "noisy" document terms by generalizing the document-query retrieval metric to make use of approximate term matching:

$$S_a(\mathbf{q}, \mathbf{d}) = \sum_{i \in \mathbf{q}} \sum_{j \in \mathbf{d}}\ s(i,j)\ \frac{q[i]}{||\mathbf{q}||}\ \frac{d[j]}{||\mathbf{d}||} \qquad (2)$$

where $s(i,j)$ is the similarity measure between query term i and document term j. We observe that the new metric (2) reduces to the original metric (1) with the appropriate similarity measure:

$$s(i,j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

If we consider j as the "noisy" output term generated by the phonetic recognizer when given the "clean" input term i, then we can view the similarity measure $s(i,j)$ as the conditional probability of observing hypothesis term j given reference term i

$$s(i,j) = p(j \mid i)$$

with the probabilistic model capturing the characteristics of the recognizer. If we assume the phones comprising each subword unit term are independent, then we can estimate this conditional probability using a dynamic programming (DP) procedure:

$$p(j \mid i) = A(l_i, l_j)$$

where l_i and l_j are the lengths of terms i and j, respectively, and A is the $l_i \times l_j$ DP matrix which can be computed recursively:

$$A(m,n) = \begin{cases} 1, & m{=}0,\ n{=}0 \\ A(0, n{-}1) \cdot \tilde{C}(\emptyset, j[n{-}1]), & m{=}0,\ n{>}0 \\ A(m{-}1, 0) \cdot \tilde{C}(i[m{-}1], \emptyset), & m{>}0,\ n{=}0 \\ \max \begin{cases} A(m{-}1, n) \cdot \tilde{C}(i[m{-}1], \emptyset) \\ A(m{-}1, n{-}1) \cdot \tilde{C}(i[m{-}1], j[n{-}1]) \\ A(m, n{-}1) \cdot \tilde{C}(\emptyset, j[n{-}1]) \end{cases}, & m{>}0,\ n{>}0 \end{cases}$$

where $\tilde{C}(r,h)$ is the probability of observing phone h given phone r and is obtained by normalizing the error confusion matrix:

$$\tilde{C}(r,h) = C(r,h) / \sum_{k \in \{h\}} C(r,k)$$

Thresholds can be placed on $p(j \mid i)$ to limit the number of approximate term matches that have to be considered when computing the retrieval score in (2). We note that other probabilistic models such as hidden Markov Models (HMMs) can also be used to estimate this conditional probability.

6.3. Expanding the Document Representation

A different approach is to modify the speech document representation by including high scoring recognition alternatives to increase the chance of capturing the correct hypothesis. This can be done by using the N-best recognition hypotheses, instead of just the single best one. If a term appears in many of the top N hypotheses, it is more likely to have actually occurred than if it appears in only a few. As a result, a simple estimate of the frequency of term i in document d, $f_d[i]$, can be obtained by considering the number of times, n_i, it appears in the top N hypotheses: $f_d[i] = n_i/N$. We note that other information from the recognizer, such as likelihood and confidence scores, can also be used to weight our belief in the accuracy of different hypotheses.

6.4. Query Modification via Automatic Relevance Feedback

The goal in relevance feedback is to iteratively refine a query by modifying it based on the results from a prior retrieval run. A commonly used query reformulation strategy is to add terms found in the retrieved relevant documents and to remove terms found in the nonrelevant documents [7]:

$$\mathbf{q}' = \alpha\, \mathbf{q} + \beta\ \left(\tfrac{1}{N_r} \sum_{i \in D_r} \mathbf{d}_i\right) - \gamma\ \left(\tfrac{1}{N_n} \sum_{i \in D_n} \mathbf{d}_i\right)$$

where D_r is the set of N_r relevant documents, D_n is the set of N_n nonrelevant documents, and α, β, and γ are tunable weight parameters. A threshold can also be placed on the number of new terms, N_t, that are added to the original query. Since there is no human user in the loop to label the initially retrieved documents as relevant and nonrelevant, an automatic variation of the above strategy can be implemented by simply taking the top N_r retrieved documents as relevant and the bottom N_n documents as nonrelevant. Modifying the query in this way not only adds new terms, but can potentially add approximate match terms that occur in the top ranked documents as well.

