International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 20, No. 1 (2006) 91–109
© World Scientific Publishing Company

World Scientific
www.worldscientific.com

VOICE RETRIEVAL OF MANDARIN BROADCAST NEWS SPEECH

BERLIN CHEN

Graduate Institute of Computer Science & Information Engineering
National Taiwan Normal University

berlin@csie.ntnu.edu.tw

This paper presents an improved framework for voice retrieval of Mandarin broadcast news speech. First, several unsupervised and data-driven approaches for broadcast news transcription were proposed to improve the speech recognition accuracy and efficiency. Then, a multiscale indexing paradigm for broadcast news retrieval was exploited to alleviate the problems caused by the speech recognition errors and the flexible wording structure of the Chinese language. Finally, we used the PDA as the platform and broadcast radio programs collected in Taiwan as the document collection to establish a speech-based multimedia information retrieval prototype system. Very encouraging results were obtained.

Keywords: Multimedia; broadcast news; speech recognition; information retrieval; multiscale indexing.

1. Introduction

Speech is the primary and the most convenient means of communication between people.[20] Due to the successful development of much smaller electronic devices and the popularity of wireless communication and networking, it is widely believed that speech will play a more active role and will serve as the major human-machine interface for the interaction between people and different kinds of smart devices in the near future. On the other hand, huge quantities of multimedia information, such as broadcast radio and television programs, voice mails, digital archives and so on, are continuously growing and filling our computers, networks and lives. It is obvious that speech is one of the most important sources of information for the great volumes of multimedia and therefore research on multimedia content processing using speech is now becoming more and more emphasized. For example, substantial efforts and very encouraging results on broadcast news transcription, retrieval and summarization have been reported in the last few years.[4,9,15]

However, in order to obtain better recognition performance, most of the transcription systems require not only large amounts of manually transcribed speech materials for acoustic training in the data preparation phase, but also much time and memory in the recognition phase. Moreover, because the subject domains and lexical regularities of the linguistic contents of news articles are very diverse and often change with time, it is extremely difficult to build well-estimated language models for speech recognition. Hence, in the recent past, several attempts have been made to investigate the possibility of achieving automatic acquisition of speech or language training data for system refinement or for rapid prototyping of a new recognition system to new domains, and very encouraging results have been obtained.[2,21,32] Quite a few studies have also explored ways to improve recognition efficiency, and many good approaches have been proposed.[1,13] On the other hand, because spoken documents are just video/audio signals accompanied with very long sequences of automatically transcribed words, for example, a 3-hour video of course lecture, a 2-hour movie, or a 1-hour news episode, and lack for organizations and structures, it is still not very easy to retrieve and identify the desired contents from the retrieved multimedia documents.[25]

With these observations in mind, in this paper we present a framework for voice retrieval of Mandarin broadcast news. Unlike Ref. 17 that focused on the delivery of multimedia information and Ref. 9 that used the spoken query to retrieve the Mandarin newswire texts, we focus here on automatic transcription and indexing of speech information, and both the query and the documents to be retrieved are in spoken form. Several unsupervised and data-driven approaches were proposed to improve the speech recognition accuracy and efficiency. Moreover, a multiscale indexing paradigm was employed to make use of the special structural properties of the Chinese language as well as to alleviate the problems caused by the speech recognition errors. All the above approaches have been successfully integrated into our speech recognition and information retrieval systems, while a prototype system for voice retrieval of Mandarin broadcast news via the PDA has also been established.

The remainder of this paper is organized as follows. In Sec. 2, we review the major constituent parts of our broadcast news system and introduce the data-driven and unsupervised approaches we present in this paper, while the multiscale indexing paradigm and the information retrieval model are described in Sec. 3. Then, the results of a series of speech recognition and information retrieval experiments are discussed in Sec. 4. The prototype system for voice retrieval of Mandarin broadcast news speech via the PDA is explained in Sec. 5. Finally, we conclude in Sec. 6.

2. The Broadcast News Transcription System

We will briefly review the major constituent components of the broadcast news transcription system developed at National Taiwan Normal University (NTNU) and describe several unsupervised and data-driven approaches to be integrated into the transcription system.[7] The overall framework is depicted in Fig. 1.

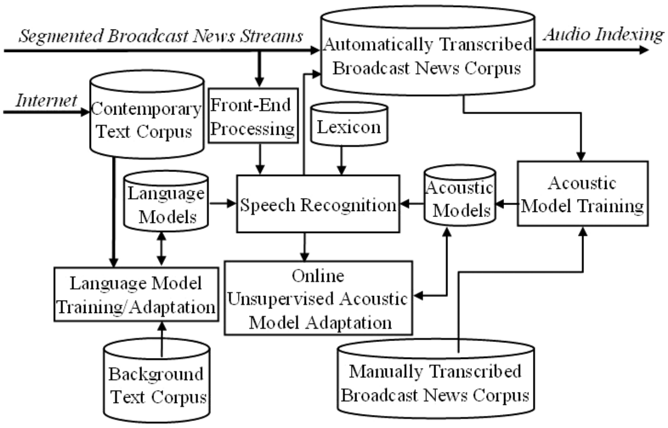

Segmented Broadcast News Streams
Automatically Transcribed Broadcast News Corpus
Audio Indexing
Internet
Contemporary Text Corpus
Front-End Processing
Lexicon
Language Models
Speech Recognition
Acoustic Models
Acoustic Model Training
Language Model Training/Adaptation
Online Unsupervised Acoustic Model Adaptation
Background Text Corpus
Manually Transcribed Broadcast News Corpus

Fig. 1. The overall framework of the NTNU broadcast news transcription system.

