# The SoVideo Mandarin Chinese Broadcast News Retrieval System

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

*Institute of Information Science, Academia Sinica, Taipei, Taiwan*

whm@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. Currently, the database consists of 177 hours of broadcast news, which yields 3264 stories by automatic story segmentation. We discuss the development of the retrieval system, and the evaluation of each component and the retrieval system.*

Keywords: broadcast news, speech recognition, story segmentation, spoken document retrieval

## 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 (http://www.informedia.cs.cmu.edu) has pioneered new approaches for automated video and audio indexing, navigation, visualization, search, and retrieval (Wactlar et al., 1996, Hauptmann et al., 2001). More recently, spoken document retrieval applications are crossing the threshold of practicality, as evidenced by Compaq's SpeechBot (http://speechbot.research.compaq.com), which is a web-based English spoken document retrieval system (Logan et al., 2000). Moreover, many other research institutes and universities have been involved in video and audio indexing and retrieval research in recent years (Jones et al., 1996, Wechsler, 1998, Makhoul et al., 2000, Ng, 2000, Renals et al., 2000). As to the Chinese language, Chen et al. (2002) and Meng et al. (2000a) have investigated the use of multi-scale units to index Chinese spoken documents in Mandarin and Cantonese, respectively.

Started in 1997, the Topic Detection and Tracking (TDT) (Wayne, 2000) research develops algorithms for discovering and threading together topically related material in streams of data such as newswire and broadcast news in both English and Mandarin Chinese. The evaluation tasks include new event detection, story link detection, story segmentation, topic detection, and topic tracking. Based on the TDT corpora, the MEI project (Meng et al., 2000b) developed one of the first cross lingual spoken document retrieval (CL-SDR) systems that can retrieve Mandarin broadcast news using English newswire articles as query exemplars. 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 some are 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) yields the 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 be translated into different Chinese words based on its pronunciation. For example, Kosovo 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., while Al Qaeda may be translated into 開打/kai1-da3/, 凱達/kai3-da2/, 蓋達/gai4-da2/, 卡達/ka3-da2/, 卡伊達/ka3-i1-da2/, 阿爾蓋達/a1-er3-gai4-da2/, etc. Different translations usually have some syllables in common, or may have exactly the same syllables.

The characteristics of the Chinese language lead to some special considerations while performing Mandarin Chinese speech recognition, e.g. syllable recognition is believed to be a key problem (Lee, 1997), performance evaluation is usually based on syllable accuracy and character accuracy rather than word accuracy, etc. 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 Chen et al. (2002), 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.

Accurate segmentation of an audio stream is a key process to improve the performance for transcription and retrieval of broadcast news. Various segmentation algorithms have been proposed in the literature. Kubala et al. (1997) proposed a decoder-guided segmentation approach. The input audio stream is first decoded. Then, the desired segments can be produced by cutting the input at the silence locations. Bakis et al. (1997) proposed to build different models, e.g. Gaussian mixture models, for a fixed set of acoustic classes, such as band-limited telephone speech, music/noise corrupted speech, pure music, speech, etc., from a training corpus. The Viterbi algorithm can then be used to trace a path through the trellis corresponding to the model, and to assign a class identity to contiguous sets of the acoustic feature vectors. Metric-based segmentation (Siegler et al., 1997) is proposed to segment the audio stream at maxima of the distances between neighboring windows placed at every sample. As pointed by Chen and Gopalakrishnan (1998), the decoder-guided segmentation only places boundaries at the silence locations, which in general has no direct connection with the acoustic change in the data. The model-based segmentation does not generalize to unseen acoustic conditions. The metric-based 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 which lack stability and robustness. Chen and Gopalakrishnan (1998) therefore proposed a maximum likelihood approach. 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 the BIC into the metric-based segmentation framework and designed a hierarchical algorithm for clustering of audio segments by using the BIC as a termination criterion. Based on these, we have also developed a simple but effective multi-pass approach for automatic story segmentation.

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. Our approaches for speech recognition, story segmentation and spoken document retrieval are discussed in Sections 3, 4, and 5, respectively. Finally, the prototype retrieval system is presented in Section 6 and conclusions are made in Section 7.

**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 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. Currently, 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”(Barras et al., 2001). 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. Speech Recognition

This section will introduce our speech recognition approach.

### *3.1. 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) (Furui, 1981) is applied to the training and testing speech.

### *3.2. 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 (Wang et al., 1997). 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 number of states and mixtures for the INITIAL/FINAL and silence models was set following our previous work (Huang et al., 2002). The acoustic models were trained by using a database with 16 hours of broadcast news speech collected from several radio stations located at Taipei and finally a total of 11004 mixtures were obtained. The broadcast news data was recorded using a wizard FM radio connected to a PC and digitized at a sampling rate of 16kHz with 16bit resolution. The data collection spanned the period December 1998 to July 1999.

