Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp. 1-18
© The Association for Computational Linguistics and Chinese Language Processing

Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription

Berlin Chen[*], Jen-Wei Kuo[*] and Wen-Hung Tsai[*]

Abstract

This article investigates the use of several lightly supervised and data-driven approaches to Mandarin broadcast news transcription. With the special structural properties of the Chinese language taken into consideration, a fast acoustic look-ahead technique for estimating the unexplored part of a speech utterance is integrated into lexical tree search to improve search efficiency. This technique is used in conjunction with the conventional language model look-ahead technique. Then, a verification-based method for automatic acoustic training data acquisition is proposed to make use of large amounts of untranscribed speech data. Finally, two alternative strategies for language model adaptation are studied with the goal of achieving accurate language model estimation. With the above approaches, the overall system was found in experiments to yield an 11.88% character error rate when applied to Mandarin broadcast news collected in Taiwan.

Keywords: acoustic look-ahead, lightly supervised acoustic model training, language model adaptation, Mandarin broadcast news

1. Introduction

With the continuing growth of the amount of multimedia information accessible over the Internet, large volumes of real-world speech information, such as that in broadcast radio and television programs, digital libraries, and so on, are now being accumulated and made available to the public. Substantial efforts and very encouraging results for broadcast news transcription, retrieval, and summarization have been reported [Woodland 2002; Gauvain *et al.* 2002; Beyerlein *et al.* 2002; Chen *et al.* 2002; Chang *et al.* 2002; Meng *et al.* 2004; Furui *et al.* 2004]. 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

[*] Graduate Institute of Computer Science and Information Engineering,
National Taiwan Normal University, Taipei, Taiwan, Republic of China
E-mail: {berlin, rogerkuo, louis}@csie.ntnu.edu.tw

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 [Kemp and Waibel 1999; Wessel and Ney 2001; Macherey and Ney 2002; Bacchiani 2003]. On the other hand, quite a few studies have also explored ways to improve recognition efficiency, and many good approaches have been proposed [Schuster 2000; Aubert 2002; Evermann and Woodland 2003]. In this paper, several lightly supervised and data-driven approaches to Mandarin broadcast news transcription are presented. First, considering the special structural properties of the Chinese language, a fast acoustic look-ahead technique that employs syllable-level heuristics is integrated into lexical tree search to improve search efficiency. It is used in conjunction with the conventional language model look-ahead technique [Ortmanns and Ney 2000]. Then, a verification-based method for automatic acoustic training data acquisition is proposed to make use of large speech corpora. Finally, two alternative strategies for language model adaptation are studied with the goal of achieving accurate language model estimation.

The remainder of this paper is organized as follows. In section 2, we review the major constituents of our broadcast news system and introduce the experimental speech and language data used in this research. The acoustic look-ahead technique using syllable-level heuristics is presented in section 3, while the lightly supervised acoustic model training and language model adaptation approaches are described in sections 4 and 5, respectively. Then, the results of a series of speech recognition experiments are discussed in section 6. Finally, conclusions are drawn in section 7.

2. The NTNU Broadcast News System

The major constituent parts of the broadcast news system developed at National Taiwan Normal University (NTNU) as well as the speech and language data used in this paper will be described in this section [Chen *et al.* 2004]. Figure 1 depicts the overall framework of the broadcast news system.

2.1 Front-End Processing

Front-end processing is conducted with two feature extraction approaches: the conventional MFCC-based (Mel-frequency Cepstral Coefficients) [Davis and Mermelstein 1980] and the data-driven LDA-based (Linear Discriminant Analysis) [Duda and Hart 1973] approaches. In the case of the MFCC-based approach, 13-dimensional cepstral coefficients derived from 18

