RETRIEVAL OF MANDARIN BROADCAST NEWS USING SPOKEN QUERIES

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

ABSTRACT

Considering the monosyllabic structure of the Chinese language, a whole class of indexing features for retrieval of Mandarin broadcast news using syllable-level statistical characteristics has been previously investigated. This paper presents the improvements achieved over the previous results. The major differences are: (1) Multi-scale character- and word-level indexing terms have been integrated with the syllable-level information. (2) Information cues from the contemporary newswire text corpus have been used to create more accurate syllable indexing terms. (3) Automatic document expansion, blind relevance feedback, and query expansion via the term association matrix have been applied in retrieval. With all these schemes, the average precision can be improved from 55.46% to 71.29%.

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 technology for retrieving speech information is, thus, becoming more and more important, and has been extensively studied in recent years [1-3]. Using the spoken queries to access the audio streams remains to be a very challenging research topic because the query terms could contain recognition errors and such errors could make the retrieval system completely mistake what the user needs especially when the query is short. There are still not many reports on this very challenging task. However, since wireless communication is becoming more and more popular, development of technology for voice retrieval of speech information is becoming more and more important.

There have been several different approaches developed for spoken document retrieval (SDR) in recent years. Word–based retrieval approaches have been very popular and successful, although with the potential problems of either having to know the query words in advance, or requiring a large enough lexicon to cover the growing dynamic contents of the diverse broadcast news [6]. Some other researchers proposed the concept of subword-based approaches, which can constrain the size of vocabulary needed to cover the language but also result in a tremendous size of indexing terms [8]. Considering the monosyllabic structure of the Chinese language, a whole class of indexing features for retrieval of Mandarin broadcast news using syllable-level statistical characteristics has been previously investigated [9]. This paper presents the improvements achieved over the previous results: First of all, the fusions of different indexing terms were extensively evaluated and the results were compared with that of the syllable-based approach. Then, information cues from the contemporary newswire text corpus were use to create more accurate syllable indexing terms. Finally, several information retrieval techniques were further applied to improve the retrieval performance. With all these schemes, the average precision can be improved from 55.46% (the best result previously reported in [9]) to 71.29%.

2. BROADCAST NEWS DATABASE

The whole broadcast news database used in this paper was collected from December 1998 to July 1999. The training data consists of 453 stories (about 4.0 hours of speech materials), and was collected from Police Radio Station (PRS), UFO Station (UFO), Voice of Han (VOH), and Voice of Taipei (VOT), all located at Taipei. All recordings were manually transcribed and segmented into stories and sentences. In this research, only the speech parts of the anchor speakers were used in the experiments. The testing data to be retrieved consists of 757 recordings (about 10.2 hours of speech materials), and was collected from Broadcasting Corporation of China (BCC). Each recording was a short news abstract (about 50 seconds) produced by an anchor speaker, and contained several news items. Some recordings in the testing data contain background music. The testing data was also manually transcribed but was not segmented into sentences. A set of 40 query terms and the corresponding relevant news recordings were manually created to support the retrieval experiments. Each query contains roughly only 4 characters (or syllables) and has on average 23.3 relevant documents among the 757 in the database, with the exact number ranging from 1 to 75. Two (one male and one female) speakers were asked to pronounce the 40 query terms respectively.

3. SPEECH RECOGNITION

3.1 Acoustic Processing and Language Modeling

Each frame of the speech data is represented by a 39 dimensional feature vector, which consists of 12 MFCCs and log energy, and their first and second differences. Utterance-based cepstral mean subtraction (CMS) is applied to all the training and testing materials. A set of 150 intra-syllable right-context-dependent sub-syllable HMMs plus one silence HMM were used for speech recognition. In addition, the syllable-based and word-based N-gram language models were trained by a newswire text corpus consisting of 80 million Chinese characters collected from Central News Agency (CNA) in 1999. Note that both the newswire text corpus and the broadcast news database were collected almost in the same time frame. Word segmentation and phonetic labeling were performed for the training materials using a 62k-word lexicon.

	Syllable	Character	Word
Spoken Queries	89.20%	84.88%	72.62%
Spoken Documents	73.37%	62.79%	43.37%

Table 1: The speech recognition results after introducing the word-level language model.