2.1. *Front-end processing*

The front-end processing is conducted with two feature extraction approaches: the conventional MFCC (Mel-frequency Cepstral Coefficients)-based approach and the data-driven LDA (Linear Discriminant Analysis)-approaches. In the case of the MFCC-based approach, 13-dimensional cepstral coefficients derived from 18 filter bank outputs are incorporated along with their first- and second-order time derivatives, which totally span nine successive speech frames. As for the LDA-based approach, the states of each HMM (Hidden Markov Model) are taken as the units for class assignment. Either the outputs of filter banks or the cepstral coefficients are chosen as the basic vectors. The basic vectors from every nine successive speech frames are spliced together to form supervectors for constructing the LDA transformation matrix, which is then used to project the supervectors to a lower feature space. The dimension of the resultant vectors is set to 39, which is just the same as that used in the MFCC-based approach. Finally, in both the MFCC- and LDA-based feature extraction approaches, utterance-based cepstral mean subtraction and variance normalization are applied.

2.2. *Speech corpus and acoustic modeling*

The speech data set consists of about 112 hours of FM radio broadcast news, which was collected from several radio stations located in Taipei during 1998–2002 using a wizard FM radio connected to a PC and digitized at a sampling rate of 16 kHz with 16-bit resolution.[9] All the speech materials were manually segmented into separate stories, each of which is a news abstract spoken by one anchor speaker. Some of these stories contain background noise and music. For 7.7 hours of speech data, we have corresponding orthographic transcripts. About 4.0 hours of this data collected from 1998 to 1999 was used to bootstrap the acoustic training, and the other 3.7 hours of data collected in September 2002 was used for testing. The remaining 104.3 hours of untranscribed speech data was reserved for unsupervised acoustic model training, which will be described in more detail in Sec. 4. The acoustic models chosen for speech recognition were 112 right-context-dependent INITIAL's and 38 context-independent FINAL's. They were selected based on consideration of the phonetic structure of Mandarin syllables.[9] Here, INITIAL means the initial consonant of a syllable and FINAL is the vowel (or diphthong) part but also includes an optional medial or nasal ending. Each INITIAL is represented by an HMM with three states, while each FINAL is represented with four states. The Gaussian mixture number per state ranges from 2 to 128, depending on the quantity of training data. In all the experiments, gender-independent models were used.

2.3. *Lexicon, text corpus and language modeling*

In the Chinese language, each character (at least 7000 characters are commonly used) is pronounced as a monosyllable and is a morpheme with its own meaning. New words are very easily generated by combining a few characters but nevertheless are tokenized into several single-character words or words with fewer characters when the text corpus is processed for language model training. This definitely makes the out-of-vocabulary problem especially serious in the case of Mandarin broadcast news transcription. In order to alleviate the degradation of speech recognition accuracy caused by the out-of-vocabulary problem, compound words must be carefully selected and added to the lexicon according to their statistical properties in the corpus. Hence, we explored the use of the geometrical average of the forward and backward bigrams of any word pair (w_i, w_j) occurring in the corpus for compound word selection[28,31]:

$$FB(w_i, w_j) = \sqrt{P_f(w_j \mid w_i) P_b(w_i \mid w_j)}, \tag{1}$$

where

$$P_f(w_j \mid w_i) = \frac{P(w_{t+1} = w_j, w_t = w_i)}{P(w_t = w_i)} \quad \text{and} \tag{2}$$

$$P_b(w_i \mid w_j) = \frac{P(w_{t+1} = w_j, w_t = w_i)}{P(w_{t+1} = w_j)}. \tag{3}$$

We started with a lexicon composed of 67K words and iteratively used the above measures with varying thresholds to find all possible word pairs which could be merged together. Eventually, a set of about 5K compound words was added to the lexicon to form a new lexicon of 72K words. The n-gram language modeling approach was adopted in the study; thus, the background language models consisted of word-based trigram and bigram models, which were estimated using a text corpus consisting of 170 million Chinese characters collected from Central News Agency (CNA) in 2000 and 2001 (the Chinese Gigaword Corpus released by LDC). On the other hand, a corpus consisting of 50 million Chinese characters in newswire texts collected from the Internet from August to October 2002 was used as a contemporary corpus for language model adaptation. The language models were trained with Kneser–Ney backoff smoothing[22] using the SRI Language Modeling Toolkit (SRILM).[30]

