SPEECH INFORMATION RETRIEVAL FOR MANDARIN CHINESE

Hsin-min Wang

Institute of Information Science, Academia Sinica
Taipei, Taiwan
E-mail: whm@iis.sinica.edu.tw

ABSTRACT

Network technology and the Internet are creating a completely new information era. With the rapidly growing use of audio and multi-media information on the Internet, an exponentially increasing number of spoken documents, such as broadcast radio and television programs, video tapes, digital libraries and so on, are now being accumulated and made available. However, most of them are simply stored there and are difficult to further reuse because of the lack of efficient retrieval technology. Development of technology for retrieving speech information is therefore becoming more and more important. Our team has been investigating audio indexing and retrieval of spoken documents in Mandarin Chinese, with a special consideration on the characteristics and monosyllabic structure of the Chinese language. This paper reports some interesting results and findings obtained in this research.

1. INTRODUCTION

Massive quantities of video and audio recordings, such as broadcast radio and television programs, are becoming available on the Internet in the global information infrastructure. The Informedia Digital Video Library project at Carnegie Mellon University has pioneered new approaches for automated video and audio indexing, navigation, visualization, search, and retrieval [1]. More recently, spoken document retrieval applications are crossing the threshold of practicality, as evidenced by Compaq's SpeechBot [2], which is a web-based English spoken document retrieval system [3]. Moreover, many other research institutes and universities have been involved in video and audio indexing and retrieval research in recent years [4-8]. In this paper, we describe our efforts towards the development of spoken document retrieval technology for Mandarin Chinese.

In Mandarin Chinese, there is an unknown number of words, though only a portion of it is commonly used. Each word is composed of one or more characters, while each character is pronounced as a monosyllable and is a morpheme with its own meaning. As a result, new words are easily generated every day by combining a few characters. For example, the combination of 電 (electricity) and 腦(brain) gives a new word, 電腦(computer). Mandarin Chinese is phonologically compact; an inventory of about 400 base syllables provides full phonological coverage of Mandarin audio. On the other hand, an inventory of about 13,000 characters provides full textual coverage of written Chinese (in Big5 code). There is a many-to-many mapping between characters and syllables. For example, the character 乾 may be pronounced as /gan1/ or /qian2/ while all of the characters 甘千柑肝竿尷疳 are also pronounced as /gan1/ and all of 前錢潛黔虔捐 are pronounced as /qian2/. Consequently, a foreign word can very often be translated into different Chinese words. For example, "Kosovo" in "As the Kosovo peace talks in France…" may be translated into 科索沃/ke1-suo3-wo4/, 科索佛/ke1-suo3-fo2/, 科索夫/ke1-suo3-fu1/, 科索伏/ke1-suo3-fu2/, 柯索佛/ke1-suo3-fo2/, etc. Accordingly, syllable recognition is believed to be a key problem in Mandarin Chinese speech recognition [9].

The characteristics of the Chinese language also lead to some special considerations for the spoken document retrieval task. Word-level indexing features possess more semantic information than subword-level features; thus word-based retrieval enhances precision. On the other hand, subword-level indexing features are more robust against Chinese word tokenization ambiguity, Chinese homophone ambiguity, the open vocabulary problem, and speech recognition errors; thus subword-based retrieval enhances recall. Consequently, there is good reason to use information fusion of indexing features of different levels. In [8], we have shown that syllable-level indexing features are very effective for Mandarin Chinese spoken document retrieval and the retrieval performance can be improved by integrating information of character-level and word-level indexing features.

The rest of this paper is organized as follows: Our approaches for speech recognition and spoken document retrieval are discussed in Sections 2 and 3, respectively. The prototype retrieval systems are presented in Section 4. Finally, conclusions are made in Section 5.

2. SPEECH RECOGNITION

This section will introduce our speech recognition approaches.

2.1. Signal Processing

In our speech recognizer, spectral analysis is applied to a 20 ms frame of speech waveform every 10 ms. For each speech frame, 12 mel-frequency cepstral coefficients (MFCC) and the logarithmic energy are extracted, and these coefficients along with their first and second time derivatives are combined to form a 39-dimensional feature vector. In addition, to compensate for channel noise, utterance-based cepstral mean subtraction (CMS) is applied to the training and testing speech.

