Improved Spoken Document Retrieval by Exploring Extra Acoustic and Linguistic Cues

Berlin Chen[1,2], Hsin-min Wang[1], and Lin-shan Lee[1,2]

[1] Institute of Information Science, Academia Sinica,
[2] Dept. of Computer Science & Information Engineering, National Taiwan University,
Taipei, Taiwan, Republic of China
E-mail: {berlin, whm, lsl}@iis.sinica.edu.tw

ABSTRACT

In this paper, we explored the use of various extra information to improve the performance of spoken document retrieval (SDR). From the speech recognition perspective, we incorporated the acoustic stress and word confusion information into the audio indexing. From the linguistic perspective, we applied the part-of-speech information in both the audio indexing and the query representation. From the information retrieval perspective, we integrated techniques such as the query expansion by word associations and the blind relevance feedback into the retrieval process. The SDR experiments were based on the Topic Detection and Tracking Corpora (TDT-2 and TDT-3). We used the Chinese newswire text stories as query exemplars and the Mandarin Chinese audio news stories as the spoken documents. With all the above acoustic and linguistic cues applied, the average precision was improved from 0.5122 to 0.6312 for the TDT-2 collection and from 0.6216 to 0.7172 for the TDT-3 collection.

1. INTRODUCTION

With the rapidly growing use of the audio and multi-media information on the Internet, an exponentially increasing number of spoken documents, such as broadcast radio and television programs, are now being accumulated and made available. Development of technologies for spoken document retrieval (SDR) is thus becoming more and more important, and has been extensively studied in recent years [1-3]. However, most of such studies simply used an automatic speech recognizer to transcribe the spoken documents into either word or subword sequences or graphs and then applied the conventional text information retrieval techniques, or used some similar approaches. Unlike the texts, the spoken documents also carry plenty of extra linguistic, acoustic and prosodic information, such as acoustic stresses, speech acts, part-of-speech (POS), and so on, which might be helpful for speech information retrieval, but have not been extensively explored yet. Some researchers have used the prosodic information for speech corpus analyses [4-5], but the use of such information in spoken document retrieval has not been investigated. Moreover, speech recognition inevitably introduces errors which naturally degrade the performance of a speech information retrieval system. Quite a few research works have been focused on the development of robust retrieval approaches against the recognition errors, including using multiple recognition hypotheses to provide more information in the audio indexing [6-8], or using the phone-level acoustic confusion information statistically collected form the training speech corpus to expand either the query or the document representations for robust matching of indexing terms [9]. Based on these observations, in this paper, we explored the use of various extra information for spoken document retrieval from different perspectives. From the speech recognition perspective, we incorporated the acoustic stress and word confusion information into the audio indexing. From the linguistic perspective, we applied the part-of-speech information in both the audio indexing and the query representation. From the information retrieval perspective, we integrated techniques such as the query expansion by word associations and the blind relevance feedback into the retrieval process. All of the schemes investigated in this paper are shown to be helpful to the spoken document retrieval task. With all the schemes applied together, the retrieval performance was significantly improved.

In this paper, all the experiments were tested on the task involving the use of an entire Chinese newswire story (text) as a query, to retrieve relevant Mandarin Chinese radio broadcast news stories (audio) in the document collection. Such a retrieval context is termed *query-by-example*.

2. EXPERIMENT SETUP

2.1 Experimental Corpora

We used two Topic Detection and Tracking (TDT) collections for this work. TDT-2 is taken as the development test set while TDT-3 is taken as the evaluation test set. The Chinese news stories (text) from Xinhua News Agency were used as queries (or query exemplars). The Mandarin news stories (audio) from Voice of America news broadcasts were used as the spoken documents. All news stories were exhaustively tagged with event-based topic labels, which served as the relevance judgments for performance evaluation. Table 1 describes the details for the corpora used in this paper.

The Dragon large-vocabulary continuous speech recognizer [10] provided Chinese word transcriptions for our Mandarin audio collections (TDT-2 and TDT-3), such that the results here may be compared with works done by other groups. We tried to assess the performance of Dragon's recognizer on the TDT corpora by comparing Dragon's recognition hypotheses with the manual transcriptions. Notice that Dragon's recognition output contains word boundaries (tokenizations) resulting from its language models and vocabulary definition while the manual transcripts are running texts without word boundaries. Since Dragon's lexicon is not available, we augmented the LDC Mandarin Chinese Lexicon with the 24k words extracted from Dragon's word recognition output, and used the augmented LDC lexicon in tokenizing the manual transcriptions for computing word error rates. We also used the augmented LDC lexicon in tokenizing the text query exemplars in the retrieval experiments. We have spot-checked a fraction of the TDT-2 development set (39.90 hours) and obtained a word error rate of 35.38%. Spot-checking approximately 76 hours of the TDT-3 test set gave a word error rate of 36.97%.