2.4. *Speech recognition*

Our baseline recognizer was implemented with left-to-right frame-synchronous tree search as well as lexical prefix tree organization of the lexicon.[1,4] Each tree arc (or phonetic arc) in the lexical tree corresponded to the HMM for an INITIAL or FINAL in Mandarin Chinese, and each tree leaf denoted a word boundary for words sharing the same pronunciation. At each speech frame, the so-called word-conditioned method was used to group the path hypotheses that shared the same history of predecessor words (or more precisely, the same search history of n-1 predecessor words for n-gram language modeling) into identical copies of the lexical tree, and they were then expanded and recombined according to the tree structure until a possible next word ending was reached. At word boundaries, the path hypotheses among the tree copies that had equivalent search histories (the same last $n - 1$ words) were recombined and then propagated into the existing tree copies or used to start new ones if none existed. Note that these tree copies were built according to a conceptual view. During the search process, only one lexical tree structure was built for reference purposes, and all path hypotheses were stored in a list structure instead. These path hypotheses were accessed by means of four-dimensional coordinates, each of which represented the history of $n - 1$ predecessor words, the tree arc in the lexical tree, the HMM state, and the speech frame, respectively. At each speech frame, a beam pruning technique, which considered the decoding scores of path hypotheses together with their corresponding language model look-ahead scores, was used to select the most promising path hypotheses. Language model look-ahead was adopted because the search structure was implemented with a lexical prefix tree and the current word identity of a particular path hypothesis could not be determined until it reached a tree leaf. In addition, language model look-ahead has the merit of early application of language model constraints, which can help guide the search process. In this research, unigram language model look-ahead was adopted. The unigram language model look-ahead score for a tree arc was defined as the maximum unigram probability over all the words that could be reached via this specific arc, which could be easily calculated and stored beforehand. Moreover, if the word hypotheses ending at each speech frame had scores that were higher than the predefined threshold, their associated decoding information, such as the word start and end speech frames, the identities of current and predecessor words, and the acoustic score, were kept in order to build a word graph for further language model rescoring.[27] Once the word graph had been built, forward-backward search with a more sophisticated language model was conducted to generate the most likely word sequence. In this study, the bigram language model was used in the tree search procedure, while the trigram language model was used in the word graph rescoring procedure.

2.5. *Acoustic look-ahead using Mandarin syllable-level heuristics*

In a baseline recognizer, language model look-ahead and beam pruning techniques can be incorporated together to help retain the most promising path hypotheses for further expansion. However, the crucial problem with such an approach is that it does not consider the potential likelihood of the unexplored portion of a speech utterance when beam pruning is applied. Thus, many unpromising path hypotheses and ambiguities will unavoidably be included during the search process. Therefore, the search efficiency may be degraded, since a large number of path hypotheses will have to be examined at each speech frame. On the other hand, the Chinese language is well known for its monosyllabic structure, in which each Chinese word is composed of one or more syllables (or characters); thus, syllables are the very important constituent units of Chinese words.[9,24] In addition, Mandarin Chinese is phonologically compact; an inventory of about 400 base syllables provides full phonological coverage of Mandarin audio data if the tonal information is further ignored. This implies that syllable recognition will be much faster than word recognition. Thus, in this study, we utilized syllable-level heuristics to enhance search efficiency. A compact syllable lattice based on the structural information of words in the lexicon was automatically built and used to estimate the likelihood of the unexplored portion of a speech utterance. Each HMM state in the syllable lattice could be easily related to its corresponding HMM states in the lexical tree, and the relation between them was a one-to-many mapping. In the first pass, the syllable lattice was calculated in a right-to-left time-synchronous manner, and at each speech frame, the acoustic scores for the HMM states in the lattice were stored and taken as the likelihood estimation for acoustic look-ahead. In the second pass, frame-synchronous tree search was performed by incorporating the language model look-ahead scores together with the acoustic look-ahead scores for beam pruning. Though speech recognition was carried out in a two-pass mode, the time spent on calculating acoustic look-ahead scores was almost negligible. The word graph rescoring procedure also could be applied after the second-pass search.

2.6. *Unsupervised acoustic model training*

The purpose of acoustic modeling is to provide a method to calculate the likelihood of a speech utterance occurring given a word sequence. In principle, a word sequence can be decomposed into a sequence of phone-like (subword, or INITIAL or FINAL in Mandarin Chinese) units, each of which is represented by an HMM, and the corresponding model parameters can be efficiently estimated from a corpus of orthographically transcribed training utterances using the Expectation-Maximum (EM) algorithm.[12] Accordingly, in order to obtain acceptable performance in speech recognition, large amounts of manually transcribed speech data are inevitably required, especially when porting the system to new application domains. However, generating manually transcribed data is an expensive process in terms of both manpower and time. Based on this observation, we investigated here the unsupervised acoustic model training approach for Mandarin broadcast news recognition. Unlike the previous approaches,[23,26] which aligned closed-captions with automatic transcripts and kept only portions that agreed for acoustic model training, in this study, we developed a verification-based method for automatic acoustic training data acquisition. The prototype system, initially trained with only four hours of manually transcribed speech corpus, was used to recognize the remaining more than one hundred hours of unannotated speech corpus, as described previously in Sec. 2.2. For each candidate word segment generated by the forward-backward search in the word graph rescoring procedure, its associated word-level posterior probability as well as subword-level acoustic verification score, or more specifically, sub-syllable-level verification score, were incorporated together.[7] The word-level posterior probability and subword-level acoustic verification score were normalized in the range of 0 to 1 and were equally weighted to form the word confidence measure, which was then used to locate the most probably correct words. By varying the word confidence thresholds, different amounts of automatically transcribed data were accordingly selected and used in combination with the original four-hour manually transcribed corpus to retrain different sets of acoustic models. The LDA transformation matrix employed in the feature extraction process needed to be reestimated, and the acoustic features were recalculated as well, according to the speech data selected for training.