2.2. Acoustic Modeling

Considering the monosyllabic structure of the Chinese language in which each syllable can be decomposed into an INITIAL/FINAL format, the acoustic units used in our speech recognizer are intra-syllable right-context-dependent INITIAL/FINAL, including 112 context-dependent INITIALs and 38 context-independent FINALs. Each INITIAL or FINAL is represented by a continuous density HMM (CDHMM) with 1 to 4 states. In addition, the silence model is a 1-state CDHMM trained by using the non-speech segments. The acoustic models for recognition of the broadcast news speech were trained by using 16 hours of broadcast news speech collected on air from several radio stations located at Taipei while the acoustic models for recognition of the speech queries were trained by using 5.3 hours of microphone speech.

2.3. Language Modeling

The syllable-based and word-based *N*-gram language models were trained by using a newswire text corpus consisting of 65 million Chinese characters collected from Central News Agency (CNA) in 1999. Word segmentation and phonetic labeling were performed for the training text corpus based on a 61521-word lexicon for training the *N*-gram language models.

2.4. Large-Vocabulary Continuous Speech Recognition

Our speech recognizer adopts a multi-pass search strategy. In the first pass, Viterbi search is performed based on the acoustic models and the syllable bigram language model, and the score at every time index is stored. In the second pass, a backward time-asynchronous A* search [10] generates the best syllable sequence based on the heuristic scores obtained from the first pass search and the syllable trigram language model. In the third pass, based on the state likelihood scores evaluated in the first pass search and the syllable boundaries of the best syllable sequence obtained in the second pass, the speech recognizer further performs Viterbi search on each utterance segment

which may be a syllable and produces several most likely syllable candidates, and a syllable lattice can thus be constructed. In the forth pass, the recognizer further constructs the word graph from the syllable lattice based on the 61521-word lexicon and performs dynamic programming on it to find the best word sequence using the word unigram and bigram language models. The finally obtained word sequence will then be automatically converted into its equivalent character-level and syllable-level sequences to be used in the retrieval task.

2.5. Keyword Spotting

The search framework of our keyword spotter is based on a lexical network and a filler model. Each arc of the lexical network represents a subword unit, thus the network is able to handle any arbitrarily assigned keyword set. Here, the syllable is chosen as the subword unit and each syllable is composed of two sub-syllabic models, as described in Section 2.2. The filler model, which is composed of a general INITIAL model and a general FINAL model, is used to cover the surrounding non-keyword part of the input utterances.

The keyword spotting process is based on a multi-pass A* search strategy. In the first pass, the filler model is respectively decoded forward with the cumulative score for the last state (denoted as the forward filler model score later) stored at every time index and backward with the cumulative score for the first state (denoted as the backward filler model score later) stored at every time index. In the second pass, a compact syllable lattice with the respective lattice node scores evaluated forward is generated from the lexical network based on the constraint grammar considering the keyword structure and the forward filler model scores. The cumulative score of the end node of each arc is stored at every time index, which will be used as the heuristic function in the following A* search. The heuristic score of the end node n_k of arc k at time t is represented as:

$$h^*(n_k,t)=\max_{0\le t_1<t} f_f(t_1)+h(n_k,t_1+1,t), \tag{1}$$

where $f_f(t_1)$ is the forward filler model score at time t_1 and $h(n_k,t_1+1,t)$ is the cumulative score of the best path which enters the syllable lattice at time t_1+1 and reaches n_k at time t. In the third pass, a backward time-asynchronous A* search based on the right filler model scores, the lexical network, and the heuristic scores is performed. The evaluation function $E_p(n_k)$ for extending a partial path p to node n_k is represented as:

$$E_p(n_k)=\max_{0<t<T} d_p(t,T-1)+h^*(n_k,t-1), \tag{2}$$

where $d_p(t,T-1)$ is the exactly decoded score of the extended partial path p, and T is the duration of the speech utterance. In every iteration of the A* search, the partial path with the maximal evaluation function is popped, extended, and stored in a stack. This process is terminated when the N-best complete paths (candidate keywords) are obtained sequentially.

3. SPOKEN DOCUMENT RETRIEVAL

This section will introduce our spoken document retrieval approaches.