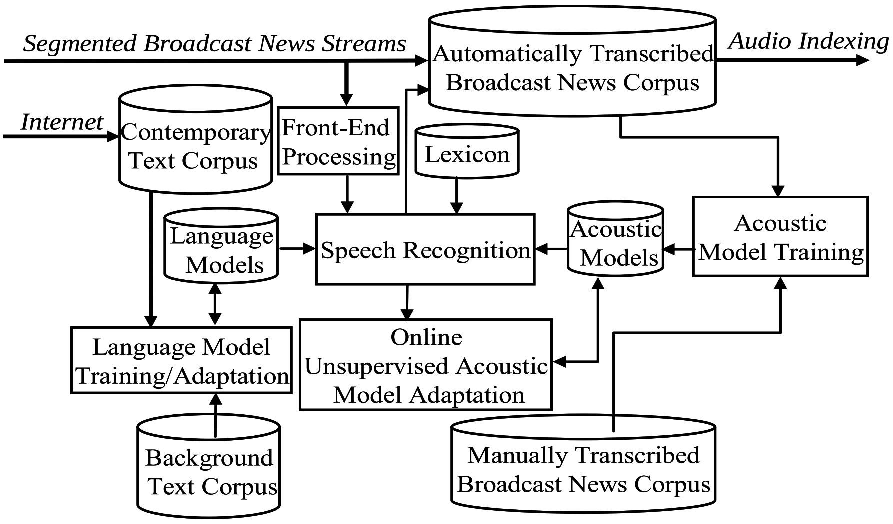
Segmented Broadcast News Streams
Automatically Transcribed Broadcast News Corpus
Audio Indexing
Internet
Contemporary Text Corpus
Front-End Processing
Lexicon
Language Models
Speech Recognition
Acoustic Models
Acoustic Model Training
Language Model Training/Adaptation
Online Unsupervised Acoustic Model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

Figure 1. The overall framework of the NTNU broadcast news system.

filter bank outputs are incorporated along with their first- and second-order time derivatives. 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 [Chen *et al.* 2002]. 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 lightly supervised acoustic model training, which will be described in more detail in section 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 [Chen *et al.* 2002]. 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 3 states, while each FINAL is represented with 4 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 7,000 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 [Saon and Padmanabhan 2001; Wang *et al.* 2002]:

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

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 [LDC 2003]). On the other hand, a corpus consisting of 50 million Chinese characters in newswire texts collected from the Internet from August to October 2002 [Chang *et al.* 2003] was used as a contemporary corpus for language model adaptation. The language models were trained with Kneser-Ney backoff smoothing [Kneser and Ney 1995] using the SRI Language Modeling Toolkit (SRILM) [Stolcke 2000].

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 [Aubert 2002; Beyerlein *et al.* 2002; Woodland 2002]. 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. Therefore, for a path hypothesis ending at speech frame t, which had a search history h and stayed at tree arc k and HMM state q, its corresponding decoding score, $D(t,h,arc_k,s_q)$, could be modified via the following equation:

$$\log\hat{D}(t,h,arc_k,s_q)=m_1\cdot\log D(t,h,arc_k,s_q)+m_2\cdot\log L_{LM}(arc_k), \tag{4}$$

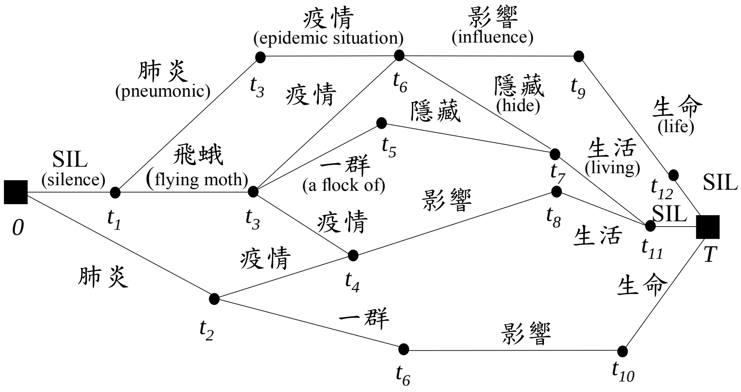

Figure 2. An illustration of a word graph, in which each arc, together with its corresponding start and end speech frames, represents a candidate word hypothesis.