2.7. *Unsupervised language model adaptation*

Statistical language modeling, which aims to capture regularities in human natural language and quantify the acceptance of a given word sequence, has been a focus of active research in speech and language processing over the past two decades. The n-gram modeling (especially the bigram and trigram modeling) approach, which determines the probability of a word given the previous $n-1$ word history, has been widely used.[3] The n-gram probabilities are usually computed based on the maximum likelihood (ML) principle. In order to tackle the inevitable data sparseness problems that occur when estimating the n-gram probabilities from a specific text corpus, a variety of smoothing or interpolation techniques have been proposed in the past several years.[11] However, for complicated speech recognition tasks, such as broadcast news transcription, it is still extremely difficult to build well-estimated language models because the subject domains and lexical characteristics of the linguistic contents of news articles are very diverse and often change with time. Various approaches have been applied to adapt language models by making use of either the contemporary corpus[14] or the recognition hypotheses cached so far.[19] Two of the most widely-used approaches to language model adaptation are count merging and model interpolation, which can be viewed as maximum *a posteriori* (MAP) language model adaptation with different parameterizations of the prior distribution and can be easily integrated into the n-gram language modeling framework to capture the local regularities of word usage in the new task domain. The adaptation formulae (e.g. for trigram modeling) for count merging and model interpolation can be, respectively, written as:

$$\hat{P}_{\text{Adapt}-1}(w_i \mid w_{i-2}w_{i-1}) = \frac{\alpha \cdot C_{d,\text{Cont}}(w_{i-2}w_{i-1}w_i) + \beta \cdot C_{d,\text{Back}}(w_{i-2}w_{i-1}w_i)}{\alpha \cdot C_{\text{Cont}}(w_{i-2}w_{i-1}) + \beta \cdot C_{\text{Back}}(w_{i-2}w_{i-1})}, \tag{4}$$

and

$$\begin{aligned}\hat{P}_{\text{Adapt}-2}(w_i \mid w_{i-2}w_{i-1}) &= \gamma \cdot P_{\text{Cont}}(w_i \mid w_{i-2}w_{i-1}) \\ &\quad + (1-\gamma) \cdot P_{\text{Back}}(w_i \mid w_{i-2}w_{i-1}).\end{aligned} \tag{5}$$

For the count merging formula in Eq. (4), $C_{d,\text{Cont}}(w_{i-2}w_{i-1}w_i)$ and $C_{d,\text{Back}}(w_{i-2}w_{i-1}w_i)$ are, respectively, the discounted trigram counts[5] accumulated from the contemporary and background text corpora, $C_{\text{Cont}}(w_{i-2}w_{i-1})$ and $C_{\text{Back}}(w_{i-2}w_{i-1})$ are, respectively, the bigram counts accumulated from the contemporary and background text corpora, and α and β are tunable weighting parameters. For the model interpolation formula in Eq. (5), $P_{\text{Cont}}(w_i \mid w_{i-2}w_{i-1})$ and $P_{\text{Back}}(w_i \mid w_{i-2}w_{i-1})$ are the trigram probabilities, respectively, estimated from the contemporary and background text corpora, and γ is a tunable weighting parameter. A more detailed derivation of Eqs. (4) and (5) also can be found in Ref. 2. In this study, we investigated the use of the above two language model adaptation approaches for Mandarin broadcast news transcription. As mentioned earlier, a corpus of contemporary Internet newswire texts collected from August to October 2002 was used for additional prediction for the linguistic events of the testing broadcast news stories collected in September 2002.

3. The Information Retrieval System

The information retrieval system is implemented in a client-server architecture, in which the broadcast news indexing and retrieval are performed at the server side and the query is posed at the client side. The considerations of the structural properties of the Chinese language, the indexing mechanism and the information retrieval model used in this paper are explained as follows.

3.1. *Considerations of the structural properties of the Chinese language*

There is an unknown number of words in Mandarin Chinese, although only some (e.g. 80 thousand, depending on the domain) are commonly used. Each word is composed of one or more characters, and 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 the characters "電 (electricity)" and "腦 (brain)" yields the word "電腦 (computer)" while the combination of "火 (fire)" and "山 (mountain)" yields the word "火山 (volcano)". As mentioned earlier, Mandarin Chinese is phonologically compact; an inventory of about 400 base syllables provides full phonological coverage of Mandarin audio, if the differences in tones are disregarded. On the other hand, an inventory of about 7000 characters provides full textual coverage of written Chinese. There is a many-to-many mapping between characters and syllables. Consequently, a foreign word can be translated into different Chinese words based on its pronunciation. For example, Kosovo may be translated as "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 as "蓋達 /gai4-da2/", "凱達 /kai3-da2/", "卡達 /ka3-da2/", "卡伊達 /ka3-i1-da2/", "阿爾蓋達 /a1-er3-gai4-da2/".[10] 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 when performing Mandarin Chinese speech recognition; e.g. Lee[24] indicated that syllable recognition is an important problem. Mandarin Chinese speech recognition evaluation is usually based on syllable and character accuracy, rather than word accuracy. The characteristics of the Chinese language also lead to some special considerations for the spoken document retrieval task.[9] Word-level indexing features possess more semantic information than subword-level features; hence, word-based retrieval enhances precision. On the other hand, subword-level indexing features behave more robustly against the Chinese word tokenization ambiguity, homophone ambiguity, open vocabulary problem, and speech recognition errors; hence, subword-based retrieval enhances recall. Accordingly, there is good reason to fuse the information obtained from indexing the features of different levels. It has been shown[9] that syllable level indexing features are very effective for Mandarin Chinese spoken document retrieval; retrieval performance can be improved further by integrating information from character-level and word-level indexing features.

3.2. *Indexing mechanism*

A retrieval index was generated over the entire collection of indexed broadcast news documents. Both the word-based and syllable(subword)-based indexing approaches were used here to represent the broadcast news documents. Every recognized word sequence was also automatically converted into its equivalent syllable-level sequence. For the word-based indexing approach, single words are taken as the index terms, while for the syllable-based approach, both the single syllables and overlapping syllable pairs are the index terms.