Thus, the training database was collected in a different way and in a different time frame from that of the target TV broadcast news database. The training database is a combination of two corpora: The first corpus contained 2 hours of field report speech and 4 hours of studio anchor speech. The transcripts have been time aligned to the phrasal level. The second corpus contained 10 hours of studio anchor speech. Each audio file is a short news abstract (50 seconds on average) produced by an anchor. Unlike the first corpus, for each audio file, only the orthographic transcripts were available but detailed time alignment was unavailable.

*3.3. 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, around the same time frame as the broadcast news training corpus was collected. Thus, the newswire text corpus for language model training was also collected in a different time frame from that of the target TV broadcast news database. Word segmentation (Chen and Liu, 1992) and phonetic labeling were performed for the training text corpus based on a 61521-word lexicon for training the *N*-gram language models. This lexicon is comprised of frequent words from the CKIP lexicon (CKIP, 1993) and new words extracted from the newswire text corpus for language model training by using the unknown-word extraction tool developed by Chen and Ma (2002).

*3.4. Speech Recognition*

Our speech recognizer adopts a multi-pass search strategy. In the first pass, Viterbi search (Rabiner and Juang, 1993) 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 (Kenny et al., 1993) 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.

*3.5. Speech Recognition Experiments*

5 one-hour shows that were randomly selected from the 40-hour carefully transcribed database were used to evaluate speech recognition performance. That is, we spot-checked ~3% (5/177*100%=2.8%) of the target database. To examine performance for different conditions, the test material was divided to various subsets:

- studio anchor speech in the presence of background music,
- studio anchor speech in the presence of background noise,
- studio anchor speech in the presence of background speech,
- clean studio anchor speech,
- field reporter speech in the presence of background music,
- field reporter speech in the presence of background noise,
- field reporter speech in the presence of background speech,
- clean field reporter speech,
- interviewee speech in the presence of background music,
- interviewee speech in the presence of background noise,
- interviewee speech in the presence of background speech,
- clean interviewee speech,
- weather anchor speech in the presence of background music,
- clean weather anchor speech.

We have conducted speech recognition experiments on these subsets separately. The recognition results are summarized in Table 1. Around 85% of the studio anchor speech is clean speech, but for the field reporter speech and the interviewee speech, the percentages for the clean speech are lower. The recognition accuracy for the anchor speech in the presence of background speech or noise is very close to the accuracy for the clean anchor speech but the recognition accuracy for the anchor speech in the presence of background music is relatively poor. We checked the test material and found that the SNRs for the studio anchor speech in the presence of background music, background noise, and background speech are 12.16db, 22.06db, and 20.14db, respectively, while the SNR for the clean studio anchor speech is 36.64db. When the studio anchor was announcing the headlines, there was always some specific music played back. Moreover, for some specific topics, when the studio anchor was enumerating short news items related to that topic, a video clip corresponding to each news item was usually played back simultaneously though the volume was relatively low. As a whole, the studio anchor speech in the presence of background sound comprised a very small percentage of the studio anchor speech. Both the field reporter speech and the interviewee speech have a very similar trend compared with the studio anchor speech except that the accuracy is lower. It is obvious from Table 1 that the background music seriously degrades the recognition accuracy but the background noise and the background speech do not degrade the recognition accuracy that much. The recognition accuracy for the interviewee speech is extremely poor but the accuracy for the studio anchor speech and the field reporter speech is relatively reasonable. As a whole, the average syllable accuracy is 51.36% while the average character accuracy is 41.87%. From Table 1, it is interesting that the accuracy for the weather anchor speech is significantly worse than the accuracy for the studio anchor speech, even though both types of speech are under the same acoustic environment condition (studio speech). The out of domain language model is probably the major reason for this situation since the newswire text corpus for training the language models did not contain any weather report at all. Anyway, the weather report part is not as important as the news stories from the information retrieval point of view.

As mentioned in Section 3.2, the training radio news database was collected in a different way from that of the target TV news database. The channel mismatch between the training database and the target database could be the major reason for the poor recognition performance. It is believed that the speech recognition accuracy will be improved by augmenting or replacing the training database using samples from the target database. However, both ways need a huge investment in

transcribing a large amount of speech from the target database, which is definitely undesired. By contrast, using un-supervised model adaptation techniques such as MLLR (Leggetter and Woodland, 1995) to improve the recognition accuracy is more applicative. But, until now, we have not applied any of these in our recognizer.

## 4. Story Segmentation

Recently, we have integrated the BIC into the metric-based segmentation framework and designed a hierarchical algorithm for clustering of audio segments by using the BIC as a termination criterion. Based on these, we have also 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. The details for our story segmentation approach will be described in the following sections.

### *4.1. 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:

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

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.