3.1. Indexing Terms

In [8], we have shown that the overlapping syllable N-grams (N=1~3) and the overlapping syllable pairs separated by *n* (*n*=1~3) syllables are very effective for Mandarin Chinese spoken document retrieval. The overlapping syllable N-grams 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. We have also shown that

retrieval performance can be improved by integrating information of overlapping character N-grams and words into indexing. As mentioned in Section 2.4, each spoken document can be transcribed into a syllable lattice, a character sequence and a word sequence. Accordingly, eight types of indexing terms can be extracted from the recognition output of a spoken document; they are syllable unigram, syllable bigram, syllable trigram, syllable pairs separated by *n* (*n*=1~3) syllables, character unigram, character bigram, character trigram, and word unigram.

3.2. Information Retrieval Model

Vector space models widely used in many text information retrieval systems are used here. A document is represented as a set of feature vectors, each consisting of information regarding one type of indexing terms. Here, eight types of indexing terms are used to construct eight feature vectors for each document d,

$$\vec{d}_j = (x_{j1}, x_{j2}, ..., x_{jt}, ..., x_{jM_j}), \quad j = 1,2,3,...,8, \tag{3}$$

where $\vec{d}_j$ is the feature vector for the *j*-th type of indexing terms, the *t*-th component of $\vec{d}_j$, x_{jt}, represents the score for a specific indexing term t, and M_j is the total number of different specific indexing terms for the *j*-th type. The value of x_{jt} is obtained by

$$x_{jt} = [1 + \ln(\sum_{i=1}^{n_t} c_t(i))] \cdot \ln(N/N_t). \tag{4}$$

For character-based or word-based indexing terms, $c_t(i)$ is set to 1. n_t is the total frequency count for the occurrence of the specific indexing term t in the document. The value of $\ln(N/N_t)$ is the Inverse Document Frequency (IDF), where N_t is the total number of documents in the collection in which the specific indexing term t appears, and N is the total number of documents in the collection. The value of x_{jt} in Equation (4) is set to zero if the specific indexing term *t* didn't appear in the document d.

For syllable-based indexing terms, $c_t(i)$, ranging from 0 to 1, is the confidence measure evaluated for the *i*-th occurrence of the specific indexing term *t* within the document d. As mentioned in Section 2.4, each spoken document will be transcribed into a syllable lattice. Each utterance segment O which may be a syllable can have several syllable candidates. For a certain syllable candidate s of the utterance segment O, the confidence measure $c(s)$ is obtained with the following Sigmoid function:

$$c(s) = \frac{2}{1 + \exp(-\alpha \times [\log p(O \mid s) - \log p(O \mid s^*)])}, \tag{5}$$

where $\log p(O \mid s)$ and $\log p(O \mid s^*)$ are the original recognition scores of syllable s and its corresponding top 1 syllable candidate s^*, respectively, and the value of α is used to control the slope of the Sigmoid function. From Equation (5), it is clear that $c(s)$=1 if $s = s^*$. Also, $c(s)$ is always between 0 and 1. The confidence measure of a specific indexing term t, c_t, is simply the average of the confidence measures for all syllables involved in the specific indexing term t.

A query is also represented by 8 feature vectors in the same way as the documents. The Cosine measure is used to estimate the query-document relevance for the *j*-th type of indexing terms:

$$R_j(\vec{q}_j, \vec{d}_j) = (\vec{q}_j \cdot \vec{d}_j) \Big/ (\|\vec{q}_j\| \times \|\vec{d}_j\|), \tag{6}$$