where $L_{LM}(arc_k)$ is the unigram language model look-ahead score for tree arc k (notice that the HMM states within the same tree arc share the same language model look-ahead score), and $\hat{D}(t,h,arc_k,s_q)$ is the modified decoding score. m_1 and m_2 are the weighting parameters, which were set to 1 and 8, respectively, in this research. During beam pruning, we first computed the modified decoding score of the best path hypothesis at each speech frame t:

$$\log \hat{D}_{\max}(t) = \max_{h,k,q} \log \hat{D}(t,h,arc_k,s_q). \qquad (5)$$

Then, an unpromising path hypothesis was pruned if the logarithm of its modified decoding score, $\log \hat{D}(t,h,arc_k,s_q)$, was lower than a predefined threshold:

$$\log \hat{D}(t,h,arc_k,s_q) < \log \hat{D}_{\max}(t) - \log f_{Thr}, \qquad (6)$$

where f_{Thr} is an empirically set pruning factor. 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 [Ortmanns *et al.* 1997]. Once the word graph had been built, as illustrated in Figure 2, 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.

3. Acoustic Look-Ahead Using 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 [Lee 1997; Chen *et al.* 2002; Meng *et al.* 2004]. 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:

$$\log\widetilde{D}(t,h,arc_k,s_q)=m_1'\cdot\log D(t,h,arc_k,s_q)+m_2'\cdot\log L_{LM}(arc_k)+m_3'\cdot\log L_{AC}(t,arc_k,s_q), \quad (7)$$

where $L_{AC}(t,arc_k,s_q)$ is the acoustic look-ahead score, and m_1', m_2' and m_3' are the weighting parameters, which were set to 1, 8 and 1, respectively, in this research. 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.

4. Lightly Supervised 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 [Dempster *et al.* 1977]. 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 lightly supervised acoustic model training approach for Mandarin broadcast news recognition. Unlike the previous approaches [Lamel *et al.* 2002; Nguyen and Xiang 2004], which aligned closed-captions with automatic transcripts and kept only portions that agreed for acoustic training, in this study, we developed a verification-based method for automatic acoustic training data acquisition. The prototype system, initially trained with only 4 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 section 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. The word-level posterior probability of a specific word segment w in the word graph with the start and end speech frames t_s and t_e, respectively, can be defined as [Wessel *et al.* 2001]

$$P_{Post}(w_{t_s}^{t_e} \mid X_1^T) = \frac{\sum_{W_1^{t_s-1}} \sum_{W_{t_e+1}^T} p(W_1^{t_s-1} \cdot w_{t_s}^{t_e} \cdot W_{t_e+1}^T, X_1^T)}{\sum_{W_1^T} p(W_1^T, X_1^T)}, \tag{8}$$

where X_1^T is the speech utterance X which starts at speech frame 1 and ends at speech frame T, $W_1^{t_s-1}$ denotes the word sequence W which starts at speech frame 1 and ends at speech frame t_s-1, and $p(W_1^{t_s-1} \cdot w_{t_s}^{t_e} \cdot W_{t_e+1}^T, X_1^T)$ denotes the joint probability of word sequence $W_1^{t_s-1} \cdot w_{t_s}^{t_e} \cdot W_{t_e+1}^T$ and speech utterance X_1^T. On the other hand, the subword-level acoustic verification score of word segment w, which starts at speech frame t_s and ends at speech frame t_e, can be expressed as [Chen *et al.* 1998]

$$Score_{AV}(w_{t_s}^{t_e}) = \frac{1}{N_w} \sum_{i=1}^{N_w} \frac{2}{1+\exp[-\tau \cdot LLR(Sub(i)) + \eta]}, \tag{9}$$