### *4.2. 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 is 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 (Siegler et al., 1997). 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:

$$deltaBIC = BIC(M_1) - BIC(M_0). \tag{2}$$

We are comparing two models: One models the two windows of cepstral vectors as two multivariate Gaussians; i.e., $M_1 : x_{t-n+1}, x_{t-n+2}, ..., x_t \sim N(\mu_1, \Sigma_1); x_{t+1}, x_{t+2}, ..., x_{t+n} \sim N(\mu_2, \Sigma_2)$. The other models the data as just one multivariate Gaussian; i.e., $M_0 : x_{t-n+1}, x_{t-n+2}, ..., x_t, x_{t+1}, x_{t+2}, ..., x_{t+n} \sim N(\mu, \Sigma)$. Here, $\mu_1$, $\mu_2$, and $\mu$ are respectively the sample mean vectors of $\{x_{t-n+1}, x_{t-n+2}, ..., x_t\}$, $\{x_{t+1}, x_{t+2}, ..., x_{t+n}\}$, and $\{x_{t-n+1}, x_{t-n+2}, ..., x_t, x_{t+1}, x_{t+2}, ..., x_{t+n}\}$ while $\Sigma_1$, $\Sigma_2$, and $\Sigma$ are respectively the sample covariance matrices. According to Equation (1), we have

$$BIC(M_0) = \log L(\{x_{t-n+1}, x_{t-n+2}, ..., x_t, x_{t+1}, x_{t+2}, ..., x_{t+n}\}, N(\mu, \Sigma)) - \lambda \frac{1}{2} \#(N(\mu, \Sigma)) \times \log(2n) \tag{3}$$

and

$$\begin{aligned} BIC(M_1) = {} & \log L(\{x_{t-n+1}, x_{t-n+2}, ..., x_t\}, N(\mu_1, \Sigma_1)) + \log L(\{x_{t+1}, x_{t+2}, ..., x_{t+n}\}, N(\mu_2, \Sigma_2)) \\ & - \lambda \frac{1}{2} \#(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)) \times \log(2n), \end{aligned} \tag{4}$$

where $\#(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)) = 2\#(N(\mu, \Sigma))$. If $BIC(M_1) < BIC(M_0)$, it is more likely that $X = \{x_{t-n+1}, x_{t-n+2}, ..., x_t, x_{t+1}, x_{t+2}, ..., x_{t+n}\}$ is a homogeneous segment; i.e., there is no change point in *X*. By contrast, if *t* is a change point, we must have $deltaBIC > 0$, which means it is better to model the two windows separately. 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, the width of a peak is defined as the time span of its neighboring points with *deltaBIC* larger than 0, and the peak width threshold that was decided empirically was used to reduce false alarms.

*4.3. 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., $M_1 : x_1^i, x_2^i, ..., x_{N_i}^i \sim N(\mu_i, \Sigma_i); x_1^j, x_2^j, ..., x_{N_j}^j \sim N(\mu_j, \Sigma_j)$. 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.

*4.4. Story Segmentation Experiments*

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. To reduce false alarm, we can apply two simple heuristic criteria to further verify the anchor segments in the anchor cluster:

1. Two anchor segments must be placed at least 5 seconds apart, otherwise they will be merged as one.

2. The length of an anchor segment must be at least 5 seconds, so that the short anchor segments will be considered as a false alarm and will be removed from the anchor cluster.

Currently, threshold values for the above two criteria were set intuitively and we have not refined them precisely.

We have conducted story segmentation experiments based on the same 5 one-hour shows used for speech recognition evaluation in Section 3.5. We tabulate the experimental results in Table 2. According to the hand-segmentation, there are 112 stories in total; i.e., 112 anchor segments. After segmentation and clustering, 134 segments were first judged as anchor segments. 10 segments were too close to their previous anchor segments and the other 10 segments were too short. Finally, after being verified by the two criteria described above, there were 114 detected anchor segments. Among them, 2 segments were not produced by anchors, which means all the 112 anchor segments were detected. Therefore, the recall and precision rates are 1.0 (112/112) and 0.982 (112/114), respectively. As a result, among the 112 stories, 2 stories were divided in two to give a total of 4 stories, and our automatic story segmentation method finally resulted in 114 stories. We have also looked at the segmentation positions. Among the 112 stories whose beginning anchor segments were correctly detected, there were 103 stories whose starting time errors were within 2 seconds and 5 stories whose errors were between 2 and 3 seconds while, for the remaining 4 stories, the errors were 5.4, 6.1, 9.4, and 16.6 seconds, respectively.

We have also further examined the speaker and acoustic change detection. Figure 2 depicts a partial transcription of a broadcast news show. The transcription has three hierarchically embedded layers of segmentation (orthographic transcription, speaker turns, and sections (stories)), plus a fourth layer of segmentation (acoustic background conditions) that is independent of the other three. Some frequent situations are as follows: the non-speech part between the speech segments of two distinct speakers could be chopped into several distinct short segments according to their acoustic foreground and background conditions. Moreover, a speaker turn could be separated into several segments by short silence segments.