3.2 Speech Recognition Results

A two-pass search strategy is used in the syllable recognition. In the first pass, Viterbi search was performed based on the acoustic models and the syllable bigram language model, and the score at every time index were stored. In the second pass, a backward time-asynchronous A* tree search generates the best syllable sequence based on the heuristic scores obtained from the first pass search and the syllable trigram language model. Based on the state likelihood scores calculated in the first pass search and the syllable boundaries of the best syllable sequence, the syllable recognizer further performs the Viterbi search on each utterance segment which may include a syllable and outputs several most possible syllable candidates, and a syllable lattice can thus be constructed. The syllable accuracies achieved for the spoken documents and the speech queries were 64.90% and 81.50% respectively, as reported in [9]. In this paper, we further construct the word graph from the syllable lattice and perform dynamic programming on the word graph to find the best word sequence using the word unigram and bigram language models. Then, we map the word sequence into the character- and syllable-level sequences. The speech recognition results are summarized in Table 1. It can be found that the syllable accuracies for the spoken documents and the speech queries can be improved to 73.37% and 89.20%, respectively while the character accuracies are 62.79% and 84.88%, and the word accuracies are 43.37% and 72.62%.

4. SYLLABLE-BASED MANDARIN SDR

4.1 Syllable-level Indexing Terms

In the Chinese language, each word is composed of from one to several characters and all the characters are monosyllabic. On the other hand, each syllable may stand for many different characters with different meanings and the combination of several specific syllables very often gives only very few, if not unique, homonym polysyllabic words. As a result, comparing the input query and the document based on the segments of several syllables may provide very good degree of similarity between them. Therefore, the syllable-level indexing terms composed of the overlapping syllable segments with length N ($S(N)$, $N=1\sim3$) and the syllable pairs separated by n syllables ($P(n)$, $n=1\sim3$) are used in the following SDR experiments. Considering a syllable string of *10* syllables $S_1\ S_2\ S_3 \ldots\ S_{10}$, examples of the former are listed on the upper half of Table 2, while examples of the latter on the lower half of Table 2. The combination of these indexing terms has been shown to be very effective for Mandarin SDR [9]. For example, the overlapping syllable segments with length N can capture the information of polysyllabic words or phrases while the syllable pairs separated by n syllables can tackle the problems arising from the flexible wording structure, abbreviation, and speech recognition errors.

4.2 Information Retrieval Model

Vector space models widely used in many text information retrieval systems were used here. In this paper, a document is represented as a set of feature vectors, each consists of one type of indexing terms ($S(1)$, $S(2)$, $S(3)$ *and* $P(1\sim3)$), instead of a single feature vector consisting of all indexing terms together as in [9]. Each component of a feature vector $\vec{v}$ is associated with the weighted statistics $ws_{\vec{v}}(t)$ of a specific indexing term t :

$$ws_{\vec{v}}(t) = \left(1 + \ln\left(\sum_{i=1}^{n_t} cm_t(i)\right)\right) \cdot \ln\left(N / N_t\right) \quad (1)$$

where $cm_t(i)$, ranging from *0* to *1*, is the acoustic confidence measure of the *i*-th occurrence of indexing term t within the document and is simply set to *1* when the document is a perfect manual transcription, and variable n_t is the total occurrences of indexing term t in the document. The value of $1+\ln\left(\sum_{i=1}^{n_t} cm_t(i)\right)$ denotes the term frequency of indexing term t and the logarithmic operation is used to condense the distribution of the term frequency. $\ln(N/N_t)$ is the Inverse Document Frequency (**IDF**), where N is the total number of documents in the collection and N_t is the number of documents that contain term t. A query is also represented by a set of feature vectors based on the same feature vector construction procedure. The Cosine measure is used to estimate the query-document similarity for each type, s, of indexing terms:

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

The overall similarity is then the weighting sum of the similarity scores of all types of indexing terms:

$$SIM_S(\vec{q}, \vec{d}) = \sum_S w_s \cdot SIM_s(\vec{q}_s, \vec{d}_s), \quad (3)$$

where w_s is a empirically tunable weighting parameter.

