LIGHTLY SUPERVISED AND DATA-DRIVEN APPROACHES TO MANDARIN BROADCAST NEWS TRANSCRIPTION

Berlin Chen, Jen-Wei Kuo, Wen-Hung Tsai
Graduate Institute of Computer Science & Information Engineering,
National Taiwan Normal University, Taipei, Taiwan, Republic of China
{berlin,rogerkuo,louis}@csie.ntnu.edu.tw

ABSTRACT

This paper investigates the use of several lightly supervised and data-driven approaches to Mandarin broadcast news transcription. First, with a consideration of the special structural properties of the Chinese language, a fast acoustic look-ahead technique for estimating the unexplored part of speech utterance was integrated into the lexical tree search to improve the search efficiency, in conjunction with the conventional language model look-ahead technique. Then, a verification-based method for automatic acoustic training data acquisition was developed to make use of the large amount of untranscribed speech data. Finally, two alternative strategies for language model adaptation were further studied for accurate language model estimation. With the above approaches, the system yielded an 11.94% character error rate on the Mandarin broadcast news collected in Taiwan.

1. INTRODUCTION

With the continuing growth of accessible multimedia information over the Internet, large volumes of real-world speech information, such as 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 on broadcast news transcription and retrieval have been reported [1-2]. However, in order to obtain better recognition performance, most of the transcription systems required not only large amounts of manually transcribed speech materials for acoustic training in the data preparation phase, but also much time expenditure and memory overhead in the recognition phase. Hence, in the recent past, several attempts have been made to investigate the possibility of unsupervised acquisition of speech or language training data for system refinement or for rapidly prototyping a new recognition system to new domains, and very encouraging results were initially achieved [3-4]. On the other hand, quite a few papers also have been working on the exploration of ways to improve the recognition efficiency, and thus many good approaches have been proposed [5]. With these observations, several lightly supervised and data-driven approaches to Mandarin broadcast news transcription were proposed in this paper. First, considering the special structural properties of the Chinese language, a fast acoustic look-ahead technique using syllable-level heuristics was integrated into the lexical tree search to improve the search efficiency, in conjunction with the conventional language model look-ahead technique. Then, a verification-based method for automatic acoustic training data acquisition was developed to make use of the large speech corpus. Finally, two alternative strategies for language model adaptation were further studied for accurate language model estimation. All these improved approaches presented in this paper have been successfully integrated into our NTNU broadcast news system while a prototype system for voice access of Mandarin broadcast news speech has also been established [2].

2. THE NTNU BROADCAST NEWS SYSTEM

The overall framework of the broadcast news system developed at National Taiwan Normal University (NTNU) is depicted in Figure 1. We review some of its main features below.

2.1 Front-End Processing

The front-end processing is conducted with two feature extraction approaches: the conventional MFCC-based (Mel-frequency Cepstral Coefficients) and the data-driven LDA-based (Linear Discriminant Analysis) approaches. For the MFCC-based approach, 13-dimensional cepstral coefficients derived from 18 filter bank outputs were augmented with their first- and second-order time derivatives. For the LDA-based based approach, the states of each HMM were taken as the unit for class assignment. Either the outputs of filter banks or the cepstral coefficients were alternatively chosen as the basic vectors. The basic vectors from every nine successive frames were spliced together to form the supervectors for the construction of the LDA transformation matrix, which was then used to project the supervectors to a lower feature space. The dimension of the resultant vectors was set to 39, which was just the same as that of the MFCC-based approach.