	TDT-2 (Development) 1998, 02~06			TDT-3 (Evaluation) 1998, 10~12		
# Spoken documents	2,265 stories, 46.03 hours of audio			3,371stories, 98.43 hours of audio		
# Distinct text queries	16 Xinhua text stories (Topics 20001~20096)			47 Xinhua text stories (Topics 30001~30060)		
	Min.	**Max.**	**Mean**	**Min.**	**Max.**	**Mean**
Doc. length (characters)	23	4841	287.1	19	3667	415.1
Query length (characters)	183	2623	532.9	98	1477	443.6
# Relevant documents per query	2	95	29.3	3	89	20.1

Table 1: Statistics of TDT-2 and TDT-3 collections used in this paper.

2.2 Information Retrieval Model

We used the word-level indexing features (indexing terms composed of single words only) and the vector space model in the SDR experiments here for initial studies, although it has been found that syllable-level indexing features [7-8] and HMM/N-gram-based models [11] may provide very good results for the problem. For a spoken document, each component in the vector representation $\vec{d}$ is given by the weighted statistics $s(t)$ of a specific word t:

$$s(t) = \left(1 + \ln\left(\sum_{i=1}^{n_t} w_t(i)\right)\right) \cdot \ln\left(N / N_t\right), \quad (1)$$

where $w_t(i)$ is the term weight of the *i*-th occurrence of the word t within the document. Below we'll try to incorporate different useful information into $w_t(i)$, though $w_t(i)$ was simply set to 1 in the baseline retrieval system. The value of $1+\ln\left(\sum_{i=1}^{n_t} w_t(i)\right)$ denotes the term frequency for the word t in the document, where the logarithmic operation is to condense the distribution of the term frequency. $\ln(N/N_t)$ is the Inverse Document Frequency (IDF), where N_t is the number of documents that include the word t and N is the total number of documents in the collection. A query is also represented as a vector $\vec{q}$ in exactly the same way. The Cosine measure is used to estimate the query-document relevance:

$$R(\vec{q},\vec{d}) = \left(\vec{q} \bullet \vec{d}\right) \Big/ \left(\left\|\vec{q}\right\| \cdot \left\|\vec{d}\right\|\right). \quad (2)$$

2.3 Baseline Experimental Results

In this paper, the manual transcriptions of the spoken documents (denoted as TD below) were also used in the retrieval experiments for reference, as compared to the erroneous transcriptions obtained from speech recognition (denoted as SD below). The retrieval results were expressed in terms of *non-interpolated average precision* [12]. As shown in Table 2, the baseline retrieval results for the SD and TD cases are respectively 0.5122 and 0.5548 for the TDT-2 collection, and 0.6216 and 0.6505 for the TDT-3 collection.

	TDT-2	TDT-3
SD	0.5122	0.6216
TD	0.5548	0.6505

Table 2: The baseline retrieval performance.

3. IMPROVEMENTS FROM THE SPEECH RECOGNITION PERSPECTIVE

3.1 The Acoustic Stress Information (AS)

In this section, we tried to explore the use of the prosodic information in spoken document retrieval by incorporating the acoustic stresses of spoken words into the audio indexing. We assumed that words carrying important linguistic clues in the spoken documents might be pronounced more slowly, more strongly, and more clearly than the other words. Therefore, the signal magnitudes, durations, and confidence measures (or speech recognition scores) of words can be applied to the audio indexing. As described in the Section 2.1, the spoken documents were furnished with recognized words from the Dragon system. These word hypotheses were accompanied with information such as the begin times, durations, and confidences. Thus, the word durations and word confidences are available. Nevertheless, we need to estimate the word magnitude information (signal magnitude of the word) from the speech samples. In this research, we tried to calculate the signal magnitude of every spoken word t using the following equation:

$$MAG'(t) = \frac{1}{E_t - B_t + 1} \sum_{i=B_t}^{E_t} \left(\frac{1}{M} \sum_{j=i-M+1}^{i} |s(i)| \right), \quad (3)$$

where B_t and E_t are the begin and the end times of the word t, $|s(\cdot)|$ is the magnitude of the corresponding speech samples, and M is the window length for calculating the average magnitude of the speech signal at a specific time index. We then transformed this magnitude of the word t to a value between 0 and 1 using a Sigmoid function:

$$MAG(t) = \frac{2}{1 + \exp(-\alpha (MAG'(t) - MAG'(t^*)))}, \quad (4)$$

where t^* is the word with the maximal magnitude in a spoken document, α is used to control the slope of the Sigmoid function. In this study, α was empirically determined. Similarly, we also respectively obtained the duration $DUR(t)$ and the confidence $CON(t)$ of the word t normalized to values between 0 and 1 via the same Sigmoid function, but with different α values. As a result, the combination of these three kinds of information could be expressed as a geometric average:

$$AS(t) = \sqrt[\gamma]{MAG(t)^{k_1} DUR(t)^{k_2} CON(t)^{k_3}}. \quad (5)$$

In this research, k_1, k_2 and k_3 were all set to 1 in the initial tests, and γ was set to 3. In addition, we have also investigated the combinations of any two of the three information sources by calculating the geometric average of the two values (e.g. $AS(t) = \sqrt{MAG(t) DUR(t)}$).

3.2 The Word-level Confusion Information (WC)

In speech recognition, a specific word may be often mis-recognized to be some other specific words according to the acoustic characteristics of the involved words, the speech data or the behavior of the speech recognizer. Such word confusion information may be helpful for the spoken document retrieval task if it is derived from a training speech corpus with similar acoustic characteristics as the spoken document collection to be retrieved using the same speech recognizer. The use of the phone-level confusion information has been shown to be helpful in improving the performance of a phone-based (kind of subword-based) spoken document retrieval system recently [9]. However, not too much work on applying the word-level confusion information to the word-based spoken document retrieval task has been reported yet. As a result, we used the TDT-2 spoken documents (the development set) as the training

	+MAG	+DUR	+CON	+ MAG+DUR	+ MAG+CON	+DUR+CON	+AS (MAG+DUR+CON)
TDT-2 (SD)	0.5150 (+0.54%)	0.5176 (+1.05%)	**0.5270** (+2.89%)	0.5166 (+0.85%)	0.5226 (+2.03%)	0.5237 (+2.24%)	**0.5258** (+2.66%)
TDT-3 (SD)	0.6263 (+0.77%)	0.6262 (+0.74%)	0.6248 (+0.51%)	0.6252 (+0.58%)	0.6300 (+1.35%)	0.6298 (+1.32%)	**0.6309** (+1.50%)

Table 3: The retrieval performance after applying the acoustic stress information in the audio indexing.

	+WC	+AS+WC
TDT-2 (SD)	0.5217 (+1.85%)	0.5299 (+3.46%)
TDT-3 (SD)	0.6355 (+2.24%)	0.6392 (+2.83%)

Table 4: The retrieval performance after applying the word confusion information and the acoustic stress information.

speech corpus to get the word-level confusion information. We aligned Dragon's word recognition outputs with the manual transcriptions using dynamic programming, and recorded the corresponding word hypothesis for every reference word. Consequently, a global word confusion matrix was built, which recorded for each word w_i the probability that it would be recognized as a word hypothesis w_j. In the retrieval process, we used this confusion information to expand the query representation by redistributing the weights of the original query terms over their corresponding top N confusing words.

3.3 Experimental Results

The retrieval results obtained by using either one or two or all of the three acoustic stress information (MAG, DUR, CON as presented in Equations (4)(5)) are shown in Table 3, in which the numbers in the parentheses are the relative improvements compared to the baseline SD cases of the TDT-2 and TDT-3 collections. It can be found that the use of any one of the acoustic stress information alone seemed to offer some slight improvements compared to the baseline retrieval performance, not very significantly though, while combining all of them in general outperformed using either one or either combination of two of the three.

The retrieval results obtained after applying the word-level confusion information alone are shown in the first column of Table 4. It can be found that the word confusion information trained from the TDT-2 collection in fact worked reasonable well on the TDT-3 collection. Furthermore, the word confusion information explored here seemed to be more effective than the acoustic stress information explored in Table 3. The right column of Table 4 shows the results achieved by using both the acoustic stress information and the word confusion information. We found that using both the acoustic stress information and the word confusion information outperformed using either alone, therefore the acoustic stress information and the word confusion information are additive. This is intuitively reasonable.

