A MANDARIN LECTURE SPEECH TRANSCRIPTION SYSTEM FOR SPEECH SUMMARIZATION

Ho Yin Chan, Justin Jian Zhang, Pascale Fung, Lu Cao

Human Language Technology Center
Electronic and Computer Engineering
University of Science and Technology
Clear Water Bay, Hong Kong
ricky@cs.ust.hk, zjustin@ust.hk,
pascale@ee.ust.hk

ABSTRACT

This paper introduces our work on mandarin lecture speech transcription. In particular, we present our work on a small database, which contains only 16 hours of audio data and 0.16M words of text data. A range of experiments have been done to improve the performances of the acoustic model and the language model, these include adapting the lecture speech data to the reading speech data for acoustic modeling and the use of lecture conference paper, power points and similar domain web data for language modeling. We also study the effects of automatic segmentation, unsupervised acoustic model adaptation and language model adaptation in our recognition system. By using a 3xRT multiple passes decoding strategy, we obtain 70.3% accuracy performance in our final system. Finally, we apply our speech transcription system into a SVM summarizer and obtain a ROUGE-L F-measure of 66.5%.

Index Terms— lecture speech transcription, model adaptation, multi-pass decoding, speech summarization

1. INTRODUCTION

Large amounts of work have been done for automatic transcription of broadcast news and telephone conversations. Research for automatic transcription of spontaneous speech in the environment of lecture presentation or conference presentation, however, has received relatively less attention in the field. In [1], [2], [3] and [4], Kawahara et al. presented their work on spontaneous Japanese lecture speech recognition. By using 37.9 hours of lecture speech for acoustic modeling and 1.48M words of transcribed text for language modeling and by incorporating speaking-rate dependent decoding and adaptation, they obtained 69.2% word accuracy in their experiments [3]. In [4], they showed improvement in lecture speech recognition by using more training data (60 hours of lecture speech and 3.15M words of transcribed text) and by applying unsupervised language model adaptation, where word error rate of 28.7% was achieved. [5] studied unsupervised language model adaptation for Japanese lecture speech transcription and showed that improvement in language model perplexity and recognition word error rate can be obtained. In [6] and [7], the ISL performed lecture transcription on English. As Mandarin is one of the most popular languages used in the world, it is interesting to study the Mandarin lecture speech transcription.

In this paper, we present our initial work on Mandarin lecture speech transcription. In particular, we collected a database, which contains 16 hours of audio data and 0.16M words of text data. Since our database is relatively small, we aim to improve our acoustic model and language model by using available data from other sources. These include the use of reading speech audio data for acoustic modeling and the use of lecture conference paper, power points and similar domain web data for language modeling. As our recognition system is performed offline, we aim to achieve the best performance by using multiple passes decoding strategy together with unsupervised acoustic model adaptation and cross system language model adaptation techniques. An application of lecture speech transcription for summarization is also implemented and compared to manual summarization.

The rest of the paper is organized as follows. In Section 2, we describe the audio corpora and the text corpora that were used in this work. Then, a baseline system is described in section 3. In section 4, model adaptation and multiple passes decoding strategy are presented. The application of our transcription system for speech summarization is presented in Section 5. Finally, conclusions are given in Section 6.

2. DATABASE

The lecture audio data collected for this work consists of 60 Mandarin oral presentations given by different speakers in the Chinese National Conference on Man-Machine Speech Communication (NCMMSC 2005). Each of the presentation lasts for 15-20 minutes and is recorded at 22 KHz and 16 bit

sampling rate. 55 presentations that amount to 16 hours of speech are chosen for training and 5 presentations that amount to 1.4 hours of speech are used for testing. All the lecture audio data are manually segmented and transcribed and down-sampled to 16 KHz.

The 863 Mandarin speech corpus and the HKU Mandarin speech corpus are used as the reading speech corpora. These corpora compose a total of 170 hours of reading speech from 250 speakers.

The text data used for language model training include the manual lecture speech transcriptions, the paper and power point presentations in the lecture conference, and web collected technical articles and conference papers. The total size of the text is 1.43M words, in which only 0.16M words are contributed from the lecture speech transcriptions and 0.17M words are contributed from the paper and power point presentations respectively.

3. BASELINE SYSTEM

3.1. Language Modeling

Chinese word segmentation is performed on the training corpora by using the HIT IR Lab Chinese Segmenter [8]. Vocabulary selection based on maximum likelihood [9] is then applied to the training data to obtain a wordlist of 6878 words. A total of 282 words, or 4.1% of the wordlist size, are English words. The out-of-vocabulary (OOV) rate of this wordlist on the test set is 1.0%. For each training data set, we built one language model with a cut-off threshold of two for the n-grams. The individual language models are then linear interpolated and merged to form a single language model. The interpolation weights are computed with the cross validation approach by dividing all the lecture speech transcription into five portions, and the estimation for each portion is done by the SRILM [10] toolkit.