According to hand-segmentation of the bottom layer (the orthographic transcription level) as shown in Figure 2, there were 2013 true change points in total. Among them, 1831 change points corresponded to audio segments that were shorter than 3 seconds and most of the short segments contained pure silence, music, or noise. For these change points, it is very hard to come up with a standard way for analyzing the detection errors. Our change detection method found 966 detected change points. Both the miss rate and the false alarm rate are shown in the second row of Table 3. A change point was considered missed if there was no detected change point within a 3-second window centered on the true change point. In this way, several neighboring change points could

relate to the same detected change point. This is why only 543 true change points were considered missed such that the miss rate was 26.97% (543/2013*100%) even though there were 2013 true change points but there were only 966 detected change points. A detected change point was counted as a false alarm if there was no true change point within a 3-second window centered on the detected change point. In this way, the false alarm rate was 15.63% (151/966*100%).

Some of the change points at the acoustic background level are synchronous with the change points at the orthographic transcription level but some are not. According to hand-segmentation of both the bottom layer (the orthographic transcription level) and the top layer (the acoustic background level) as shown in Figure 2, there were 2248 change points in total; i.e., 235 (2248-2013) extra change points were obtained from the acoustic background level. As shown in the third row of Table 3, the miss rate was 32.30% (726/2248*100%). 183 (726-543) of the extra 235 change points were missed. Because the background signals were often with lower energy compared with the foreground speech, most of the extra change points were difficult to detect. As shown in the third row of Table 3, the false alarm rate was reduced to 14.49% (140/966*100%). 11 (151-140) false alarms with respect to the change points at the orthographic transcription level in fact respectively corresponded to a change point at the acoustic background level.

We have also evaluated change detection at the speaker turn level; i.e., we used the time stamps of the second layer from the bottom as shown in Figure 2 as the ground truth in this case. The results are summarized in the fourth row of Table 3. There were 739 true change points and only 70 of them corresponded to audio segments shorter than 3 seconds. The miss rate was 22.33% (165/739*100%) while the false alarm rate was 40.17% (388/966*100%). We found that for the 76 speaker change points that were considered missed, there respectively existed one detected change point located inside an extended range[1] of the non-speech between the speech segments of the two speakers. If these 76 change points were not counted, 89 speaker change points were still missed and the miss rate was 12.04% (89/739*100%). Furthermore, 248 detected change points that were judged as a false alarm in fact respectively corresponded to a change point at the orthographic transcription level or the acoustic background level. Therefore, it is reasonable that the false alarm rate was 14.49% (exactly the same as the false alarm rate evaluated at both the orthographic

[1] The extended range of a speaker change point includes the non-speech encompassing the point plus an additional 1.5 seconds either side of the non-speech.

transcription level and the acoustic background level) rather than 40.17%. As shown in the fifth row of Table 3, when the detection performance was evaluated at the speaker turn level but hand-segmentation of the orthographic transcription level and the acoustic background level was also taken into account, the miss rate and the false alarm rate were 12.04% and 14.49%, respectively.

## 5. Spoken Document Retrieval

This section will introduce our spoken document retrieval approaches.

### *5.1. Indexing Terms*

In Chen et al. (2002), 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 3.4, 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.

### *5.2. Information Retrieval Model*

Vector space models widely used in 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 $d$,

$$\vec{d}_j = (x_{j1}, x_{j2}, ..., x_{jt}, ..., x_{jM_j}), \quad j = 1,2,3,...,8, \tag{5}$$

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

$$x_{jt} = 1 + \ln(\sum_{i=1}^{n_t} c_t(i)). \tag{6}$$

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 $x_{jt}$ in Equation (6) is set to zero if the specific indexing term $t$ did not 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 $d$. If $\sum_{i=1}^{n_t} c_t(i) < 1$, then $x_{jt} = \sum_{i=1}^{n_t} c_t(i)$. As mentioned in Section 3.4, 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:

$$c(s) = \frac{2}{1 + \exp(-\alpha \times [\log p(O \mid s) - \log p(O \mid s^*)])}, \tag{7}$$

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 $\alpha$ is used to control the slope of the Sigmoid function. From Equation (7), 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 a similar way as the documents. First, the corresponding word, character, and syllable sequences are derived from a query using the same word segmentation method as in Section 3.3. Then, eight types of indexing terms are used to construct eight feature vectors for a query $q$,