Syllable Segments	Examples
S(N), N=1	$(s_1)\ (s_2)\ \ldots(s_{10})$
S(N), N=2	$(s_1\ s_2)\ (s_2\ s_3)\ldots(s_9\ s_{10})$
S(N), N=3	$(s_1\ s_2\ s_3)\ (s_2\ s_3\ s_4)\ldots(s_8\ s_9\ s_{10})$
Syllable Pair Separated by *n* Syllables	Examples
P(n), n=1	$(s_1\ s_3)\ (s_2\ s_4)\ \ldots(s_8\ s_{10})$
P(n), n=2	$(s_1\ s_4)\ (s_2\ s_5)\ \ldots(s_7\ s_{10})$
P(n), n=3	$(s_1\ s_5)\ (s_2\ s_6)\ \ldots(s_6\ s_{10})$

Table 2: The indexing terms extracted from an example syllable string $S_1\ S_2\ S_3 \ldots\ S_{10}$.

4.3 Experimental Results

The IR experiments using the perfect manual transcriptions of the queries and documents were evaluated for reference. The perfect query and document manual transcriptions are denoted as **TQ** (**T**ext **Q**ueries) and **TD** (**T**ext **D**ocuments) while the erroneous transcriptions from speech recognition are denoted as **SQ** (**S**poken **Q**ueries) and **SD** (**S**poken **D**ocuments) correspondingly. The baseline syllable-based SDR results are shown in the second column of Table 3 in terms of *non-interpolated average precision* (**nAP**). Each entry denotes the result obtained by combining all indexing term types *S(N), N=1~3,* and *P(n), n=1~3*, while the results obtained by merely using *S(N), N=1*, or *S(N), N=1~2*, are also presented in the parenthesis for reference. For the SQ/SD case, the average precision is improved from 55.46% (the best result achieved in [9]) to 67.39%, and the improvements mainly come from the higher syllable accuracies of the spoken documents and the speech queries and the modifications in the vector space representations.

Indexing Approaches	Syllable-based	Character-based	Word-based	Fusions
TQ/TD	97.40 (47.43, 96.56)	97.78 (76.80, 96.04)	90.27 (88.04, 90.03)	97.97 (SC: 97.95, CW: 98.03, SW: 97.58)
SQ/TD	89.82 (41.37, 88,98)	88.11 (66.71, 86.76)	77.55 (74.89, 76.83)	90.22 (SC: 90.45, CW: 87.88, SW: 90.06)
TQ/SD	71.48 (34.56, 70.09)	69.88 (55.77, 68.72)	61.60 (59.88, 61.38)	72.67 (SC: 72.60, CW: 70.03, SW: 72.13)
SQ/SD	67.39 (31.20, 65.83)	65.15 (51.36, 64.29)	55.49 (51.36, 55.34)	68.14 (SC: 68.29, CW: 65.13, SW: 67.80)

Table 3: The SDR results (in **nAP** (%)) of the syllable-, character-, and word-based approaches and their fusions.
(SC, CW and SW denote the fusions of the syllable- and character-level, character- and word-level, and syllable- and word-level information, respectively)

5. FUSION OF SYLABBLE-, CHARACTER- AND WORD-LEVEL INFORMATION

At first, the retrieval performance of the syllable-, character- and word-based indexing approaches are extensively examined, and the results are presented in Columns 2, 3, and 4 of Table 3, respectively. It is obvious that the retrieval performance for the syllable-based indexing approach is very similar to that of the character-based approach when both queries and documents are the perfect transcriptions (97.40% vs. 97.78%), and is even better when either queries or documents or both are the erroneous transcriptions (89.82% vs. 88.11%, 71.48% vs. 69.88%, and 67.39% vs. 65.15%), while the word-based approach always has the worst results. Furthermore, for the word-based approach, the combination of the higher order indexing terms (larger N and n for $S(N)$ and $P(n)$ respectively) only achieved very few improvements. From these results, we can conclude that the subword-based approach (including the character-based and syllable-based approaches) is better than the word-based approach for the Mandarin SDR task, though many research results have indicated that the word-based approach is very useful for the SDR tasks for western languages, such as English [8].