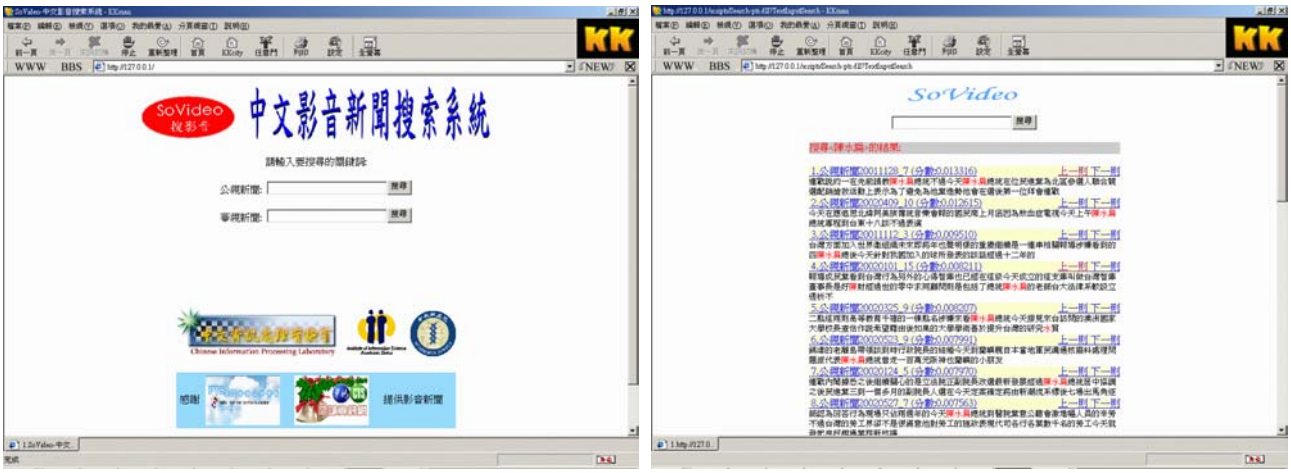

Figure 1. The SoVideo Web-based Mandarin Chinese broadcast news retrieval system.

where $\vec{q}_j$ is the feature vector for the query using the j-th type of indexing terms. The overall relevance measure is then the weighted sum of the relevance measures of all types of indexing terms:

$$R(\vec{q},\vec{d}) = \sum_j w_j \times R_j(\vec{q}_j,\vec{d}_j), \tag{7}$$

where w_j is a weighting parameter obtained empirically.

4. PROTOTYPE SYSTEMS

4.1. Sovideo: A Web-based Mandarin Chinese Broadcast News Retrieval System

We have implemented a web-based Mandarin Chinese broadcast news retrieval system, SoVideo. As shown in Figure 1, it functions as an audio search engine, which allows users to input search terms to search for their desired news stories from the broadcast news database. Given the search terms, the IR server will first tokenize them and output the corresponding word and syllable strings. Then, the indexing feature vectors corresponding to the word/character/syllable N-grams can be constructed respectively and used to compute the query-document relevance. Finally, the IR server will return a HTML file containing the ranking results and the URLs of all the relevant spoken documents.

Currently, the target database consists of 177 Mandarin Chinese broadcast news recordings collected from Public Television Service Foundation (Taiwan). Each recording consists of a broadcast news episode of 60 minutes. There is a total of 3264 stories by automatic story segmentation. Recordings are in mono with 16kHz sampling rate and 16 bit resolution.

4.1.1. Speech recognition evaluation

We have conducted some speech recognition experiments based on 4 one-hour shows which have been carefully transcribed and annotated. The recognition results are summarized in Table 1. The average character accuracy is 41.59% while the average syllable accuracy is 51.09%. The accuracy for the interviewees' speech is extremely poor but the accuracy for the anchors' speech is relatively reasonable. Also, it's obvious from Table 1 that the background music seriously degrades the recognition accuracy. The recognition accuracy can be further improved by using un-supervised model adaptation techniques such as MLLR, but, until now, we have not applied any of these in our recognizer.

	Anchor		Reporter		Interviewee	Average
	Without background music	With background music	Without background music	With background music		
Syllable	69.48%	47.73%	60.02%	42.14%	26.54%	51.09%
Character	64.02%	38.68%	50.22%	30.01%	15.68%	41.59%

Table 1. Recognition accuracy for the broadcast news speech.

4.1.2. Information retrieval evaluation

We have tested SoVideo using a set of 40 keyword queries. On average, each query contains 4.0 characters. For each query, the system returned 20 documents. Because the relevance judgment is not available, we are not able to obtain the traditional recall/precision graph. Two performance measures are used instead, namely the mean average precision and the percentage of queries for which the relevant documents are ranked in the very top group. Here, the mean average precision is defined as follows: Given a fixed number of retrieved documents, the precision average for each query is obtained by computing the precision after every retrieved relevant document and then averaging these precisions over the total number of retrieved relevant documents. These query averages are then averaged across all queries. The mean average precisions obtained in this way are 0.975, 0.944, 0.911, 0.894, and 0.871, respectively, when 1, 5, 10, 15, and 20 returned documents were considered. Using the second performance measure, we found that 97.5% of the queries are able to get the relevant document when only one document is returned while 100% of queries are able to get at least one relevant documents within 3 retrieved documents. By taking advantage of this situation, we can apply relevance feedback techniques to enhance retrieval performance.