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

where $\vec{q}_j$ is the feature vector for the *j*-th type of indexing terms, and the value of $x_{jt}$ is obtained by

$$x_{jt} = [1 + \ln(n_t)] \cdot \ln(\frac{N+1}{N_t}). \tag{9}$$

Here, we use $[1 + \ln(n_t)]$ instead of $[1 + \ln(\sum_{i=1}^{n_t} c_t(i))]$ since $c_t(i) = 1$ for all kinds of indexing terms. The value of $\ln(\frac{N+1}{N_t})$ 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 Cosine measure is used to estimate the query-document relevance for the *j*-th type of indexing terms:

$$R_j(\vec{q}_j, \vec{d}_j) = (\vec{q}_j \cdot \vec{d}_j) \big/ (\|\vec{q}_j\| \times \|\vec{d}_j\|). \tag{10}$$

The overall relevance measure is then the weighted sum of the relevance measures of all types of indexing terms:

$$R(\vec{q}, \vec{d}) = \sum_j w_j \times R_j(\vec{q}_j, \vec{d}_j), \tag{11}$$

where $w_j$ is a weighting parameter obtained empirically.

## 6. The SoVideo Broadcast News Retrieval System

We have successfully implemented a web-based Mandarin Chinese broadcast news retrieval system, SoVideo, by integrating the above speech recognition, story segmentation, and spoken document retrieval approaches. As depicted in Figure 3, 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.

We have implemented Equation (11) in a hierarchical way following our previous work (Chen et al. 2002). The overall relevance measure is the weighted sum of the relevance measures for syllable-, character-, and word-level indexing terms, which are respectively the weighted sum of the relevance measures for the corresponding index terms such as unigrams, bigrams, etc. The weighting parameters for syllable-, character-, and word-level indexing terms are in the ratio 1:0.3:0.5 while the weighting parameters for unigram, bigram, and trigram are in the ratio 0.1:0.7:0.3.

*6.1 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. Whether a retrieved document is relevant or not was judged by one of the authors who conducted the experiments. 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 (12):

$$\text{mAP} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{N_i}\sum_{k=1}^{N_i}\frac{k}{rank_{ik}} \tag{12}$$

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. The detailed retrieval performance is plotted in Figure 4. The raw average precision (rAP) obtained by directly averaging the precision for each query at a fixed number of retrieved documents is also depicted for reference. The rAP is calculated using Equation (13):

$$\text{rAP} = \frac{1}{L}\sum_{i=1}^{L}\frac{N_i}{N} \tag{13}$$

where $L$ is the total number of queries, $N_i$ is the total number of relevant documents contained in the $N$ retrieved documents for query $q_i$. The rAPs obtained in this way are 0.975, 0.835, 0.69, 0.577, and 0.501, respectively, when 1, 5, 10, 15, and 20 returned documents were considered. The relatively low average precisions when the number of returned documents increased are obviously because the current target database only contains 3264 stories and some queries only have a few relevant documents. When we looked at the returned documents, we found that for some queries only the top few documents are relevant while the rest are all irrelevant. The high mAPs indicate that the retrieved relevant documents tend to gather in the top group of documents retrieved. 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 or query-by-exemplar techniques to enhance retrieval performance.

**7. 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. In this paper, we presented our speech recognition, story segmentation, and information retrieval approaches and reported on the evaluation of each component and the retrieval system. Though the speech recognition accuracy is not very good, the retrieval performance is quite impressive. Presently, the retrieval system is definitely far away from

practical use because of the small database. However, the target database is continually growing. In addition, our ongoing research is focused on improving the usability of the retrieval system in various ways, mainly by reducing recognition errors using acoustic model and language model adaptation, improving retrieval performance by using relevance feedback or query-by-exemplar techniques, improving story segmentation by considering both acoustic and linguistic clues, and reducing indexing errors by using language identification techniques.

**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.

**References**

Bakis, R., Chen, S., Gopalakrishnan, P., Gopinath, R., Maes, S., Polymenakos, L., and Franz, M. (1997). Transcription of broadcast news shows with the IBM large vocabulary speech recognition system. *Proceeding of 1997 DARPA Speech Recognition Workshop*.

Barras, C., Geoffrois, E., Wu, Z. B., and Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33: 5-22.

Chen, B., Wang, H. M., and Lee, L. S. (2002). Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese. *IEEE Transactions on Speech and Audio Processing*, 10(5): 303-314.

Chen, K. J. and Liu, S. H. (1992). Word identification for Mandarin Chinese sentences. *COLING1992 Proceedings*, pp. 101-107.

Chen, K. J. and Ma, W. Y. (2002). Unknown word extraction for Chinese documents. *COLING2002 Proceedings.*

Chen, S. and Gopalakrishnan, P. S. (1998). Speaker, environment and channel change detection and clustering via the Bayesian information criterion. *Proceedings of 1998 DARPA Broadcast News Transcription and Understanding Workshop*.

CKIP group. (1993). Analysis of syntactic categories for Chinese. *CKIP Technical Report*, No. 93-05, Institute of Information Science, Academia Sinica, Taipei.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics Speech and Signal Processing*, 29: 254-272.