Table 1: Perplexity for different language models

Language Model	Bigram	Trigram
(I) Lecture transcription	319	395
(II) Paper and power point	743	834
(III) Web data	1020	1167
Data mixing (I) (II)	231	247
Data mixing (I), (II), (III)	296	301
Interpolate model (I), (II)	222	233
Interpolate model (I), (II), (III)	213	215

Table 1 gives the bigram and trigram perplexities for language models under different training set. As can be seen, language models trained solely from lecture speech transcription gives very large perplexities due to insufficient data. By adding the paper and power point presentations from the lecture conference, the language models are improved. Further use of web collected data gives only little improvement. Since the perplexities of the produced bigram language models are smaller than trigram language models, we used the interpolated bigram language model from all training data in our subsequent baseline experiments.

3.2. Acoustic Modeling

Our system uses tied-state cross-word triphone HMMs that are constructed by decision tree clustering. The system uses up to 3500 tied states in total and each state contains 16 Gaussian mixture components. For every shift of 10ms, a 25ms window of input speech is represented by a feature vector that includes 13 MFCC cepstral parameters (including C0) and their 1st and 2nd order derivatives. Cepstral mean normalization (CMN) is applied on each speech segment. The number of phones for the system is 67, where each of the 27 Mandarin initial phones and silence are modeled by three states left-to-right HMM with no state-skipping, and each of the 37 Mandarin final phones, noise and unlabelled English word are modeled by five states left-to-right HMM. For the Mandarin final HMMs, state transitions are added such that a minimum of three frames are allowed for matching of short finals. During the training, English to Mandarin phone mapping is applied to a dictionary such that transcribed English words can be trained. Acoustic model training using ML criterion is done for three training sets: read speech only, lecture speech only, read speech and lecture speech.

Table 2: Character accuracy for different acoustic models

Acoustic model (# tied states)	Data Size (hrs)	Acc (%)
Reading speech (3.5k)	170	47.7
Lecture speech (2k)	16	65.5
Mixed Reading and Lecture speech (3.5k)	186	65.0
Lecture speech adapt mixed data model (3.5k)	186	66.6

Table 2 shows the performance of the acoustic models tested under the supervision of the interpolated bigram language model. As can be seen, the model trained purely from 170 hours of reading speech gives very poor performance due to speaking style mismatch between read speech and lecture speech. By using 16 hours of lecture speech only, we obtained 65.5% character accuracy. The results also show that mixing large amounts of reading speech data to lecture speech data performs 0.5% worse than the acoustic model trained merely from lecture speech data. We obtain the best model by adapting the lecture speech data to the mixed data acoustic model. This is done by using the maximum likelihood linear regression (MLLR) [11] follows by the maximum a posteriori (MAP) [12] criterion, where a 1.1% absolute improvement is obtained compare to the model

trained from lecture speech only. This model is used for testing in our later experiments.

3.3. Decoding and Automatic Segmentation

The baseline system uses a single pass decoder. The decoder performs time-synchronous Viterbi beam search through a lexical tree and runs in a total of 1xRT for Chinese word bigram decoding by using a 1.86GHz dual core processor and 1GB memory. Automatic segmentation for the lecture audio is also performed and compared to manual segmentation. We used all the lecture speech training data to train five events: silence, noise, Mandarin initial, Mandarin final and unlabelled English word. For each event, we trained a GMM with 256 components. The silence and noise events are modeled by three-state HMMs while the Mandarin initial, Mandarin final and unlabelled English word events are modeled by seven-state HMMs. A grammar based Viterbi decoder is used to produce the GMM sequences for the audio. The GMM sequences are then relabeled to speech/non-speech labels and post-processed with the BBN approach [13]. Several combinations of the gap-bridging parameter ***g*** for non-speech regions and silence padding ***P*** at either end of each created speech segment are tried. Table 3 gives the comparison between the automatic segmentation and the manual segmentation. Using our chosen parameters, the automatic segmentation gives more segments in total but shorter average length. We observed 1.0%-1.1% absolute degrade in accuracy performance by using automatic segmentation. But the degradation is eliminated after acoustic model adaptation (section 4).