4. IMPROVEMENTS FROM THE LINGUISTIC PERSPECTIVE

4.1 Part-of-Speech (POS) Information

In information retrieval, the proper nouns, such as the locations, organizations, personal names, etc., usually play more important roles than the adjectives, adverbs and verbs. Thus, they should be emphasized with higher weights. Consequently, the following POS weights for indexing terms were applied in our system:

Proper Nouns	1.2
Adjectives and Adverbs	0.8
Verbs	0.9
The rest of words	1.0

The part-of-speech tags of words were referred to our Mandarin Chinese lexicon. These term weights were applied to both the audio indexing and the query representation.

4.2 Experimental Results

The retrieval results after applying the POS information are shown in the first column of Table 5. Compared to the baseline retrieval performance, we found that the POS information is rather effective for both the SD and TD cases for the TDT-2 and TDT-3 collections.

5. IMPROVEMENTS FROM THE INFORMATION RETRIEVAL PERSPECTIVE

Like conventional text information retrieval, spoken document retrieval has to deal with the indexing term mismatch problem too. That is, the system may fail to retrieve the desired relevant documents given the specified query terms, simply because the indexing terms are really not matched. Therefore, in this section two prevailing information retrieval techniques were revisited to tackle the indexing term mismatch problem. They are the query expansion by word associations and the blind relevance feedback.

5.1 Word Associations (WA)

Words that co-occur frequently in the same spoken document can be assumed to have some degree of synonymity association [13]. Thus, we can build a global word association matrix, in which each entry $WA(i,j)$ stands for the correlation factor between words t_i and t_j, and is expressed as:

$$WA(i,j) = \frac{\hat{c}_{i,j}}{c_i + c_j - \hat{c}_{i,j}}, \quad (6)$$

where c_i and c_j are respectively the total number of documents in the document collection which include the words t_i and t_j respectively, and $\hat{c}_{i,j}$ is the total number of documents in the document collection which include both the words t_i and t_j within the same document. For example, $WA(i,j)=1$ if words t_i and t_j always appear in the same document, and $WA(i,j)=0$ if words t_i and t_j never appear in the same document. The query feature vector is then reformulated by including in the new query expression a limited number of extra indexing terms, which have the highest synonymity association to the indexing terms existing in the original query expression. In this research, two global word association matrices were automatically built for the SD and TD cases, respectively, based on Dragon's word recognition outputs and the manual transcriptions of the TDT-2 spoken document collection.

5.2 Blind Relevance Feedback (BREF)

It has been found that some indexing terms not appearing in the query may still act as useful cues for relevance judgments. For example, the information from the relevant or irrelevant

		+POS	+WA	+BREF	+ POS+WA+BREF	+AS+WC+POS+WA+BREF
TDT-2	**SD**	0.5414 (+5.70%)	0.5362 (+4.69%)	0.5650 (+10.31%)	0.6211 (+21.26%)	**0.6312** (+23.23%)
	TD	0.5870 (+5.80%)	0.5760 (+3.82%)	0.6331 (+14.11%)	**0.6540** (+17.80%)	-
TDT-3	**SD**	0.6374 (+2.54%)	0.6271 (+0.88%)	0.6723 (+8.16%)	0.7016 (+12.87%)	**0.7172** (+15.38%)
	TD	0.6659 (+2.37%)	0.6570 (+1.00%)	0.6793 (+4.43%)	**0.7086** (+8.93%)	-

Table 5: The retrieval performance after applying the acoustic and linguistic information and the advanced retrieval techniques.

documents selected or deleted in the first stage retrieval can be further used to identify the indexing terms relevant to the user's intention. In this research, a blind relevance feedback procedure was used to reformulate the initial query expression automatically based on the modified Rocchio's formula [13]:

$$\vec{q}' = \alpha \cdot \vec{q} + \beta \cdot \sum_{\vec{d}_i \in D_r} \vec{d}_i - \gamma \sum_{\vec{d}_j \in D_{irr}} \vec{d}_j, \quad (7)$$

where $\vec{q}$ and $\vec{q}'$ are the initial and modified query feature vectors, $\vec{d}_j$ is the vector representation of the documents, D_r and D_{irr} are the sets of relevant and irrelevant documents, and α, β and γ are empirically adjustable weighting parameters.