where N_w is the number of subword (INITIAL or FINAL) units involved in the word segment w; $\frac{2}{1+\exp[-\tau \cdot LLR(Sub(i)) + \eta]}$ is a sigmoid function which provides the acoustic verification score for the subword unit $Sub(i)$; τ and η are used to control the slope and shift of the sigmoid function, respectively; and $LLR(Sub(i))$ is the log likelihood ratio for $Sub(i)$. In this research, τ and η were set to 0.5 and zero, respectively. The value of $LLR(Sub(i))$ can be calculated using the following equation:

$$LLR\,(Sub\,(i)) = \log \frac{p(X_{t_1}^{t_2} \mid Sub\,(i))}{\max\limits_{Sub^*} p(X_{t_1}^{t_2} \mid Sub^*)}, \tag{10}$$

where t_1 and t_2 are, respectively, the start and end speech frames of subword unit $Sub(i)$, $P\left(X_{t_1}^{t_2} \mid Sub(i)\right)$ is the likelihood that the speech segment $X_{t_1}^{t_2}$ will generated by $Sub(i)$, and $\max_{Sub^*} p(X_{t_1}^{t_2} \mid Sub^*)$ is the likelihood that $X_{t_1}^{t_2}$ will be generated by the corresponding top 1 subword unit, which acts here as the competing subword unit. From Equations (9) and (10), it is clear that the subword-level acoustic verification score for $Sub(i)$ becomes 1 if $Sub(i)$ is just the top 1 candidate and decreases to zero as $P\left(X_{t_1}^{t_2} \mid Sub(i)\right)$ becomes much smaller than $\max_{Sub^*} p(X_{t_1}^{t_2} \mid Sub^*)$. The word-level posterior probability and subword-level acoustic verification score were set within the range of 0 to 1 and can be weighted to form the word confidence measure:

$$CM(w_{t_s}^{t_e}) = c_1 \cdot P_{Post}(w_{t_s}^{t_e} \Big| X_1^T) + c_2 \cdot Score_{AV}(w_{t_s}^{t_e}), \tag{11}$$

where c_1 and c_2 are weighting parameters, whose values were set here to be equal, that is, c_1=c_2=0.5. Thus, we can use the word confidence measure to locate the most probably correct words. As the word confidence thresholds were varied, different amounts of automatically transcribed data were accordingly selected and used in combination with the original 4-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.

5. 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 [Rosenfeld 2000; Goodman 2001; Bellegarda 2004]. The *n*-gram probabilities are usually computed based on either the maximum likelihood (ML) principle or the maximum entropy (ME) principle [Berger *et al.* 1996]. However, 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 [Chen and Goodman 1999; Chen and Rosenfeld 2000]. In addition, statistical language modeling was also introduced to information retrieval (IR) problems in the late 1990s, and research at a number of sites has confirmed that such a modeling paradigm does provide a theoretically attractive and potentially very effective probabilistic framework for building IR systems [Croft and Lafferty 2003; Liu and Croft 2005; Zhai and Lafferty 2004]. 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 [Federico and Bertoldi 2001] or the recognition hypotheses cached so far [Jelinek *et al.* 1991]. 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

$$\hat{P}_{Adapt-1}\left(w_i \middle| w_{i-2}w_{i-1}\right)=\frac{\alpha\cdot C_{d,Cont}\left(w_{i-2}w_{i-1}w_i\right)+\beta\cdot C_{d,Back}\left(w_{i-2}w_{i-1}w_i\right)}{\alpha\cdot C_{Cont}\left(w_{i-2}w_{i-1}\right)+\beta\cdot C_{Back}\left(w_{i-2}w_{i-1}\right)}, \tag{12}$$

and

$$\hat{P}_{Adapt-2}\left(w_i \middle| w_{i-2}w_{i-1}\right)=\gamma\cdot P_{Cont}\left(w_i \middle| w_{i-2}w_{i-1}\right)+\left(1-\gamma\right)\cdot P_{Back}\left(w_i \middle| w_{i-2}w_{i-1}\right). \tag{13}$$