3.3. *Information retrieval model*

The vector space model (VSM) widely used in many text information retrieval systems was used here for simplicity,[9] although it has been shown that the HMM model[10] or the TMM (Topical Mixture Model)[6] gave better performance for

Mandarin spoken document retrieval. In VSM, a document D can be represented by a set of feature vectors $\vec{d}_j$, each consisting of information for one type of indexing terms, such as a single word, a single syllable or a syllable pair. Each component z_{jt} of the feature vector $\vec{d}_j$ for a document D is associated with the weighted statistics of a specific indexing term t:

$$z_{jt} = (1 + \ln(c(t))) \cdot \ln(N/N_t), \tag{6}$$

where $c(t)$ denotes the occurrence count of the term t within the document D, and the natural logarithmic operation is to compress its distribution. The term weighting scheme, $1 + \ln(c(t))$, which is a variation to the conventional schemes, is used to measure the intra-document weight for the term t and its performance has been extensively studied previously.[29] 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. IDF is to measure the inter-document discriminability for the term t, reflecting that indexing terms appearing in many documents are less useful in identifying the relevant documents. A query Q is also represented by a set of feature vectors $\vec{q}_j$ constructed in the same way. The Cosine measure is used to estimate the query-document relevance for each type of indexing terms:

$$R_j(\vec{q}_j, \vec{d}_j) = (\vec{q}_j \bullet \vec{d}_j)/(\|\vec{q}_j\| \cdot \|\vec{d}_j\|). \tag{7}$$

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

$$R(Q, D) = \sum_j w_j \cdot R_j(\vec{q}_j, \vec{d}_j), \tag{8}$$

where w_j are empirically tunable weights.

4. Experimental Results

4.1. *Broadcast news transcription*

In the following subsections, a series of experiments was performed to assess recognition performance as a function of feature extraction, decoding method, as well as unsupervised acoustic and language model training.

4.1.1. *Baseline*

The baseline broadcast news system was alternatively configured using the conventional MFCC-based and data-driven LDA-based feature extraction approaches. The results are shown in Rows 2 to 4 of Table 1, where the second (MFCC) row lists the results obtained using the MFCC-based approach, and the third (LDA-1) and fourth (LDA-2) rows list, respectively, the results obtained when different sets of basic vectors were adopted during the construction of the LDA transformation matrix. In LDA-1, the cepstral coefficients are taken as the basic vector, while in

Table 1. The baseline character error rate achieved with respective to different feature extraction approaches.

	Character Error Rate (%)	
	TC	WG
MFCC	26.34	22.55
LDA-1	23.10	19.90
LDA-2	23.13	19.97
LDA-2 + Acoustic Look-ahead	23.24	20.12

LDA-2, the outputs of filter banks as the basic vector. As can be seen in Table 1, the character error rates obtained, respectively, using the two variant LDA-based approaches, after either tree search (TS) or word-graph rescoring (WG), were significantly better than those obtained using the standard MFCC-based approach. Moreover, LDA-2, which uses the filter bank outputs directly as the basic vector, was even more efficient than the MFCC-based approach due to the fact that the discrete cosine transform as well as the first- and second-order time derivative operations could be excluded from front-end processing. The LDA-2 features were, thus, chosen as the default acoustic features for the experiments described below.

4.1.2. *Experiment on acoustic look-ahead using syllable-level heuristics*

The recognition performance and efficiency, after the acoustic look-ahead technique was integrated into the system, were evaluated. These results were obtained by using the same beam pruning threshold as that previously reported in Sec. 4.1.1 and were run on an ordinary 2.6 GHz Pentium IV PC. The search efficiency results are shown in Columns 2 to 6 of Table 2, which list, respectively, the real time factors for feature extraction and HMM state emission probability calculation (FE), acoustic look-ahead (AL), tree search (TS), word-graph rescoring (WG), and the overall recognition time (Total), while the recognition accuracy results are shown in the last row of Table 1. The numbers in the parentheses in the last row of Table 2 are the relative speedups achieved compared to the results shown in the second row. Comparing the results shown in the last two rows of Table 1, it can be found that the recognition accuracy was slightly degraded (e.g. the character error rate increased from 19.97% to 20.12% after word-graph rescoring) when acoustic look-ahead was used. However, according to the results shown in Table 2, the recognition

Table 2. Recognition efficiency achieved as acoustic model look-ahead was further applied. The recognition efficiency is expressed in terms of the real time factor.

	FE	AL	TC	WG	Total
Without Acoustic Look-ahead	0.323	0.000	1.264	0.196	1.783
With Acoustic Look-ahead	0.323	0.004	0.738 (41.61%)	0.149 (23.98%)	1.214 (31.91%)

efficiency for tree search improved significantly (a relative improvement of 41.61% was obtained) while the time spent on acoustic look-ahead (0.004 real time factor) was almost negligible. In summary, the acoustic look-ahead method proposed here achieves an overall speedup of more than 31% and enables the whole system to run almost in real time.

4.1.3. *Experiment on unsupervised acoustic model training and adaptation*

Table 3 summarizes the performance of unsupervised acoustic model training. Column 2 (WG) shows the recognition results achieved using several sets of acoustic models, which were trained by selectively combining different amounts of automatically transcribed speech data with the original four-hour manually transcribed speech data. Column 1 indicates the actual sizes of the automatically transcribed speech data selected, and the numbers in parentheses are the corresponding word confidence thresholds used. In addition, the third column presents the results obtained when online unsupervised MLLR (Maximum Likelihood Linear Regression) speaker adaptation was further included.[16] It can been found from Table 3, that with careful selection of automatically transcribed speech data, the character error rate could be effectively reduced from 20.12% to 15.34% (a relative improvement of 23.76% was obtained) when a total of 21 hours of automatically transcribed data were selected for acoustic training, in combination with the original four-hour manually transcribed data. Use of the word confidence measure aided selection of the best subset of automatically transcribed data for acoustic model training. Meanwhile, use of the online unsupervised MLLR speaker adaptation technique also resulted in additional performance gains under all experimental conditions.