5.3 Experimental Results

Columns 2 and 3 of Table 5 respectively show the retrieval results when the query expansion by word associations (WA) and the blind relevance feedback (BREF) were introduced into the retrieval process. It can be found from Column 3 that the blind reference feedback (BREF) is equally effective for both the TDT-2 and TDT-3 collections. However, the query expansion by word associations (WA) in Column 2 seems to be more effective for the TDT-2 collection than for the TDT-3 collection. One possible reason is that the topics of the TDT-2 and TDT-3 are not very related to each other, therefore the word association information derived form the TDT-2 collection is not very helpful to TDT-3. The rest parts of Table 5 are the results when the different approaches mentioned above in this paper were all applied together. For the SD cases, the average precision was finally improved from 0.5122 to 0.6312 (23.23% relative improvement) for the TDT-2 collection while from 0.6216 to 0.7172 (15.38% relative improvement) for the TDT-3 collection. Figure 1 depicts the comparison of retrieval performance for the TDT-3 evaluation set by using the baseline configuration (the solid curve) and the best configuration (all approaches, last column of Table 5, the dotted curve) in the SD case for all the 47 queries respectively. Some observations can be drawn. For the queries with very poor retrieval performance in the baseline system, the use of the extra information developed here was still not very helpful in general. For the queries with reasonable retrieval performance in the baseline system, the approaches developed here were quite effective in many cases. For the queries already with very good retrieval performance in the baseline system, the best configuration developed here worked either just as well or even better.

6. CONCLUSIONS

In this paper, we explored various approaches to use extra information for spoken document retrieval (SDR) from different perspectives. From the speech recognition perspective, we incorporated the acoustic stress and word confusion information into the audio indexing. From the linguistic perspective, we applied the part-of-speech information in both the audio indexing and the query representation. From the information retrieval perspective, we integrated techniques such as the query expansion by word associations and the blind relevance feedback into the retrieval process. All these schemes investigated here are more or less helpful to the spoken document retrieval task. With all these schemes applied together, the retrieval performance could be improved significantly.

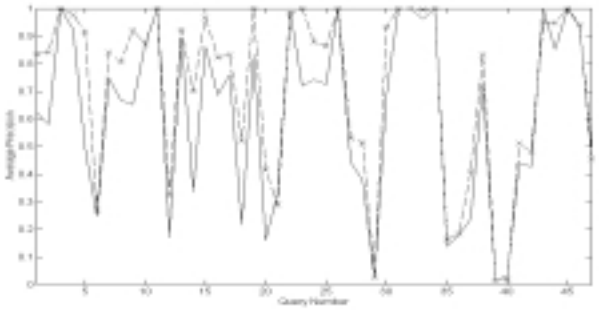

Figure 1: The comparison of retrieval performance for the TDT-3 evaluation set by using the baseline configuration and the best configuration (the dotted curve) in the SD case.

7. REFERENCES

[1] P. Jourlin, S. E. Jonson, K. Spärck Jones, P. C. Woodland, "Spoken Document Representations for Probabilistic Retrieval," *Speech Communication*, 32, pp. 21-36, 2000.

[2] CMU Informedia Digital Video Library project.

[3] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, "Speech and Language Techniques for Audio Indexing and Retrieval," *Proc. IEEE*, Vol. 88, No. 8, Aug. 2000.

[4] R. Silipo and S. Greenberg, "Automatic Transcription of Prosodic Stress for Spontaneous English Discourse," in Proc. ICPhS, 1999.

[5] R. Silipo and F. Crestani, "Prosodic Stress and Topic Detection in Spoken Sentences, " Technical Report, International Computer Science Institute, Berkeley, 2000.

[6] K. Ng, "Information Fusion for Spoken Document Retrieval," in *Proc.* ICASSP 2000.

[7] B. Chen, H. M. Wang, and L. S. Lee, "Retrieval of Broadcast News Speech in Mandarin Chinese Collected in Taiwan Using Syllable-Level Statistical Characteristics," in *Proc.* ICASSP 2000.

[8] H. M. Wang, H. Meng, P. Schone, B. Chen and W. K. Lo, "Multi-Scale Audio Indexing for Translingual Spoken Document Retrieval, " in *Proc.* ICASSP 2001.

[9] S. Srinivasan and D. Petkovic, "Phonetic Confusion Matrix Based Spoken document Retrieval," in *Proc.* ACM SIGIR 2000.

[10] P. Zhan, S. Wegmann, and L. Gillick, "Dragon Systems' 1998 Broadcast News Transcription System for Mandarin," in *Proc. of the DARPA Broadcast News Workshop*, 1999.

[11] B. Chen, H. M. Wang, and L. S. Lee, "An HMM/N-gram-based Linguistic Processing Approach for Mandarin Spoken Document Retrieval," submitted to Eurospeech 2001.

[12] D. Harman, Overview of the Fourth Text Retrieval Conference (TREC-4). 1995.

[13] Baeza-Yates Ricardo and Ribeiro-Neto Berthier. Modern Information Retrieval. 1999.