The SoVideo broadcast news retrieval system for Mandarin Chinese

Hsin-min Wang, Shi-sian Cheng, and Yong-cheng Chen

Institute of Information Science, Academia Sinica
Taipei, Taiwan

Email: {whm, sscheng, yc_chen}@iis.sinica.edu.tw

ABSTRACT

This paper describes the SoVideo broadcast news retrieval system for Mandarin Chinese. The system is based on technologies such as large-vocabulary continuous speech recognition for Mandarin Chinese, automatic story segmentation, and information retrieval. Until now, the database consisted of 177 hours of broadcast news, which yielded 3264 stories by automatic story segmentation. We discuss the development of the retrieval system, and the evaluation of each component and the retrieval system.

1. INTRODUCTION

Massive quantities of audio and multimedia content, such as broadcast radio and television programs, are becoming available on the Internet in the global information infrastructure. As a result, spoken document retrieval (SDR) has been extensively studied in recent years [1-3]. Recently, spoken document retrieval applications are crossing the threshold of practicality, as evidenced by Compaq's SpeechBot [4], which is a web-based English spoken document retrieval system.

In this paper, we describe our efforts towards the development of a broadcast news retrieval system for Mandarin Chinese. By integrating technologies such as large-vocabulary continuous speech recognition, story segmentation, and spoken document retrieval, we have successfully implemented a web-based Mandarin Chinese broadcast news retrieval system, SoVideo. Currently, the target database consists of 177 hours of Mandarin Chinese broadcast news, which yields 3264 stories by automatic story segmentation. The preliminary test results when using a set of 40 keyword queries show 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 document within 3 retrieved documents.

The rest of this paper is organized as follows: The broadcast news corpus used in this paper is described in Section 2. The characteristics of the Chinese language are briefly introduced in Section 3. Our approaches for speech recognition, story segmentation and spoken document retrieval are discussed in Sections 4, 5, and 6, respectively. Finally, the prototype retrieval system is presented in Section 7 and conclusions are made in Section 8.

2. DATA COLLECTION

In August 2001, our group started a speech corpus collection project. We expect to collect and annotate 220 hours of Mandarin Chinese broadcast news speech over 3 years. Public Television Service Foundation (Taiwan) has kindly agreed to share their broadcast news with us. A Digital Audio Tape (DAT) recorder, which is connected to the broadcasting machine using the XLR balanced cable, has been set up in the TV broadcasting studio. That is, the broadcast news speech was recorded synchronously while broadcasting to avoid the modulation effect. Recordings are on DATs in stereo with 44kHz sampling rate and 16 bit resolution. Each recording consists of a broadcast news episode of 60 minutes. Each DAT was manually processed to transfer the digital speech samples into a single Microsoft Windows wave file and stored in the hard disk. Then, the signal was down-sampled to 16kHz with a resolution of 16 bits. During this operation, only the left channel was selected. Until now, about 200 hours of broadcast news have been recorded in this way.

The corpus has been segmented, labeled and transcribed manually using a tool developed by DGA (Délégation Générale pour l'Armement, France) and LDC, called "Transcriber"[5]. The transcripts are in BIG5-encoded form, with SGML tagging to annotate acoustic conditions, background conditions, story boundaries, speaker turn boundaries and audible acoustic events, such as hesitations, repetitions, vocal non-speech events, external noises, etc. These tags include time stamps to align the text with the speech data. The speech segments from anchors, reporters, interviewees, etc. are carefully transcribed while the remaining segments containing advertising or pure music are just annotated with time stamps without orthographic transcripts. The first interim 40-hour

corpus has been completed, on which we can conduct speech recognition evaluation and story segmentation evaluation, while the transcription and annotation work for the remaining material is still in progress.

3. CHARACTERISTICS OF THE CHINESE LANGUAGE

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 甘干柑肝竿尷疳 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 and a multi-pass search strategy is usually adopted [6].

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 [3], 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 details of our multi-scale audio indexing and retrieval method will be presented in Section 6.

4. SPEECH RECOGNITION

This section will introduce our speech recognition approach.

4.1. 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.

4.2. 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. The Gaussian mixture number per state ranges from 4 to 64, depending on the amount of corresponding training data available. In addition, the silence model is a 1-state CDHMM with 64 Gaussian mixtures trained by using the non-speech segments. The acoustic models were trained by using a database with 16 hours of broadcast news speech collected on air from several radio stations located at Taipei and finally a total of 11004 mixtures were obtained.

4.3. 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.

4.4. 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* tree search generates the best syllable

	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. The syllable and character recognition accuracies for the broadcast news speech.

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.

4.5. Speech Recognition Experiments

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.