2.2 Acoustic Training

The speech data set consists of about 112 hours of FM radio broadcast news, which were collected from several radio stations located at Taipei in the 1998-2002 period [2]. All the speech materials were manually segmented into separate stories, and each of them is a news abstract pronounced by one anchor speaker. Some stories contain background noise and music. Only 7.7 hours of speech data is equipped with corresponding orthographic transcripts, in which about 4.0 hours of data collected during 1998 to 1999 is used to bootstrap the acoustic training and the other 3.7 hours of data collected in September 2002 is for testing. The rest 104.3 hours of untranscribed speech data is reserved for unsupervised acoustic training. The acoustic models chosen for speech recognition are 112 right-context-dependent INITIAL's and 38 context-independent FINAL's, specially considering the phonetic structure of Mandarin syllables [2]. Here INITIAL is the initial consonant of the syllable and

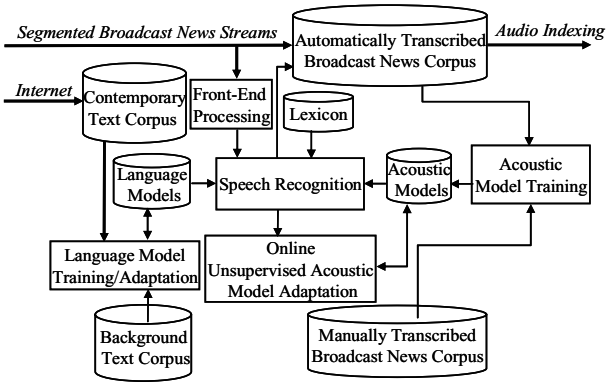

Figure 1: The overall framework of the NTNU broadcast news system.

FINAL is the vowel (or diphthong) part but including optional medial or nasal ending. Each INITIAL is represented by an HMM with 3 states while each FINAL with 4 states. The Gaussian mixture number per state ranges from 2 to 128, depending on the quantity of training data. For all experiments, gender-independent models were used.

2.3 Lexicon and Language Modeling

In the Chinese language, each of the large number of characters (at least 7,000 are common used) is pronounced as a monosyllable and is morpheme with its own meaning. New words are very easily generated by combining a few characters, but nevertheless would be tokenized into several single-character words or words with fewer characters as the text corpus is processed for language model training. This definitely makes the out-of-vocabulary problem especially important for Mandarin broadcast news transcription. In order to alleviate the out-of-vocabulary problem, compound words should be carefully selected and added to the lexicon in correspondence with 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:

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

, where

$$P_f(w_j \mid w_i) = \frac{P(W_{t+1} = w_j, W_t = w_i)}{P(W_t = w_i)}, \tag{2}$$

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

We started with a lexicon composed of 67K words, and iteratively used the above measures with varying thresholds to find the possible word pairs which can 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 background language models used in this paper consist 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. On the other hand, a corpus consisting of 50 million Chinese characters of newswire texts collected from the Internet during August to October 2002 is used as the contemporary corpus for unsupervised language model adaptation. The language models were trained with Kneser-Ney backoff smoothing using the SRI Language Modeling Toolkit (SRILM) [6].

2.4 Speech Recognition

Like most conventional approaches, our baseline recognizer was implemented with a left-to-right frame-synchronous tree search as well as a lexical prefix tree organization of the lexicon [5]. At each speech frame, the so-called word-conditioned method grouped the path hypotheses that shared the same history of predecessor words to the same copies of the lexical tree, and expanded and recombined them according to the tree structure until reaching a possible word ending. At word boundaries, the path hypotheses among the tree copies that had the equivalent search history were recombined, and were then propagated into the existing tree copies or used to start up new ones in case that they did not exist yet. Note that these tree copies were just built in a mental view. During the search process, only one lexical tree structure was built for reference, 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 predecessor words, the phonetic arc (i.e. an INITIAL or FINAL in Mandarin Chinese) in the lexical tree, the model 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 unigram language model look-ahead scores, was used to select the most promising path hypotheses. Moreover, if the word hypotheses ending at each speech frame had scores higher than a predefined threshold, their associated decoding information, such as the word begin and end frames, the identities of current and predecessor words, and the acoustic score, will be kept in order to build a word graph for further language model rescoring. Once the word graph had been built, the forward-backward search with a more sophisticated language model was conducted on it to generate the most likely word sequence. In this paper, the word bigram language model was used in the tree search procedure while the trigram language was used in the word graph rescoring procedure.