Hauptmann, A., Thornton, S., Houghton, R., Qi, Y., Ng, D., Papernick, N., and Jin, R. (2001). Video retrieval with the Informedia digital video library system. *Proceedings of The Tenth Text REtrieval Conference*.

Huang, M. F., Chen, K. T., and Wang, H. M. (2002). Towards retrieval of video archives based on the speech content. *Proceedings of International Symposium on Chinese Spoken Language Processing*.

Jones, K. S., Jones, G. J. F., Foote, J. T., and Young, S. J. (1996). Experiments on Spoken Document Retrieval. *Information Processing & Management*, 32(4): 399-417.

Kenny, P., Hollan, R., Gupta, V. N., Lennig, M., Mermelstein, P., and O'Shaughnessy, D. (1993). A*-admissible heuristics for rapid lexical access. *IEEE Transactions on Speech and Audio Processing*, 1(1): 49-58.

Kubala F., Jin, H., Matsoukas, S., Nguyen, L., Schwartz, R., and Makhoul, J. (1997). The 1996 BBN Byblos Hub-4 transcription system. *Proceeding of 1997 DARPA Speech Recognition Workshop*.

Lee, L. S. (1997). Voice dictation of Mandarin Chinese. *IEEE Signal Processing Magazine*, 14(4): 63-101.

Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models. *Computer, Speech and Language*, 9: 171-185.

Logan, B. Moreno, P., van Thong, J. M., and Whittaker, E. (2000). An experimental study of an audio indexing system for the Web. *ICSLP2000 Proceedings*.

Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R., and Srivastava, A. (2000). Speech and language techniques for audio indexing and retrieval. *Proceedings of the IEEE*, 88(8): 1338-1353.

Meng, H., Lo, W. K., Li, Y. C., and Ching, P. C. (2000a). Multi-scale audio indexing for Chinese spoken document retrieval. *ICSLP2000 Proceedings*.

Meng, H. et al. (2000b). Mandarin-English Information (MEI): Investigating translingual speech retrieval. *Technical Report, The Johns Hopkins University Summer Workshop 2000,* http://www.clsp.jhu.edu/ws2000/final_reports/mei/ws00mei.pdf.

Ng, K. (2000). Subword-based approaches for spoken document retrieval. Ph.D. thesis, MIT.

Rabiner, L. R. and Juang, B. H. (1993). *Fundamentals of Speech Recognition*. NJ: Prentice-Hall.

Renals, S., Abberley, A., Kirby, D., and Robinson, T. (2000). Indexing and retrieval of broadcast news", *Speech Communication*, 32(1-2): 5-20.

Siegler, M., Jain, U., Ray, B., and Stern, R. (1997). Automatic segmentation, classification and clustering of broadcast news audio. *Proceeding of 1997 DARPA Speech Recognition Workshop*.

Wactlar, H. D., Kanade, T., Smith, M. A., and Stevens, S. M. (1996)**.** Intelligent access to digital video: Informedia project. *IEEE Computer*, 29(5): 46-52.

Wang, H. M. et al. (1997). Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data. *IEEE Trans. on Speech and Audio Processing*, 5(2): 195-200.

Wayne, C. L. (2000). Multilingual topic detection and tracking: successful research enabled by corpora and evaluation. *LREC2000 Proceedings*.

Wechsler, M., (1998). Spoken document retrieval based on phoneme recognition. Ph.D. thesis, Swiss Federal Institute of Technology (ETH).

| | | Length in seconds | Syllable accuracy | Character accuracy |
|---|---|---|---|---|
| Studio anchor speech | In background music | 220.4 | 49.63% | 40.06% |
| | In background noise | 105.8 | 69.63% | 60.30% |
| | In background speech | 55.5 | 69.92% | 65.73% |
| | Clean speech | 2203.2 | 69.87% | 63.79% |
| | Overall estimate | 2584.9 | 69.86% | 63.67% |
| Field reporter speech | In background music | 850.7 | 43.89% | 32.35% |
| | In background noise | 2726.3 | 54.85% | 44.12% |
| | In background speech | 609.0 | 63.03% | 55.13% |
| | Clean speech | 2323.6 | 62.78% | 54.71% |
| | Overall estimate | 6509.6 | 57.34% | 47.67% |
| Interviewee speech | In background music | 52.2 | 15.79% | 9.79% |
| | In background noise | 1517.2 | 23.71% | 12.86% |
| | In background speech | 98.5 | 23.58% | 8.19% |
| | Clean speech | 2386.6 | 30.48% | 18.50% |
| | Overall estimate | 4054.5 | 26.54% | 15.68% |
| Weather anchor speech | In background music | 240.6 | 41.97% | 32.76% |
| | Clean speech | 357.9 | 44.51% | 36.37% |
| | Overall estimate | 598.5 | 43.40% | 34.91% |

Table 1. The syllable and character recognition accuracy for the broadcast news speech.