4.1.4. *Experiment on unsupervised language model adaptation*

The language adaptation results obtained using the contemporary text corpus are shown in Table 4. The second row shows the character error rates and perplexity for the system without language model adaptation. It can be seen that the character

Table 3. The character error rate (%) achieved with different amounts of unsupervised training data.

	Character Error Rate	
	WG	+MLLR
Original 4 Hours	20.12	18.77
+5 Hours (Thr = 0.9)	16.60	15.84
+21 Hours (Thr = 0.8)	15.34	14.71
+33 Hours (Thr = 0.7)	15.78	15.02
+48 Hours (Thr = 0.6)	15.62	14.93
+54 Hours (Thr = 0.5)	15.60	14.92
+60 Hours (Thr = 0.4)	15.49	14.84

Table 4. The character error rate (%) and perplexity achieved as the language models are adapted with contemporary text corpus using either the count merging and model interpolation strategies.

	Character Error Rate		
	WG	+MLLR	Perplexity
No LM Adaptation	15.34	14.71	670.23
Count Merging	12.89	12.17	367.18
Model Interpolation	12.49	11.91	359.26

error rates are the best ones shown in Table 3, and that the initially achieved perplexity value was 670.23. This high perplexity value was probably obtained because the local word regularity properties of the tested broadcast news stories were not modeled very well by the background language models. The third and fourth rows are respectively the results for the systems when either the count merging strategy or the model interpolation strategy was used. As can be seen, the model interpolation adaptation strategy is slightly better than the count merging one both in character error rates and in perplexity, which can significantly reduce the character rate from 14.71 to 11.91 (+MLLR) and can give a reduction of almost a half of the perplexity as well. The above results reveal that the local word regularity (or contextual) information that can be obtained from the contemporary corpus is vital for the task of Mandarin broadcast news recognition, whereas the subject domains or topical information embedded in the contemporary corpus may be worth taking into account and exploring further when performing language model adaptation.[8]

4.2. *Information retrieval*

There were totally about 18,600 broadcast news documents (more than 100 hours) used here for information retrieval. Both the word-based and syllable-based indexing approaches described previously in Sec. 3 were evaluated. On the other hand, a set of 20 simple queries, in both spoken and written forms, and their corresponding relevant news documents were manually created to support the retrieval experiments. Four speakers (two males and two females) were instructed to produce the 20 queries, respectively, over an Acer n20 PDA using the original microphone and in a quiet recording environment. To recognize the spoken queries, another read–speech database consisting of about 2.5 hours of speech produced by other 21 male and 14 female speakers over the same type of PDA was used for training the speaker-independent HMMs for automatic recognition of the spoken queries. The recognition results are presented in Table 5. It can be found that the word error rate for the spoken queries is 47.84%, while the character and syllable error rates are 37.31% and 29.88%, respectively. The results are not as good as that of broadcast news transcription reported earlier, it is mainly due to the amount of acoustic training data used for the PDA platform being much smaller than that for broadcast news transcription, and most of the test queries contain one to several

Table 5. The recognition results for the set of 20 queries spoken by two male speakers and two female speakers.

Word Error Rate (%)	Character Error Rate (%)	Syllable Error Rate (%)
47.84	37.31	29.88

Table 6. The retrieval results evaluated in terms of the mean average precision at different document cutoff values.

	Word	Syllable	Word + Syllable
Document Cutoff 10			
Text Query	0.9309	0.8885	0.9580
Spoken Query	0.5533	0.6036	0.6617
Document Cutoff 30			
Text Query	0.8838	0.8165	0.9224
Spoken Query	0.5270	0.5435	0.6465
Document Cutoff 50			
Text Query	0.8656	0.7834	0.9065
Spoken Query	0.5242	0.5212	0.6386

out-of-vocabulary (OOV) words, such as personal names and new organization or event names, which apparently occur much more frequently in the queries than in the broadcast news documents and may degrade the speech recognition performance severely. The final retrieval results are evaluated in terms of the mean average precision at different document cutoff values k, which computes the mean average precision when the top k documents have been presented to the user. The formula can be expressed as[18]:

$$mAP_{\text{cutoff}_k} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{j}{r_{i,j}}, \tag{9}$$

where m is the number of queries, n_i is the total number of documents that are relevant to query i and appear among the top k documents, and $r_{i,j}$ is the position of the jth document that is relevant to query i and appear among the top k documents, counting down from the top of the ranked list. The retrieval results are shown in Table 6. Columns 3–5 respectively show the results using the word-level indexing features, syllable-level indexing features and both are evaluated at different document cutoff values and with either text or spoken queries. As can be seen, the word-level indexing features are better than the syllable-level features for the text queries, while using both gives significant improvements over using any of them alone. Moreover, the retrieval results for the spoken queries are much worse than those of the text queries, but the combination of word-level and syllable-level features helps to reduce the performance gap between the spoken and the text queries.