3. ACOUSTIC LOOK-AHEAD USING SYLLABLE-LEVEL HEURISTICS

In the baseline recognizer, language model look-ahead and beam pruning techniques can be incorporated together to help to retain the most promising path hypotheses for further expansion. However, the crucial problem for such an approach is that it does not consider the potential likelihood of the unexplored portion of speech utterance as beam pruning is applied, which will unavoidably include many unpromising path hypotheses and increase ambiguities during the search process. Therefore, the search efficiency will be degraded, since too many path hypotheses have to be examined at every speech frame. On the other hand, as it is widely known, the Chinese language is attributed to its monosyllabic structure, in which each Chinese word is composed of one to several syllables (or characters) and thus syllables are the very important constituent units of Chinese words [2]. Besides, there are only about 400 syllables in the Chinese language, if the tonal information is further ignored. It implies that the

	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

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

syllable recognition is much faster than the word recognition. Thus, in this paper, we proposed to utilize the syllable-level heuristics to enhance the search efficiency. A compact syllable lattice based on the structural information of words in the lexicon was automatically built, which was used to estimate the likelihood of the unexplored portion of speech utterance. Each HMM state in the syllable lattice can 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 as the likelihood estimation for acoustic look-ahead. In the second pass, the frame-synchronous search was performed by incorporating the language model look-ahead scores together with the acoustic look-ahead scores for beam pruning. Though speech recognition is now in a two-pass mode, the time spent on the calculation of acoustic look-ahead scores is almost negligible. The word graph rescoring procedure also can be applied after the second-pass search.

4. UNSUPERVISED ACOUSTIC MODEL TRAINING

Speech recognition essentially relies on large amounts of manually transcribed speech data for acoustic training, especially when porting the system to new application domains. However, generating the manually transcribed data is an expensive process in terms of both manpower and time. Based on this observation, we investigated here the unsupervised acoustic training technique for Mandarin broadcast news recognition. 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. 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, the sub-syllable-level verification score [2]) were incorporated together. 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 the automatically transcribed data were accordingly selected, and were used in combination with the original 4-hour manually transcribed corpus to retrain different sets of acoustic models. The LDA transformation matrix used in feature extraction should be reestimated and the acoustic features were recalculated as well, in according to the speech data selected for training.

	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 (41.6%)	0.149 (24.0%)	1.214 (31.9%)

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.

5. UNSUPERVISED LANGUAGE MODEL ADAPTATION

For complex speech recognition tasks such as broadcast news transcription, it is extremely difficult to build well-estimated language models, because the statistical characteristics for the linguistic contents of news articles are very diverse and are often changing with time. Due to the dynamic nature of this task, we attempted to alleviate such an inconsistence problem in language modeling by investigating different ways to combine the online collected temporally consistent text corpus with the original background text corpus for more accurate language model estimation. Both count merging as well as model interpolation, which can be respectively viewed as a maximum *a posteriori* (MAP) adaptation with a different parameterization of the prior distribution [5], were investigated here as the strategies for unsupervised language model adaptation. As mentioned previously, a corpus of contemporary Internet newswire texts collected during August to October 2002 was used to provide an additional prediction for the linguistic events of the testing broadcast news collected in September 2002.

6. EXPERIMENTAL RESULTS

In this section a series of experiments was performed to assess recognition performance as a function of the feature extraction approaches, the decoding methods as well as the unsupervised acoustic and language learning approaches.

6.1 The Baseline Results

The baseline broadcast news system was alternatively configured with the conventional MFCC-based and the data-driven LDA-based feature extraction approaches. The results are shown in Rows 3 to 5 of Table 1, where the third (MFCC) row stands for the results using the MFCC-based approach and the fourth (LDA-1) and the fifth (LDA-2) rows are respectively the results when different basic vectors were adopted in the LDA matrix construction. In LDA-1, the cepstral coefficients are selected as the basic vectors, while in LDA-2, the filter bank outputs as the basic vectors. As can be seen in Table 1, the character error rates for the two variant LDA-based approaches, either after the tree-copy search (TC) or word-graph rescoring (WG), are significantly better than that of the standard MFCC-based approach. Moreover, the LDA-2, which used the filter bank outputs directly as the basic vectors, is considered to be even more efficient than the MFCC-based for feature extraction. The LDA-2 features are thus chosen as the default acoustic features for the following experiments.