6.5. Fusion of Multiple Subword Representations

Different subword unit representations can capture different types of information. For example, longer subword units can capture word or phrase information while shorter units can only model word fragments. The tradeoff is that the shorter units are more robust to errors and word variants than the longer units. One simple way to combine the different information is to form a new document-query retrieval score by linearly combining the individual retrieval scores obtained with the separate subword units:

$$S_f(\mathbf{q}, \mathbf{d}) = \sum_n\ w_n\ S_e^n(\mathbf{q}, \mathbf{d})$$

where $S_e^n(\mathbf{q}, \mathbf{d})$ is the document-query score (1) obtained using subword representation n and w_n is a tunable weight parameter. An alternate "fusion" method is to create and use a heterogeneous set of indexing terms by pooling the different subword units.

7. EXPERIMENTS AND RESULTS

In this section, we first examine the behavior and performance of the robust indexing and retrieval methods individually and then explore the effects of combining the different methods.

Figure 2A shows retrieval performance, measured in average precision, for the different phonetic subword units (n=2,3,4,5,6) using the query expansion method described in Section 6.1 as the threshold is lowered to include more approximate match terms. At a threshold value of 100, the query is the original one with no added terms. We first note that subword units of intermediate length (n=3,4) perform better than short (n=2) or long (n=5,6) units; this is due to a better tradeoff between being too short and matching too many terms and being too long and not matching enough terms [5]. As the threshold is lowered and more terms are added, performance of the short subword unit (n=2) becomes worse; this is due to spurious matches from the additional terms. However, the longer subword units (n=4,5,6) are much improved with expanded queries; more terms are being matched while the

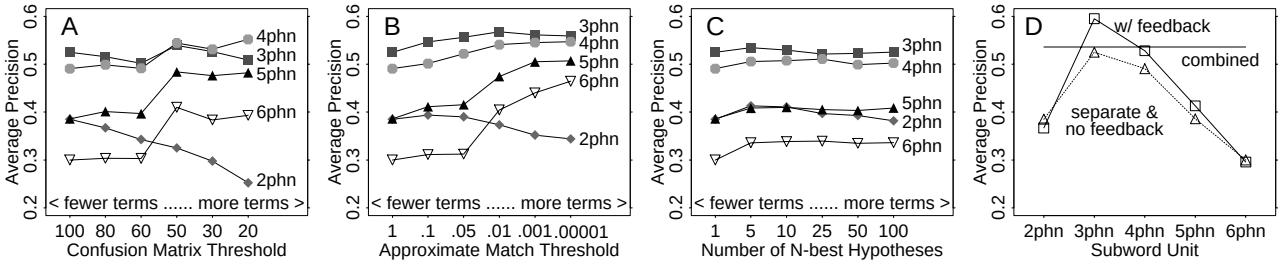

Figure 2: (A) Query expansion (B) Approximate match retrieval (C) Document expansion (D) Automatic feedback & Subword fusion.

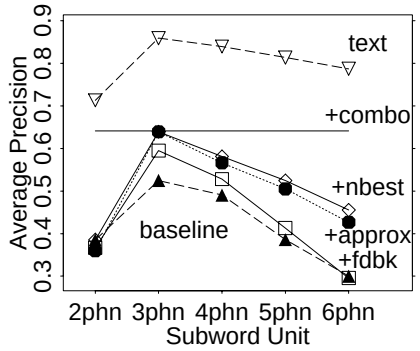

Figure 3: Successive combination of the robust methods.

longer sequence length makes it more difficult to get spurious matches. Performance for n=3 stays about the same.

Figure 2B shows retrieval performance for the different phonetic subword units using the new document-query retrieval metric in (2) as the threshold on $p(j \mid i)$ is lowered to consider more approximate matches. The performance behavior is very similar to that observed in Figure 2A with improvements for the longer subword units and losses for the short ones as the threshold is lowered. The main differences are that the overall performance gains are better and that performance of the n=3 subword unit is improved. Overall, implementing approximate match using the new document-query metric is superior to adding terms to the query.