5. The Prototype System

5.1. *System description*

We have implemented a prototype system that allows the user to search for Mandarin broadcast news via the PDA using a spoken natural language query. The framework of the system is shown in Fig. 2. There is a small client program on the PDA, as illustrated in Fig. 3, which transmits the speech waveform or acoustic feature data of the spoken query to the information retrieval server. The information retrieval server then passes the speech waveform or acoustic feature data to the large vocabulary continuous speech recognition (LVCSR) server, which works in a similar way as the broadcast news transcription system shown earlier in Sec. 2. The recognition result is then passed back to the information retrieval server to act as the query to generate a ranked list of relevant documents. When the retrieval

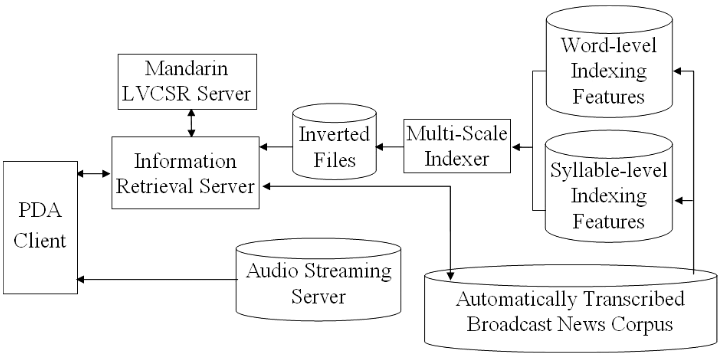
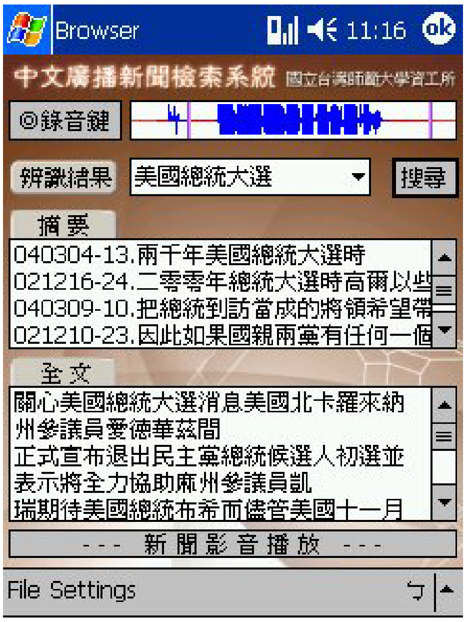

Mandarin LVCSR Server
Information Retrieval Server
PDA Client
Inverted Files
Multi-Scale Indexer
Word-level Indexing Features
Syllable-level Indexing Features
Audio Streaming Server
Automatically Transcribed Broadcast News Corpus

Fig. 2. A prototype system for voice retrieval of Mandarin broadcast news via the PDA.



辨識結果 美國總統大選

040304-13.兩千年美國總統大選時
021216-24.二零零年總統大選時高爾以些
040309-10.把總統到訪當成的將領希望帶
021210-23.因此如果國親兩黨有任何一個

關心美國總統大選消息美國北卡羅來納
州參議員愛德華茲間
正式宣布退出民主黨總統候選人初選並
表示將全力協助麻州參議員凱
瑞期待美國總統布希而儘管美國十一月

Fig. 3. The user interface of the PDA client.

results are sent back to the PDA, the user can first browse the summaries of the retrieved documents and then click to read the speech transcripts of the relevant broadcast news documents or play the corresponding audio files from the audio streaming server. The summaries were automatically generated with a dynamic programming procedure by using the word-level linguistic score, significance score, as well as confidence score.[15] On the other hand, the huge collection of broadcast news documents, as described previously in Sec. 4, is offline recognized by the broadcast news transcription system and the resultant transcripts are then utilized by the multiscale indexer to generate the word-level and syllable-level indexing terms. The final retrieval indices, including the vocabularies and document occurrences of indexing terms of different types (word- and syllable-level indexing terms), are stored as inverted files for efficient searching and comparison.

5.2. *PDA programming and multimedia streaming*

The PDA client system was programmed to run on the WinCE 4.0 Pocket PC2003 Operating System, and the development environment is the Embedded Visual C++ 4.0, which is a freeware released by the Microsoft Corporation. The client system is connected with the server side via the wireless network and following the IEEE 802.11b/g protocol. On the other hand, a streaming server runs on the Windows Server 2000 Operating System and resides at the server side, while the client system programmed with the Active Template Library (ATL) hosts the Windows CE Multimedia Player to play the streaming files. The Media Encoder, which is also a freeware released by the Microsoft Corporation, was used to process the huge collections of broadcast news recordings and to encode them into the WMA 9.0 format.

6. Conclusions

This paper presents the results of a long-term research project towards automatic recognition, retrieval and organization of Mandarin speech information. Several data-driven and unsupervised approaches to Mandarin broadcast news speech recognition and retrieval have been properly integrated into the prototype system, which allows the user to search for Mandarin broadcast news via the PDA using a spoken natural language query. Very encouraging experimental results were demonstrated.

References

1. X. L. Aubert, An overview of decoding techniques for large vocabulary continuous speech recognition, *Comput. Speech Lang.* **16** (2002) 89–114.
2. M. Bacchiani and B. Roark, Unsupervised language model adaptation, in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing I* (2003), pp. 224–227.
3. J. R. Bellegarda, Statistical language model adaptation: review and perspectives, *Speech Commun.* **42** (2004) 93–108.