Then, we explore the fusion of syllable- character- and word-level information. The similarity measure between the query and document is modified as:

$$SIM(\vec{q},\vec{d}) = SIM_S(\vec{q},\vec{d}) + SIM_C(\vec{q},\vec{d}) + SIM_W(\vec{q},\vec{d}), \qquad (4)$$

which is simply the sum of the similarity scores of syllable-, character-, word-based indexing approaches and the results are shown in the fifth column of Table 3. The results for the fusion of any two types of information are also provided in the parentheses. It can be found that the fusion of syllable-level information and the character-level information or the word-level information or both achieve better results than using the syllable-level information only. On the other hand, the fusion of the char- and word-level information is still worse than other combinations. The fusion of all the three types of information or the fusion of the syllable- and character-level information can achieve the best results. However, the improvements are in general not significant.

6. INDEXING TERM REDUCTION

The syllable-based indexing approach suffers from a huge size of indexing terms constructed from the multiple syllable hypotheses. It is encouraging to develop techniques to effectively condense the number of indexing terms without degrading the retrieval performance. This section will introduce two techniques that we have used in the SDR experiments.

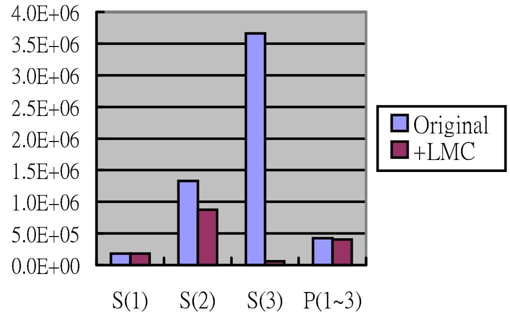

Figure1: The number of different types of syllable-based indexing terms before or after applying the language model constraints for pruning.

6.1 Language Model Constraint (LMC)

First of all, we use the information cues from the contemporary newswire text corpus to filter off the infrequent indexing terms. We assume that the statistical characteristics of syllables in the contemporary newswire text corpus are similar to that of the spoken document collection. The distributions of indexing terms $S(N)$, $N=2\sim3$, and $P(n)$, $n=1\sim3$, in the newswire text corpus were recorded beforehand as the language model constraints (**LMC**) for pruning. Figure 1 plots the number of different types of indexing terms before and after pruning, where the pruning thresholds are different for different types of indexing terms. The number of $S(3)$ (or named the syllable trigram) is dramatically reduced and the total number of indexing terms is reduced to only 27.23% of the original size, while the average precision for the SQ/SD case can be improved from 67.39% to 68.75% as shown in the third row of Table 4.

6.2 Stop Terms (ST)

In word-based SDR systems, usually, a stop word list is used to remove the non-content words. For the syllable-based approach, we could build a syllable-based "stop term" list based on the IDF scores. The M indexing terms with the lowest IDF scores were thought of as the stop terms, and they were removed from the indexing representations. We can see from the fourth row of Table 4 that the retrieval performance can be further improved to 69.01% by simultaneously applying the language model constraints and the stop term list to prune away the indexing terms with very low frequency or IDF scores.

7. FURTHER IMPROVEMENTS IN SYLLAB-BASED SDR

Although most of the information retrieval techniques have been proved to be effective as well for the word-based SDR systems [5-6], there is still little knowledge about their effectiveness in the subword-based retrieval systems. In this section, several information retrieval techniques will be investigated.

7.1 Query Expansion (QE)

7.1.1 Blind Relevance Feedback (BRF)

Some indexing terms, which do not appear in the query, may still act as good predictors of relevance judgments [7-8]. For example, the information from the relevant or irrelevant documents achieved in the first stage retrieval could be used to identify which indexing terms are good predictors. In this paper, a blind relevance feedback procedure is devised to automatically reformulate the initial query expression without the user intervention by reweighing the original indexing terms or adding the new terms based on the modified Rocchio formula [5]:

$$\vec{q}' = \alpha \cdot \vec{q} + \beta \cdot \sum_{\vec{d}_i \in D_r} \vec{d}_i - \gamma \sum_{\vec{d}_j \in D_{irr}} \vec{d}_j, \qquad (5)$$

where $\vec{q}$ is the initial query expression, $\vec{q}'$ is the modified query expression, D_r is the set of relevant documents, D_{irr} is the set of irrelevant documents, and α, β and γ are empirically adjustable weighting parameters.