Retrieval performance for the different subword units as the document representation is expanded to include the N-best recognition hypotheses is shown in Figure 2C. Performance improves slightly for all the subword units as N increases but then levels off after N=5 or 10. It appears that most of the useful recognition variants occur within the first few hypotheses.

Retrieval performance with ($\square$) and without ($\triangle$) the use of automatic relevance feedback is shown in Figure 2D for the different subword units. The following relevance feedback parameters are used: N_r=1, N_n=10, α=β=γ=1, and N_t=50. Performance is significantly improved for subword units of length n=3,4,5 but remains about the same for units of length n=2,6. This illustrates again the tradeoff advantages of intermediate length units [5].

Using the method described in Section 6.5 to linearly combine the separate retrieval scores with equal weights (w_n=0.2, n =2,3,4,5,6) results in performance that is slightly worse than just using the best performing subword unit (n=3). Changing the weights to favor the better performing units (w_3=0.5, w_4=0.2, $w_{2,5,6}$=0.1) is only marginally better than the n=3 subword unit as shown by the solid line in Figure 2D. Maybe the use of more sophisticated non-linear combination methods will be better.

Starting with the baseline retrieval performance of the different subword units, we successively combine the various robust methods to see how performance improves. As shown in Figure 3, adding automatic relevance feedback (+fdbk) improves performance for the n=3,4,5 subword units; using the approximate match retrieval metric (+approx) further improves performance for all subword units except for n=2; expanding the documents using the top N=10 recognition hypotheses (+nbest) improves performance for the longer subword units; finally combining the scores of the different subword units (+combo) gives performance similar to that of the best performing subword unit (n=3). The final result is that information retrieval performance, measured in average precision, improves from p=0.52 (for the initial n=3 subword unit) to p=0.64, a gain of about 23%. There remains, however, a large performance gap when compared to subword units derived from error-free phonetic transcriptions (text).

8. CONCLUSIONS AND FUTURE WORK

In this paper, we investigate a number of robust methods in an effort to improve spoken document retrieval performance when there are speech recognition errors. We study the different methods individually and then explore the effects of combining them. We find that using a new approximate match retrieval metric, modifying the queries via automatic relevance feedback, and expanding the documents with N-best recognition hypotheses improves performance; subword unit fusion, however, resulted in only marginal gains. Combining the approaches results in additive performance improvements. Future work in this area include investigating more sophisticated probabilistic models for approximate matching; exploring non-linear methods for combining different subword units; and examining the use of recognizer likelihood and confidence scores in the indexing and retrieval process.

9. ACKNOWLEDGMENTS

I would like to thank Dr. Victor Zue for supervising this research and for reading and providing comments on this paper.

10. REFERENCES

[1] J. Glass, *et. al.*, "A Probabilistic Framework for Feature-Based Speech Recognition," *ICSLP 1996*, pp. 2277-2280.

[2] D. K. Harman, ed., *Fifth Text REtrieval Conference (TREC-5)* Gaithersburg, MD, USA, NIST, 1996. NIST-SP 500-238.

[3] A.G. Hauptmann and H.D. Wactlar "Indexing and Search of Multimodal Information," *ICASSP 1997*, pp. 195-198.

[4] K. Marukawa, *et. al.*, "Document retrieval tolerating character recognition errors – evaluation and application," *Pattern Recognition*, vol. 30, no. 8, pp. 1361–1371, 1997.

[5] K. Ng and V. Zue, "Subword Unit Representations for Spoken Document Retrieval," *Eurospeech 1997*, pp. 1607-1610.

[6] K. Ng and V. Zue, "Phonetic Recognition for Spoken Document Retrieval," *ICASSP 1998*, pp. 325-328.

[7] G. Salton and M. McGill, "Introduction to Modern Information Retrieval," McGraw-Hill, NY, 1983.

[8] S.J. Young, *et. al.*, "Acoustic indexing for multimedia retrieval and browsing," *ICASSP 1997*, pp. 199-202.