For the count merging formula in Equation (12), $C_{d,Cont}\left(w_{i-2}w_{i-1}w_i\right)$ and $C_{d,Back}\left(w_{i-2}w_{i-1}w_i\right)$ are, respectively, the discounted trigram counts [Chen and Goodman 1999] accumulated from the contemporary and background text corpora; $C_{Cont}\left(w_{i-2}w_{i-1}\right)$ and $C_{Back}\left(w_{i-2}w_{i-1}\right)$ 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 Equation (13), $P_{Cont}(w_i|w_{i-2}w_{i-1})$ and $P_{Back}\left(w_i \middle| w_{i-2}w_{i-1}\right)$ 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 Equations (12)-(13) also can be found in [Bacchiani and Roark 2003]. 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.

6. Experimental Results

In this section, we will present a series of experiments performed to assess recognition performance as a function of the feature extraction approaches, the decoding methods, and the acoustic learning and language adaptation approaches.

Table 1. The baseline character error rates (%) achieved using different feature extraction approaches.

	Character Error Rate (%)	
	TS	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

6.1 The Baseline Results

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 3 to 5 of Table 1, where the third (MFCC) row lists the results obtained using the MFCC-based approach, and the fourth (LDA-1) and fifth (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 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.

6.2 Experiments 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 section 6.1 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 (L_{AC}), 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.

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	L_{AC}	TS	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%)

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 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.

Table 3. The character error rates (%) achieved with different amounts of automatically transcribed speech 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

6.3 Experiments on Lightly Supervised Acoustic Model Training

Table 3 summarizes the performance of lightly supervised 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 4-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 [Gales and Woodland 1996]. 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 4-hour manually transcribed data. Use of the word confidence measure aided selection of the best subset of automatically transcribed data for acoustic training. Meanwhile, use of the online unsupervised MLLR speaker adaptation technique also resulted in additional performance gains under all experimental conditions.

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

		Character Error Rate (%)		Perplexity
		WG	+MLLR	
No LM Adaptation		15.34	14.71	670.23
Count Merging	α =1, β =1	13.22	12.60	437.87
	α =3, β =1	12.89	12.17	367.18
	α =5, β =1	12.95	12.22	397.80
	α =7, β =1	13.06	12.36	425.22
	α =9, β =1	13.15	12.46	450.99
Model Interpolation	γ =0.1	13.19	12.48	517.12
	γ =0.3	12.63	11.99	411.62
	γ =0.5	12.47	11.88	373.92
	γ =0.7	12.49	11.91	359.26
	γ =0.9	12.68	12.06	363.34

6.4 Experiments on 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 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 rest of the rows show, respectively, the results obtained for the systems when either the count merging adaptation strategy or the model interpolation adaptation strategy was adopted. In this study, for count merging, the value of weighting parameter β was fixed at 1, and the value of weighting parameter α was varied from 1 to 9 with a step size of 2; meanwhile, for model interpolation, the value of weighting parameter γ was varied from 0.1 to 0.9 with a step size of 0.2. Comparatively speaking, the best results for model interpolation ($\gamma = 0.5$ or $\gamma = 0.7$) were slightly better than those for count merging ($\alpha = 3, \beta = 1$),in terms of either the character error rate or perplexity reductions. The character rate decreased significantly from 14.71% to 11.88% ($\gamma = 0.5$ and +MLLR), and the perplexity value also can be reduced from 670.23 to 359.26 ($\gamma = 0.7$), which is just about a half of the original perplexity value. 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.

7. Conclusions

This paper has presented the initial results of a long-term research project on automatic recognition, indexing and summarization of Mandarin speech information. Several improved approaches to Mandarin broadcast news speech recognition have been presented. With the special structural properties of the Chinese language taken into consideration, a fast acoustic look-ahead technique using syllable-level heuristics has been proposed, and an overall speedup of more than 31% has been achieved in experiments. A verification-based method for automatic acoustic data acquisition has also been proposed to make use of large amount of untranscribed speech data, and very encouraging recognition results have been obtained. Two alternative strategies for language model adaptation have also been shown to be helpful in reducing both the character error rate and perplexity. The broadcast news system finally yielded an 11.88% character error rate when applied to a Mandarin broadcast news test set.