7.1.2 Term Association Matrix (TAM)

The indexing terms co-occurring frequently in the spoken documents were assumed to have the synonymity association [5]. Thus, we could build a global association matrix of the indexing terms, in which each entry (i, j) stands for the correlation factor of indexing terms t_i and t_j, and is expressed as:

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

where c_i and c_j are respectively the total occurrences of indexing term t_i and t_j in the document collection, and $\hat{c}_{i,j}$ is the co-occurrences of indexing term t_i and t_j within the same document. In this paper, Equation (6) is calculated directly from the spoken document collection. The query expression is then reformulated by adding the synonymity terms of each query term into the initial query expression.

7.2 Document Expansion (DE)

In this paper, we also explore document expansion [5-6] directly from the spoken document collection. The concept of document expansion is quite similar to that of blind relevant feedback: each document is thought of as a query and used to retrieve its relevant documents within the document collection and, then, the document expression is reformulated by reweighing the original indexing terms or adding the new indexing terms according to the retrieved relevant documents.

7.3 Experimental Results

The detailed results after applying the above schemes are summarized in Table 4. Firstly, we can find from the fifth and sixth rows of Table 4 that before applying the blind reference feedback techniques, the stop term list should be applied first to prune away the indexing terms with very low IDF scores. Secondly, it's obvious that document expansion and query expansion based on the term association matrix are not as effect as blind reference feedback. Generally speaking, the performance improvement achieved by applying these IR schemes is not as significant as what have been reported in other evaluations [5-7]. In this study, each spoken document is a news abstract which contains several news items. As a result, though these schemes could introduce some relevant terms into the initial document or query expression, they could also bring it a huge amount of irrelevant terms at the same time. Of course, the recognition errors could be one of the reasons, too. Nevertheless, with all these approaches, the average precision can be improved to 71.29%.

Spoken Queries retrieving Spoken Documents	nAP (%)
Baseline	67.39
+LMC	68.75
+LMC+ST	69.01
+LMC+ST+BRF	71.11
+LMC+ BRF	69.82
+LMC+ST+ TAM	69.41
+LMC+ST+ DE	68.97
+LMC+ST +TAM+BRF	71.23
+LMC+ST+DE +TAM+BRF	71.29

Table 4: The **nAP** (%) of syllable-based SDR with combinations of different techniques.

8. CONCLUSIONS

This paper presented the recent experiments that we have done to improve our syllable-based Mandarin broadcast news retrieval system. We have explored the fusion of multi-scale indexing terms, indexing term reduction, document expansion, and query expansion. We found that both document expansion and query expansion could introduce a huge amount of irrelevant terms into the initial document or query expression because each document contains several news items in our task. The passage-level retrieval techniques [5] are believed to be very useful for this task if the document can be automatically segmented into several passages.

9. REFERENCES

[1] K. Spärck Jones, G. J. F. Jones, J. T. Foote and S. J. Young, "Experiments on Spoken Document Retrieval," Information Processing & Management, 32(4), pp. 399-417, 1996.

[2] CMU Informedia Digital Video Library project http://www.informedia.cs.cmu.edu/

[3] M. Wechsler, "Spoken Document Retrieval Based on Phoneme Recognition," Ph.D. thesis, Swiss Federal Institute of Technology (ETH), Zurich, 1998.

[4] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, "Modern Information Retrieval," 1999.

[5] Amit Singhal and Fernando Pereira, "Document Expansion for Speech Retrieval," SIGIR 1999.

[6] P.C. Woodland, S.E. Johnson, P. Jourlin, and K. Spärck Jones, "Effects of Out of Vocabulary Words in Spoken Document Retrieval," SIGIR 2000.

[7] Rila Mandala, Takenobu Tokunaga, and Hozumi Tankaka, "Query Expansion Using Heterogeneous Thesauri," Information Processing & Management 36, pp. 361-378, 1996.

[8] Kenney Ng, "Information Fusion for Spoken Document Retrieval," ICASSP 2000.

[9] Berlin Chen, Hsin-min Wang, and Lin-shan Lee, "Retrieval of Broadcast News Speech In Mandarin Chinese Collected In Taiwan Using Syllable-Level Statistical Characteristics," ICASSP 2000.