4.2. A Voice-Activated Mandarin Chinese Broadcast News Retrieval System

We have also implemented a voice-activated Mandarin Chinese broadcast news retrieval system. This system accepts both speech queries and text queries and provides some voice command control functions, such as “stop playback”, “play the next story”, “play the previous story”, etc. Currently, the target database consists of 6646 Mandarin Chinese broadcast news stories (56 hours) collected on air from several radio stations located at Taipei. Recordings are in mono with 16kHz sampling rate and 16 bit resolution. The user interface of this system is depicted in Figure 2.

5. CONCLUSIONS

This paper described our efforts towards the development of spoken document retrieval technology for Mandarin Chinese. We have successfully implemented prototype broadcast news retrieval systems. We presented our speech recognition and information retrieval approaches and reported on some preliminary evaluation of these systems.

6. ACKNOWLEDGEMENTS

The author would like to thank all members of his team for their contributions to this research and Public Television Service Foundation and the Chinese Television System for providing the TV news. The author would also like to acknowledge financial support from National Science Council, Institute for Information Industry, and Academia Sinica.

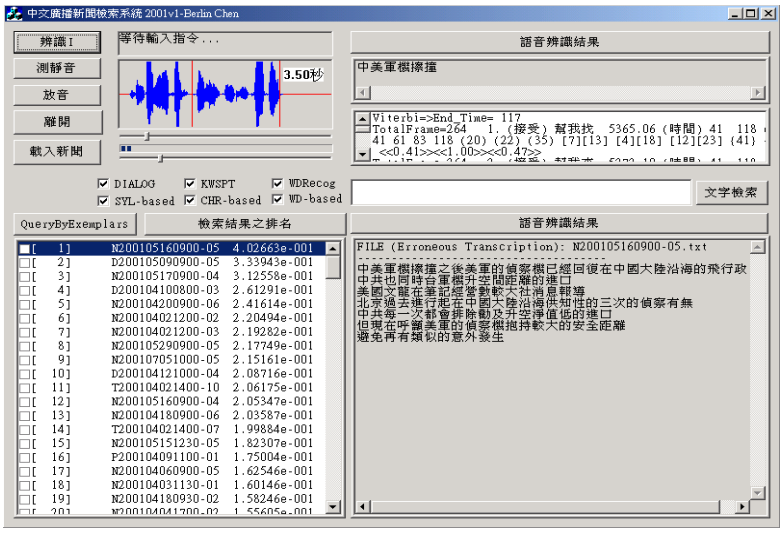

Figure 2. The voice-activated Mandarin Chinese broadcast news retrieval system.

7. REFERENCES

[1] Informedia Website, "http://www.informedia.cs.cmu.edu/".

[2] B. Logan, P. Moreno, J. M. Van Thong, and E. Whittaker, "An Experimental Study of An Audio Indexing System for the Web," *ICSLP2000*.

[3] SpeechBot Website, "http://speechbot.research.compaq.com/".

[4] 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.

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

[6] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, A. Srivastava, "Speech and Language Techniques for Audio Indexing and Retrieval," *Proc. IEEE*, 88(8), pp. 1338-1353, 2000.

[7] S. Renals, A. Abberley, D. Kirby, T. Robinson, "Indexing and Retrieval of Broadcast News", *Speech Communication*, 32(1-2), pp. 5-20, 2000.

[8] B. Chen, H. M. Wang, and L. S. Lee, "Discriminating Capabilities of Syllable-based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese," *IEEE Trans. on Speech and Audio Processing*, 10(5), pp. 303-314, 2002.

[9] L. S. Lee, "Voice Dictation of Mandarin Chinese," *IEEE Signal Processing Magazine*, 14(4), pp. 63-101, 1997.

[10] P. Kenny, R. Hollan, V. N. Gupta, M. Lennig, P. Mermelstein, and D. O'Shaughnessy, "A*-Admissible Heuristics for Rapid Lexical Access", *IEEE Trans. on Speech and Audio Processing*, 1(1), pp. 49-58, 1993.