4. P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, H. Ney, M. Pitz and A. Sixtus, Large vocabulary continuous speech recognition of broadcast news — the Philips/RWTH approach, *Speech Commun.* **37** (2002) 109–131.
5. E. Chang, F. Seide, H. Meng, Z. Chen, Y. Shi and Y. C. Li, A system for spoken query information retrieval on mobile devices, *IEEE Trans. Speech Aud. Process.* **10**(5) (2002) 531–541.
6. B. Chen, Exploring the use of latent topical information for statistical Chinese spoken document retrieval, *Patt. Recogn. Lett.* **27**(1) (2006) 9–18.
7. B. Chen, J. W. Kuo and W. H. Tsai, Lightly supervised and data-driven approaches to mandarin broadcast news transcription, *Int. J. Comput. Linguist. Chinese Lang. Process.* **10**(1) (March 2005) 1–18.
8. B. Chen, W. H. Tsai and J. W. Kuo, Statistical language model adaptation for Mandarin broadcast news transcription, *Proc. Int. Symp. Chinese Spoken Language Processing* (2004), pp. 313–316.
9. 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. Speech Aud. Process.* **10**(5) (2002) 303–314.
10. B. Chen, H. M. Wang and L. S. Lee, A discriminative HMM/N-gram-based retrieval approach for Mandarin spoken documents, *ACM Trans. Asian Lang. Inform. Process.* **3**(2) (2004) 128–145.
11. S. F. Chen and J. Goodman, An empirical study of smoothing techniques for language modeling, *Comput. Speech Lang.* **13** (1999) 359–394.
12. A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. B* **39**(1) (1977) 1–38.
13. G. Evermann and P. C. Woodland, Design of fast LVCSR systems, *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding* (2003), pp. 7–12.
14. M. Federico and N. Bertoldi, Broadcast news LM adaptation using cotemporary texts, *Proc. European Conf. Speech Communication and Technology 1* (2001), pp. 239–342.
15. S. Furui, T. Kikuchi, Y. Shinnaka and C. Hori, Speech-to-text and speech-to-speech summarization of spontaneous speech, *IEEE Trans. Speech Aud. Process.* **12**(4) (2004) 401–408.
16. M. J. F. Gales and P. C. Woodland, Mean and variance adaptation within the MLLR framework, *Comput. Speech Lang.* **10** (1996) 249–264.
17. D. Gibbon and L. Begeja, Multimedia processing for enhanced information delivery on mobile devices, *Proc. MobEA* (2004).
18. D. Harman, Overview of the fourth text retrieval conference (TREC-4), *Proc. Fourth Text Retrieval Conf.* (1995), pp. 1–23.
19. F. Jelinek, B. Merialdo, S. Roukos and M. Strauss, A dynamic language model for speech recognition, *Proc. Speech and Natural Language DARPA Workshop* (1991), pp. 293–295.
20. B. H. Juang and S. Furui, Automatic recognition and understanding of spoken language — a first step toward natural human-machine communication, *Proc. IEEE* **88**(8) (2000) 1142–1165.
21. T. Kemp and A. Waibel, Unsupervised training of a speech recognizer: recent experiments, *Proc. European Conf. Speech Communication and Technology 6* (1999), pp. 2725–2728.

22. R. Kneser and H. Ney, Improved backing-off for M-gram language modeling, in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing I* (1995), pp. 181–184.
23. L. Lamel, J. L. Gauvain and G. Adda, Lightly supervised and unsupervised acoustic model training, *Comput. Speech Lang.* **16**(1) (2002) 115–229.
24. L. S. Lee, Voice dictation of Mandarin Chinese, *IEEE Sign. Process. Mag.* **14**(4) (1997) 63–101.
25. L. S. Lee and B. Chen, Spoken document understanding and organization, *IEEE Sign. Process. Mag.* **22**(5) (2005) 42–60.
26. L. Nguyen and B. Xiang, Light supervision in acoustic model training, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing I* (2004), pp. 185–188.
27. S. Ortmanns, H. Ney and X. Aubert, A word graph algorithm for large vocabulary continuous speech recognition, *Comput. Speech Lang.* **11** (1997) 43–72.
28. G. Saon and M. Padmanabhan, Data-driven approach to designing compound words for continuous speech recognition, *IEEE Trans. Speech Aud. Process.* **9**(4) (2001) 327–332.
29. A. Singhal, F. Pereira, Document expansion for speech retrieval, *Proc. ACM SIGIR Conf. R&D in Information Retrieval* (1999), pp. 34–41.
30. A. Stolcke, *SRI Language Modeling Toolkit*, version 1.3.3, 2000. http://www.speech.sri.com/projects/srilm/.
31. C. J. Wang, B. Chen and L. S. Lee, Improved Chinese spoken document retrieval with hybrid modeling and data-driven indexing features, *Proc. Int. Conf. Spoken Language Processing* (2002), pp. 1985–1988.
32. F. Wessel and H. Ney, Unsupervised training of acoustic models for large vocabulary continuous speech recognition, *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding* (2001), pp. 307–310.

Berlin Chen received his B.S. and M.S. degrees in computer science and information engineering from National Chiao Tung University, Hsin Chu, Taiwan, in 1994 and 1996, respectively. He entered National Taiwan University, Taipei, in 1998 and received his Ph.D. degree in computer science and information engineering in 2001. He was with the Institute of Information Science, Academia Sinica, Taipei, from 1996–2001, and then with the Graduate Institute of Communication Engineering, National Taiwan University, from 2001–2002. In 2002, he joined the Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, where he is currently an assistant professor.

His research interests include acoustic and language modeling, search algorithms for large-vocabulary continuous speech recognition, and speech IR and organization.