| | | Show 1 | Show 2 | Show 3 | Show 4 | Show 5 | Overall |
|---|---|---|---|---|---|---|---|
| # true stories (anchor segments) | | 23 | 23 | 22 | 20 | 24 | 112 |
| # detected anchor segments after clustering | | 27 | 28 | 24 | 25 | 30 | 134 |
| # detected anchor segments being merged (too close to their previous anchor segments) | | 3 | 1 | 0 | 3 | 3 | 10 |
| # detected anchor segments being ignored (too short) | | 1 | 4 | 1 | 2 | 2 | 10 |
| # finally detected stories (anchor segments) | | 23 | 23 | 23 | 20 | 25 | 114 |
| # inserted stories (false alarm) | | 0 | 0 | 1 | 0 | 1 | 2 |
| # deleted stories (miss) | | 0 | 0 | 0 | 0 | 0 | 0 |
| Story starting time error | <=2 seconds | 21 | 20 | 21 | 19 | 22 | 103 |
| | 2-3 seconds | 1 | 2 | 1 | 0 | 1 | 5 |
| | >3 seconds | 1 | 1 | 0 | 1 | 1 | 4 |

Table 2. The story segmentation results.

| | # change points | Miss rate | False alarm rate |
|---|---|---|---|
| The orthographic transcription level | 2013 | 26.97% | 15.63% |
| Both the orthographic transcription level and the acoustic background level | 2248 | 32.30% | 14.49% |
| The speaker turn level | 739 | 22.33% | 40.17% |
| The speaker turn level* | 739 | 12.04% | 14.49% |

Table 3. The change detection results. For the "The speaker turn level*" case, the miss rate and the false alarm rate were obtained by also considering hand-segmentation of the orthographic transcription level and the acoustic background level