Table 3: Comparison for different segmentations

Seg (g, P)	**# Segment**	**Average length**	**Acc (%)**
v1 (0.1, 0.1)	2339	1.8s	65.46
v2 (0.2, 0.1)	1916	2.19s	65.51
v3 (0.3, 0.2)	1331	3.14s	65.59
Manual	1254	3.91s	66.60

4. ADAPTATION AND MULTI-PASS DECODING

4.1. Unsupervised Language Model Adaptation

Since our recognition system runs in offline, it allows us to perform language model adaptation by using all recognized text. We first built a language model from the recognized text of the ASR system and then merged this language model with individual language models in our baseline system. To estimate the interpolation weights, we divided all the correct lecture speech transcriptions from the training data into five portions and then used the cross-validation method. Table 4 gives the perplexities of the adapted language models by using different automatic recognized text for adaptation and the corresponding character accuracy performances on the speech segments from manual segmentation. As can be seen, significant perplexity reductions are obtained by using the adapted models, and trigram language models now give smaller perplexities and higher recognition accuracies than bigram language models. The results also show that cross system language model adaptation gives better performance. For bigram testing, the language model adapted from the trigram recognized text gives better accuracy performance than the language model adapted from the bigram recognized text and vice versa for the trigram testing. The best adapted language models are obtained by adapting mixed recognized text from a bigram system and a trigram system, both using the acoustic models with MLLR adaptation. Compare with the un-adapted language models, we obtained 31% and 37.8% perplexity reductions from the adapted bigram language model and the adapted trigram language model respectively. The absolute character accuracy improvement by using the adapted bigram language model and the adapted trigram language model are 1.0% and 1.4% respectively.

Table 4: Perplexity and character accuracy for unsupervised LM adaptation with manual segmentation (a) bigram LM testing, (b) trigram LM testing.

(a)

Language model	Perplexity	Acc (%)
unadapted LM	213	66.6
adapt asr bg transcript	156	66.7
adapt asr tg transcript	155	67.0
adapt asr mllr transcript	147	67.6

(b)

Language model	Perplexity	Acc (%)
unadapted LM	215	66.5
adapt asr bg transcript	144	67.2
adapt asr tg transcript	144	67.1
adapt asr mllr transcript	135	68.0

4.2. Unsupervised Acoustic Model Adaptation

We also implemented unsupervised acoustic model adaptation and ran it offline with an iterative approach [14]. For each lecture presentation, a global maximum likelihood linear regression (MLLR) transform [11] is first estimated with the speech segments. The MLLR adaptation is then repeated by estimating four transforms, where a regression class tree is used to cluster the Gaussian components into four sets. Table 5 gives the character accuracy of our system under different stage of adaptation. By applying acoustic model adaptation only, we got a maximum of 2.4% and 3.6% absolute character error rate reduction for the manual segmentation system and the automatic segmentation system respectively. Moreover, we can see that after performing

acoustic model adaptation, the automatic segmentation system gives better performances than the manual segmentation system. After applying language model adaptation with the mixed recognized text from the bigram system and the trigram system which used the adapted acoustic models, we further improve our systems, where 1.0% and 1.5% absolute character error rate reductions are obtained for the bigram system and the trigram system respectively.

Table 5: Character accuracy under different stage of adaptation (a) bigram LM testing, (b) trigram LM testing

(a)

Model	Manual Seg	Auto Seg
Unadapted	66.7	65.6
Adapt AM	68.9	69.2
Adapt AM+LM	69.6	70.2

(b)

Model	Manual Seg	Auto Seg
Unadapted	66.5	65.7
Adapt AM	68.9	69.0
Adapt AM+LM	69.9	70.5

4.3. Multi-pass decoding strategy

All of the experiments described before were run by full decoding. In order to speed up the overall recognition process, we make use of the lattice during the multi-pass decoding. The decoding strategy of the final system is similar to [15]. In the first pass, a full decoding with un-adapted bigram language model and acoustic model is applied to produce 1-best result and a lattice. Lattice rescoring is then performed on trigram language model to obtain another 1-best result. Then, a bigram branch and a trigram branch are created and acoustic model adaptation with the MLLR approach is applied on each branch. Lattice rescoring is then performed on each branch with the adapted acoustic models, and produces 1-best recognition results. The recognized texts from the braches are mixed and then unsupervised trigram language model adaptation is performed. A final re-decoding is done by using the adapted acoustic model and the adapted trigram language model. By using this multi-pass decoding strategy, our final recognition system runs in a total of 3xRT. The character accuracy performances on the manual segmentation system and the automatic segmentation system are 69.7% and 70.3% respectively.