5. STORY SEGMENTATION

Accurate segmentation of an audio stream is a key process to improve the performance for transcription and retrieval of broadcast news. Recently, we have developed a simple but effective multi-pass approach for automatic story segmentation. The first pass performs speaker and environment change detection. The second pass conducts hierarchical clustering of audio segments. We assume that the largest cluster is the anchor cluster and every anchor speech segment is the first segment of a story. In this way, the number of anchor segments corresponds to the number of stories in the audio stream, and the starting time of a story is the starting time of its anchor segment. In this approach, speaker and environment change detection and audio segment clustering both play key roles.

Various segmentation algorithms have been proposed in the literature [7-8]. Metric-based segmentation is proposed to segment the audio stream at maxima of the distances between neighboring windows placed at every sample [7]. This method is flexible since no or little prior knowledge about the audio signal is needed to decide on the segmentation points, but it relies on thresholding of measurements. Chen and Gopalakrishnan proposed a maximum likelihood approach [8]. The audio stream is modeled as a Gaussian process in the cepstral space and the *Bayesian Information Criterion* (BIC), a model selection criterion well-known in the statistics literature, is applied to detect turns of a Gaussian process. Recently, we have integrated BIC into the metric-based segmentation framework and designed a hierarchical algorithm for clustering of audio segments by using BIC as a termination criterion. The details for our story segmentation approach will be described in the following sections.

5.1. Model Selection via BIC

The problem of model selection is to choose one among a set of candidate models to describe a given data set. Let $X = \{x_1, x_2, ..., x_N\}$ be the data set we are modeling and $M = \{M_1, M_2, ..., M_K\}$ be the candidate model set. The BIC is then defined as:

$$BIC(M_i) = \log L(X, M_i) - \lambda \frac{1}{2} \#(M_i) \times \log(N), \quad (1)$$

where $L(X, M_i)$ is the maximum likelihood of X under M_i, $\#(M_i)$ is the number of parameters in model M_i, while the penalty weight $\lambda = 1$. The BIC procedure is to choose the model for which the BIC value is maximized.

5.2. Metric-based Change Detection via BIC

For metric-based speech segmentation approaches, the audio stream is first encoded in terms of cepstral vectors. Then the distance between each pair of consecutive windows of cepstral vectors is measured. Since it's complicated to directly measure the distance between two collections of vectors, both windows of features are often individually first modeled parametrically by distributions such as Gaussian, and then many distance measures between two parametric statistical models can be applied, e.g. the KL2 distance [7]. Our metric-based change detection approach is depicted in Figure 1. It calculates the *deltaBIC* value instead of the distance between each pair of consecutive windows of cepstral features. The *deltaBIC* is defined as:

$$deltaBIC = BIC(M_1) - BIC(M_0). \quad (2)$$

We are comparing two models: One models the two windows of cepstral vectors as two multivariate Gaussians; i.e., $M_1 : x_{t-n+1},...,x_t \sim N(\mu_1, \Sigma_1); x_{t+1},...,x_{t+n} \sim N(\mu_2, \Sigma_2)$. The other models the data as just one multivariate Gaussian; i.e., $M_0 : x_{t-n+1},...,x_t, x_{t+1},...,x_{t+n} \sim N(\mu, \Sigma)$. If t is a change point, we must have $deltaBIC > 0$. The local peaks of the *deltaBIC* curve with a peak width larger than 0.4 of the window width are detected as the segmentation points. Here the window width is 3 seconds and the width of a peak is defined as the time span of its neighboring points with *deltaBIC* larger than 0.

5.3. Clustering via BIC

Let $S = \{s_1, s_2, ..., s_L\}$ be the collection of audio segments we want to cluster, each segment s_i is associated with a sequence of cepstral vectors $X^i = \{x_1^i, x_2^i, ..., x_{N_i}^i\}$. Given two segments, s_i and s_j, we are comparing two models: One models the data as two multivariate Gaussians; i.e., $M_1 : x_1^i,...,x_{N_i}^i \sim N(\mu_i, \Sigma_i); x_1^j,...,x_{N_j}^j \sim N(\mu_j, \Sigma_j)$. The other models the data as just one multivariate Gaussian; i.e., $M_0 : x_1^i, x_2^i,...,x_{N_i}^i, x_1^j, x_2^j,...,x_{N_j}^j \sim N(\mu, \Sigma)$. If $deltaBIC = BIC(M_1) - BIC(M_0) < 0$, s_i and s_j will be merged as one segment.

Based on the BIC merging criterion, we have developed a bottom-up hierarchical clustering algorithm. To speed up clustering and maintain the clustering accuracy, the candidate pairs to be tested are first ranked according to their KL2 distances. The merging test starts from the top of the candidate pair list. Once two segments are merged, the remaining candidate pairs containing any of these segments will be skipped. The merging test stops when the whole list is completed. Then, the clustering algorithm ranks candidate pairs again and performs merging tests according to the rank list. The procedures run iteratively till no pair can be merged.