6.2 Experiment on Acoustic Look-Ahead Using Syllable-Level Heuristics

The recognition performance and efficiency, as the acoustic look-ahead technique was further integrated into the system, were evaluated. These results were obtained by using the same beam pruning threshold as those reported before, and were measured on an ordinary 2.6 GHz Pentium IV PC. The results in search efficiency are shown in Columns 2 to 6 of Table 2, which are respectively the real time factors for feature extraction and HMM state emission probability calculation (FE), acoustic look-ahead (AL), tree-copy search (TC), word-graph rescoring (WG), and overall recognition time (Total), while the results in recognition accuracy are shown in the last row of Table 1. The numbers in the parentheses of the last row of Table 2 are the relative speedups when compared with the results shown in the second row. By comparing the results in the last two rows of Table 1, it can be seen that the recognition accuracy was slightly degraded as acoustic look-ahead was used. However, according to the results shown in Table 2, the recognition efficiency for lexical tree search is significantly improved while the time spent on acoustic look-ahead is almost negligible. In summary, the acoustic look-ahead method proposed here achieves an overall speedup of more than 31% and makes the whole system run almost in real time.

6.3 Experiment on Unsupervised Acoustic Model Training and Adaptation

Table 3 summarizes the performance for unsupervised acoustic training. Column 2 (WG) shows the recognition results achieved by using several sets of acoustic models, which are 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 selected unsupervised acoustic training data and the numbers in the parentheses are the corresponding word confidence thresholds used. In addition, the third column presents the results when online unsupervised MLLR (Maximum Likelihood Linear Regression) speaker adaptation was further included. As can been seen in Table 3, with a careful selection of automatically transcribed speech data, the character error rate can be effectively reduced from 20.12% to 15.34%, as a total amount of 21 hours of automatically transcribed data was used for acoustic training, in combination with the original 4-hour manually transcribed data. The use of the word confidence measure actually can help to select the best subset of automatically transcribed data for unsupervised acoustic training. Meanwhile, the online unsupervised MLLR speaker adaptation also gives additional improvements for all the experimental conditions.

6.4 Experiment on Unsupervised Language Model Adaptation

The results for unsupervised language adaptation 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, in which the character error rates are just the best ones shown in Table 3. The third and forth 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 11.94 (+MLLR) and can give a reduction of almost a half of the perplexity as well.

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

	Character Error Rate		Perplexity
	WG	+MLLR	
No LM Adaptation	15.34	14.71	670.23
Count Merging	12.94	12.20	367.34
Model Interpolation	12.59	11.94	357.53

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.

7. CONCLUSIONS

We have presented several improved approaches to Mandarin broadcast news speech recognition, including fast acoustic look-ahead, unsupervised acoustic model training and language adaptation. Very encouraging results were achieved. The broadcast news system finally yielded an 11.94% character error rate on the Mandarin broadcast news test set.

8. ACKNOWLEDGEMENTS

The authors would like to thank the NTU Speech Processing Lab for providing the necessary speech and language data.

9. REFERENCES

[1] P. Beyerlein et al., “Large vocabulary continuous speech recognition of Broadcast News – The Philips/RWTH approach,” Speech Communication, May 2002.

[2] B. Chen, H-M Wang, and L-S Lee, "Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese", IEEE Trans. on Speech and Audio Processing, July 2002.

[3] W. Macherey et al., “Towards Automatic Corpus Preparation for A German Broadcast News Transcription System,” in Proc. ICASSP 2002.

[4] M. Bacchiani et al., “Unsupervised Language Model Adaptation,” in Proc. ICASSP 2003.

[5] X. L. Aubert, “An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition,” Computer Speech and Language, January 2002.

[6] A. Stolcke, “SRI language Modeling Toolkit,” version 1.3.3, http://www.speech.sri.com/projects/srilm/.