Acknowledgements

The authors are thankful to the National Science Council, R.O.C., for financial supports of this work (grant no. 91-2218-E-003-002 and 92-2213-E-003-008). They also thank the NTU Speech Processing Lab for providing the necessary speech and language data.

References

Aubert, X. L., “An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition,” *Computer Speech and Language*, vol. 16, 2002, pp. 89-114.

Bacchiani, M. and B. Roark, "Unsupervised Language Model Adaptation," *IEEE International Conference on Acoustics, Speech, Signal processing*, vol. I, 2003, pp. 224-227.

Bellegarda, J. R., "Statistical Language Model Adaptation: Review and Perspectives," *Speech Communication*, vol. 42, 2004, pp. 93-108.

Berger, A., S. D. Pietra and V. D. Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, vol. 22, no. 1, 1996, pp. 39-71.

Beyerlein, P., 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 Communication*, vol. 37, 2002, pp. 109-131.

Chang, E., F. Seide, H. Meng, Z. Chen, Y. Shi and Y.C. Li, "A System for Spoken Query Information Retrieval on Mobile Devices," *IEEE Trans. on Speech and Audio Processing,* vol. 10, no. 5, 2002, pp. 531-541.

Chang, P.C., S.P. Liao and L.S. Lee, "Improved Chinese Broadcast News Transcription by Language Modeling with Temporally Consistent Training Corpora and Iterative Phrase Extraction," *Proc. European Conference on Speech Communication and Technology,* 2003, pp. 421-424.

Chen, L., J.-L. Gauvain, L. Lamel and G. Adda, "Unsupervised Language Model Adaptation for Broadcast News," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal processing,* vol. I, 2003, pp. 220-223.

Chen, B., H.M. Wang, L.F. Chien and L.S. Lee, "A*-Admissible Key-Phrase Spotting with Sub-Syllable Level Utterance Verification," P*roc. International Conference on Spoken Language Processing,* 1998, CD-ROM.

Chen, B., H.M. Wang and L.S. Lee, "Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese," *IEEE Trans. on Speech and Audio Processing,* vol. 10, no. 5, 2002, pp. 303-314.

Chen, B., J.W. Kuo and W.H. Tsai, "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," *Proc. IEEE International Conference on Acoustics, Speech, Signal processing,* vol. I, 2004, pp. 777-780.

Chen, S.F. and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," *Computer Speech and Language,* vol. 13, 1999, pp. 359-394.

Chen, S.F. and R. Rosenfeld, "A Survey of Smoothing Techniques for ME Models," *IEEE Trans. on Speech and Audio Processing,* vol. 8, no. 1, 2000, pp. 37-50.

Croft, W.B., (editor) and J. Lafferty (editor), Language Modeling for Information Retrieval, Kluwer-Academic Publishers, 2003.

Davis, S.B. and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustic, Speech, and Signal Processing,* vol. 28, no. 4, 1980, pp. 357-366.

Dempster, A.P., N. M. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of Royal Statistical Society B*, 1977, vol. 39, no. 1, pp. 1-38.

Duda, R.O. and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, New York, 1973.

Gales, M.J.F. and P.C. Woodland, "Mean and Variance Adaptation within the MLLR Framework," *Computer Speech and Language,*" vol. 10, 1996, pp. 249-264.

Gauvain, J.-L., L. Lamel and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication,* vol. 37, 2002, pp. 89-108.

Goodman, J., "A Bit of Progress in Language Modeling," *Computer Speech and Language,* vol. 15, 2001, pp. 403-434.

Evermann, G. and P.C. Woodland, "Design of Fast LVCSR Systems," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding,* 2003, pp. 7-12.