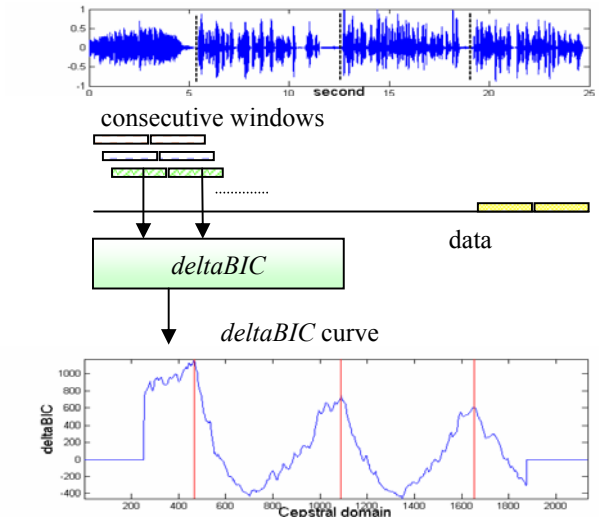

Figure 1. The procedures of metric-based segmentation via BIC

5.4. Story Segmentation Experiments

We have conducted some story segmentation experiments based on the same 4 one-hour shows used for speech recognition evaluation in section 4.5. There are 88 stories in total; i.e., 88 anchor segments. After segmentation and clustering, 92 segments were judged as anchor segments. Among them, 4 segments were not produced by anchors, which means all the 88 anchor segments were detected. Therefore, the recall and precision rates are 1.0 (88/88) and 0.957 (88/92), respectively. As a result, among the 88 stories, 4 stories were divided in two to give a total of 8 stories, and the automatic story segmentation method finally resulted in 92 stories. We have also looked at the segmentation positions. Among the 88 stories whose beginning anchor segments were correctly detected, there were 81 stories whose starting time errors were within 2 seconds and 4 stories whose errors were between 2 and 3 seconds while, for the remaining 3 stories, the errors were 5.4, 6.1, and 16.6 seconds, respectively.

6. SPOKEN DOCUMENT RETRIEVAL

This section will introduce our spoken document retrieval approach.

6.1. Indexing Terms

In [3], 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 4.4, 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.

6.2. 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,

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

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

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

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 4.4, 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:

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

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:

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

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:

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

where w_j is a weighting parameter obtained empirically.

7. THE SOVIDEO BROADCAST NEWS RETRIEVAL SYSTEM

We have successfully implemented a web-based Mandarin Chinese broadcast news retrieval system, SoVideo (http://sovideo.iis.sinica.edu.tw), by integrating the above speech recognition, story segmentation, and spoken document retrieval approaches. SoVideo 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 similarities between the query and the documents. 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 hours of Mandarin Chinese broadcast news, which yields 3264 stories by automatic story segmentation. Since the recognition accuracy for the field reports is obviously worse than that for the anchor speech, speech recognition was only applied to the anchor speech and the indexing is only based on the anchor speech.

7.1 Retrieval Experiments

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 (mAP) and the percentage of queries for which the relevant documents are ranked in the very top group. Here, the mAP 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. To be more specific, the mAP is calculated using Equation (8):

$$\text{mAP} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{N_i}\sum_{k=1}^{N_i}\frac{k}{rank_{ik}} \qquad (8)$$

where L is the total number of queries, N_i is the total number of relevant documents contained in the retrieved documents for query q_i, and $rank_{ik}$ is the rank of the k-th relevant document for query q_i. The mAPs 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.

8. CONCLUSIONS

We have successfully implemented a web-based Mandarin Chinese broadcast news retrieval system, SoVideo, by integrating technologies such as large-vocabulary continuous speech recognition, story segmentation, and spoken document retrieval. We presented our speech recognition, story segmentation, and information retrieval approaches and reported on the preliminary evaluation of the system.

9. ACKNOWLEDGEMENTS

This work was funded by Academia Sinica and the National Science Council of the Republic of China under grant No. NSC 91-2219-E-001-009. The corpus collection project was funded by the National Science Council of the Republic of China under grant No. NSC 90-2213-E-009-109. The authors would like to thank Public Television Service Foundation (Taiwan) for sharing their broadcast news with us.

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

[3] 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, July 2002.

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

[5] C. Barras, E. Geoffrois, Z. B. Wu, M. Liberman, "Transcriber: Development and Use of S tool for Assisting Speech Corpora Production," Speech Communication, 33, pp. 5-22, 2001.

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

[7] M. Siegler, U. Jain, B. Ray and R. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," *DARPA Speech Recognition Workshop*, 1997.

[8] S. Chen and P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.