5. SPEECH SUMMARIZATION

We consider the extractive summarization as a binary classification problem; that is to say, we predict whether each sentence of the lecture transcription should be in a summary or not. We built the SVM classifier as our summarizer on acoustic features, lexical features and Poisson Noun as discourse feature [16][17]. We used a total of 6049 sentences from 34 presentations in the lecture speech training data to train our SVM classifier. We evaluate the summarizer's performance by ROUGE-L (summary-level Longest Common Subsequence) F-measure. Three sets of transcriptions are evaluated: manual transcriptions based on manual segmentation, ASR transcriptions based on manual segmentation and ASR transcriptions based on automatic segmentation. For the manually segmented system, the best summarization results are obtained by using a combination of acoustic and lexical features. For the automatically segmented system, the best result is obtained by using lexical features. The result is shown in Table 6. From the table, we can see that our summarizer yields good performance by using ASR transcriptions with automatic segmentation: a ROUGE-L F-measure of 0.665 is obtained, which is the same as the result produced by ASR transcriptions with manual segmentation and only 0.6% less than the result produced by manual transcriptions with manual segmentation.

Table 6: The Summarizer's Performance Evaluation

Manual Seg & trans	Manual Seg & ASR trans	Auto Seg & ASR trans
0.671	0.665	0.665

6. CONCLUSION

This paper presents our initial work on Mandarin lecture speech transcription. We collected a relatively small database with only 16 hours of lecture speech data. A range of experiments have been done for improving acoustic model and language model. In a 3xRT multiple passes decoding architecture, we obtain 70.3% character accuracy in our transcription system. The recognized results are further applied to lecture speech summarization in a SVR summarizer and produce a ROUGE-L F-measure of 66.5%. In future work, we will collect more data and perform discriminative training on the acoustic model. We will also investigate system combination from different branches of the recognition system.

7. REFERENCES

[1] T. Kawahara, H. Nanjo, and S. Furui, "Automatic transcription of spontaneous lecture speech," *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on,* pp. 186–189, 2001.

[2] H. Nanjo, K. Kato, and T. Kawahara, "Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition," *Proc. EUROSPEECH,* pp. 2531–2534, 2001.

[3] H. Nanjo and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," *International Conference on Acoustics, Speech, and Signal Processing, 2002. Proceedings.(ICASSP'02). IEEE International Conference on,* vol. 1, 2002.

[4] H. Nanjo and T. Kawahara, "Unsupervised language model adaptation for lecture speech recognition," *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition,* pp. 75–78, 2003.

[5] T. Niesler and D. Willett, "Unsupervised language model adaptation for lecture speech transcription," *Proc. ICSLP,* 2002.

[6] M. Wolfel and S. Burger, "The ISL Baseline Lecture Transcription System for the TED Corpus," Eurospeech, 2005.

[7] C. Fugen, M. Wolfel, J. McDonough, S. Ikbal, F. Kraft, K. Laskowski, M. Ostendorf, S. Stuker, and K. Kumatani, "Advances in lecture recognition: The ISL RT-06s evaluation system," Proc. Interspeech, 2006.

[8] H. Zhang, T. Liu, J. Ma and X. Liao, "Chinese Word Segmentation with Multiple Postprocessors in HIT-IRLab," *SIGHAN*. 2005

[9] A. Venkataraman and W. Wang, "Techniques for effective vocabulary selection," *Arxiv preprint cs.CL/0306022,* 2003.

[10] A. Stolcke, "Srilm-an extensible language modeling toolkit," *Proc. ICSLP,* vol. 2, pp. 901–904, Sept 2002.

[11] M.J.F. Gales, "Maximum likelihood linear transformations for HMM based speech recognition," *Computer, Speech and Language,* vol. 12, no. 2, pp. 75–98, 1998.

[12] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *Speech and Audio Processing, IEEE Transactions on,* vol. 2, no. 2, pp. 291–298, 1994.

[13] D. Liu and F. Kubala, "A cross-channel modeling approach for automatic segmentation of conversational telephone speech," *Proceedings of the 2003 ASRU Workshop, St. Thomas, November,* 2003.

[14] P.C. Woodland, D. Pye, and M.J.F. Gales, "Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression," *Proc. ICSLP 96,* pp. 1133–1136, 1996.

[15] G. Evermann and P.C. Woodland, "Design of fast LVCSR systems," *Automatic Speech Recognition and Understanding, 2003. ASRU'03. IEEE Workshop on,* pp. 7–12, 2003.

[16] J. Zhang, H.Y. Chan, P. Fung, and L. Cao, "A comparative study on speech summarization of broadcast news and lecture speech," *Interspeech 2007: Eight Annual Conference of the International Speech Communication Association,* 2007.

[17] J. Zhang, H.Y. Chan, and P. Fung, "Improving lecture speech summarization using rhetorical information," *To Appear: Automatic Speech Recognition and Understanding, 2007. ASRU'07. IEEE Workshop on,* 2007.