Federico, M. and N. Bertoldi, "Broadcast News LM adaptation Using Cotemporary Texts," *Proc. European Conference on Speech Communication and Technology,* vol. 1, 2001, pp. 239-342.

Furui, S., T. Kikuchi, Y. Shinnaka and C. Hori, "Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech," *IEEE Trans. on Speech and Audio Processing,* vol. 12, no. 4, 2004, pp. 401-408.

Jelinek, F., 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.

Kemp, T. and A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," *Proc. European Conference on Speech Communication and Technology,* vol. 6, 1999, pp. 2725-2728.

Kneser, R. and H. Ney, "Improved Backing-off for M-gram Language Modeling," *Proc. IEEE International Conference on Acoustics, Speech, Signal processing,* vol. I, 1995, pp. 181-184.

Lamel, L., J.L. Gauvain and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer Speech and Language*, vol. 16, no.1, 2002, pp. 115-229.

LDC 2003, Chinese Gigaword Corpus: http://www.ldc.upenn.edu.

Lee, L.S., "Voice Dictation of Mandarin Chinese," *IEEE Signal Processing Magazine*, vol. 14 no. 4, 1997, pp. 63-101.

Liu, X. and W.B. Croft, "Statistical Language Modeling for Information Retrieval," to appear in *Annual Review of Information Science and Technology,* vol. 39, 2005.

Macherey, W. and H. Ney, "Towards Automatic Corpus Preparation for a German Broadcast News Transcription System," *Proc. IEEE International Conference on Acoustics, Speech, Signal processing,* vol. I, 2002, pp. 733-736.

Meng, H., B. Chen, S. Khudanpur, G. A. Levow, W. K. Lo, D. Oard, P. Schone, K. Tang, H.M. Wang and J. Wang, "Mandarin English Information (MEI): Investigating Translingual Speech Retrieval," *Computer Speech and Language,* vol. 18, no. 2, 2004, pp. 163-179.

Nguyen, L. and B. Xiang, "Light Supervision in Acoustic Model Training," *Proc. IEEE International Conference on Acoustics, Speech, Signal processing*, vol. I, 2004, pp. 185-188.

Ortmanns, S., H. Ney and X. Aubert, "A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition," *Computer Speech and Language,* vol. 11, 1997, pp. 43-72.

Ortmanns, S. and H. Ney, "Look-ahead Techniques for Fast Beam Search," *Computer Speech and Language,* vol. 14, 2000, pp. 15-32.

Rosenfeld, R., "Two Decades of Statistical Language Modeling: Where Do We Go from Here," *Proc. IEEE*, vol. 88, no. 8, 2000, pp. 1270-1278.

Saon, G. and M. Padmanabhan, "Data-Driven Approach to Designing Compound Words for Continuous Speech Recognition," *IEEE Trans. on Speech And Audio Processing,* vol. 9, no. 4, 2001, pp. 327-332.

Schuster, M., "Memory-efficient LVCSR Search Using a One-Pass Stack Decoder," *Computer Speech and Language,* vol. 14, 2000, pp. 47-77.

Stolcke, A., SRI language Modeling Toolkit, version 1.3.3, 2000. http://www.speech.sri.com/projects/srilm/.

Wang, C.J., B. Chen and L.S. Lee, "Improved Chinese Spoken Document Retrieval with Hybrid Modeling and Data-driven Indexing Features," *Proc. International Conference on Spoken Language Processing,* 2002, pp. 1985-1988.

Wessel, F. 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.

Wessel, F., R. Schluter, K. Macherey and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," *IEEE Trans. on Speech and Audio Processing,* vol. 9, no 3, 2001, pp. 288-298.

Woodland, P.C., "The Development of the HTK Broadcast News Transcription System: An Overview," *Speech Communication,* vol. 37, 2002, pp. 47-67.

Zhai, C.X. and J. Lafferty, "A Study of Smoothing Methods for Language Models Applied to Information Retrieval," *ACM Trans. on Information Systems,* vol. 22, no. 2, 2004, pp. 179-214.