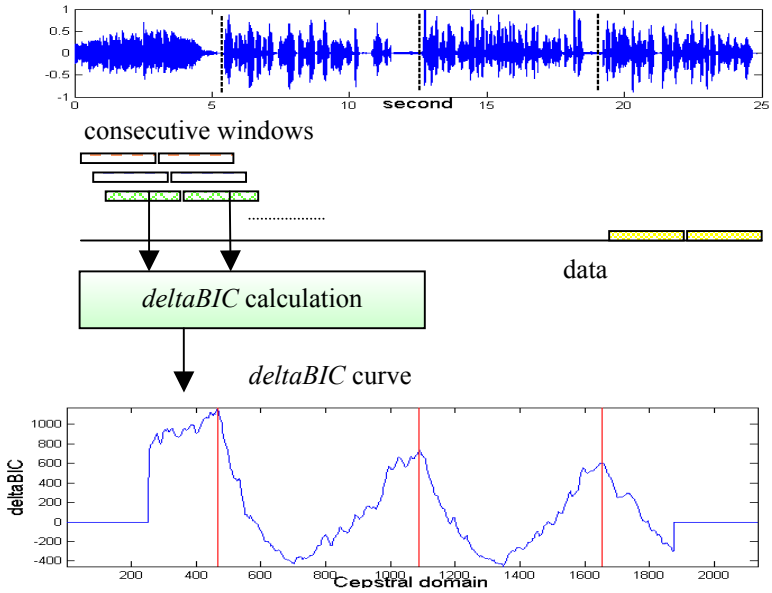


Figure 1. The procedures of metric-based segmentation via BIC

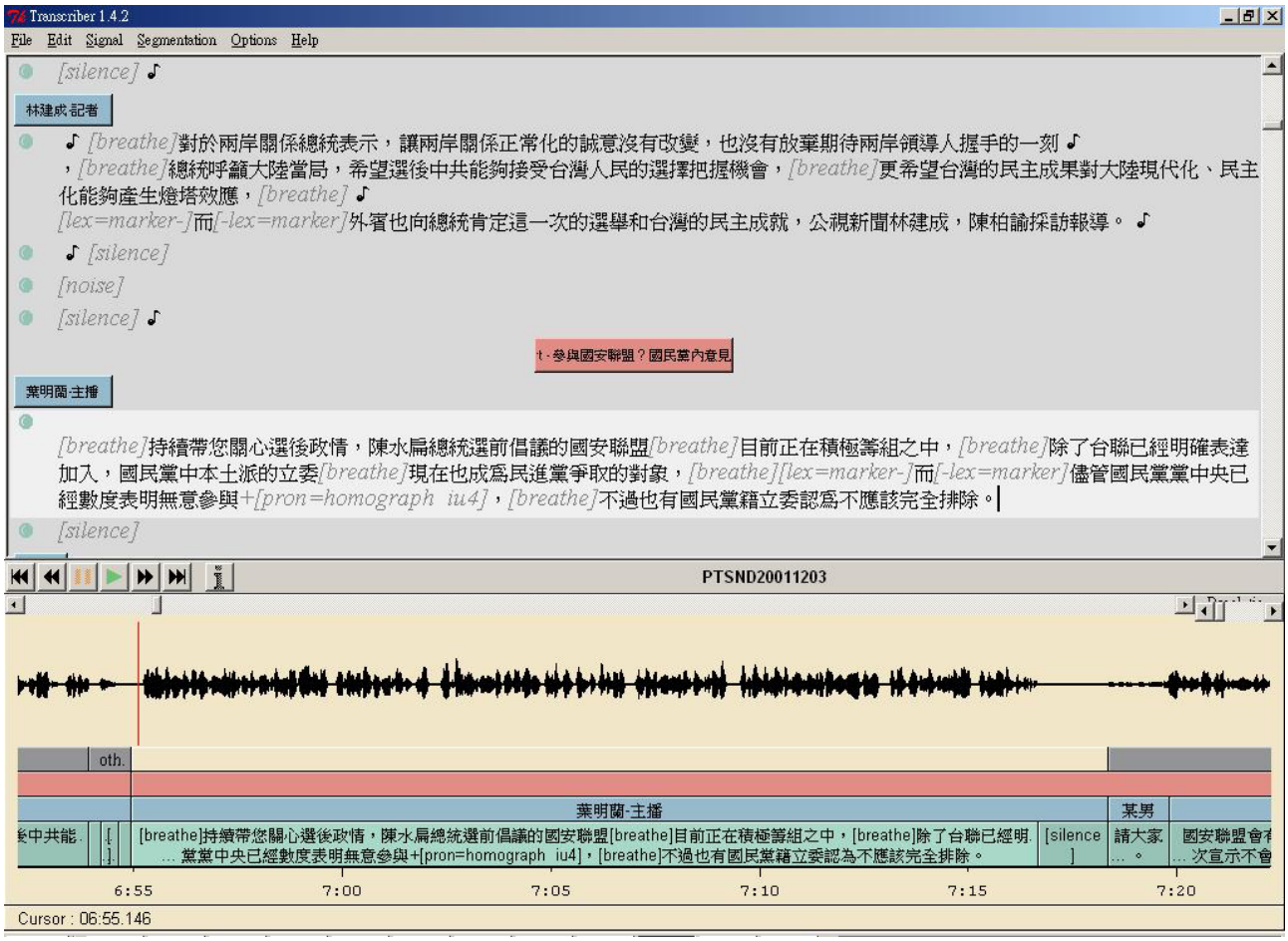


Figure 2. A partial transcription of a broadcast news show.


Figure 3. The SoVideo Web-based Mandarin Chinese broadcast news retrieval system.

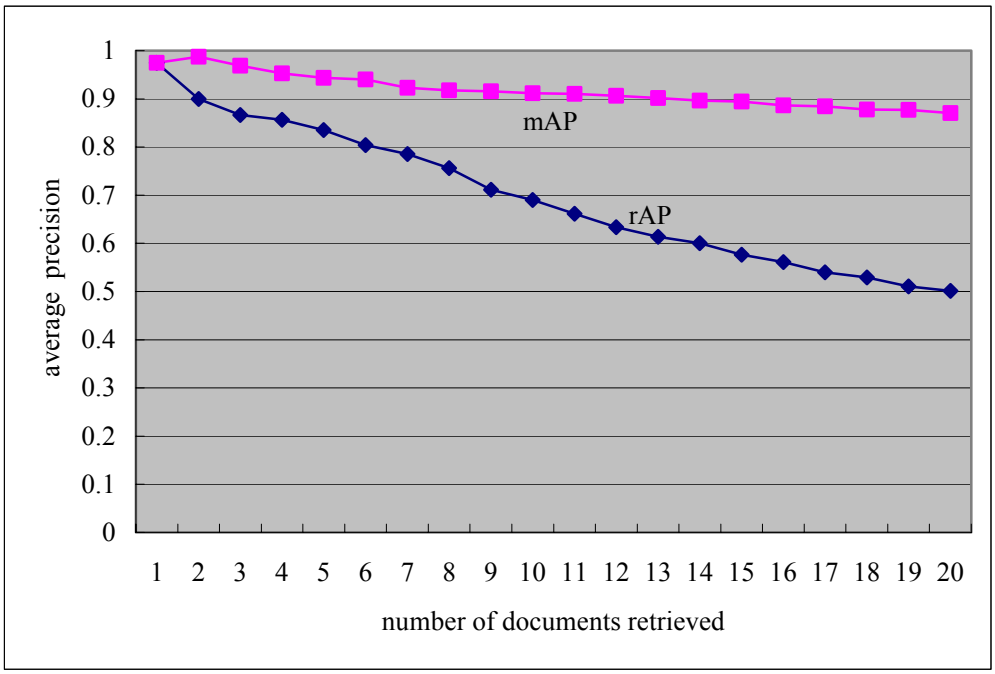


Figure 4. The retrieval performance of the SoVideo system.