# Corpus and Transcription System of Chinese Lecture Room

*Sheng Li, Yuya Akita, Tatsuya Kawahara*

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
Email: lisheng@ar.media.kyoto-u.ac.jp

## Abstract

The paper introduces our project on automatic speech recognition (ASR) of Chinese lectures. For a comprehensive study on spontaneous Chinese, we compile a corpus of Chinese Lecture Room (CCLR), which has faithful transcripts and caption texts. Based on the annotated alignment of these texts, we conduct analysis on linguistic phenomena of spontaneous Chinese speech. We also develop a baseline ASR system with this corpus, and refine it with the DNN-HMM framework. By exploiting the lecture data without faithful transcripts and conducting unsupervised speaker adaptation, significant improvement of ASR accuracy is achieved.

**Index Terms:** speech recognition, acoustic model, lecture

## 1. Introduction

Online lecture services such as TED and edX are booming nowadays and they will bring revolutionary changes to both studying and teaching. Automatic speech recognition (ASR) technology will be useful for fast indexing and captioning of the large-scale and ever increasing media archives.

Automatic transcription of spontaneous speech is still challenging both acoustically and linguistically compared to common ASR tasks [1]. Although intensive studies have been conducted [1, 2, 3], current state-of-the-art transcription systems have not yet achieved usable results on spoken lectures. The natural speech transcription is much different from conventional speech recognition tasks such as broadcast news, because of acoustic variations, disfluencies and colloquial expressions.

While there are projects on spoken lecture transcription in English [2], Japanese [4] and European language [5], works on Chinese lectures or spontaneous speech in general are limited [6, 7]. There is no large corpus public available in this category.

For a comprehensive study on ASR of spontaneous Chinese, we compile a corpus of Chinese spoken lectures and investigate ASR technology for them. In this paper, an overview of this corpus and some linguistic analysis are presented. Then, we investigate an ASR system using this corpus and also incorporating deep neural network and lightly supervised training.

## 2. Corpus of Chinese Lecture Room (CCLR)

### 2.1. Corpus description

The spoken lectures are selected from "Lecture Room" (百家講壇), which is a very popular academic lecture program of China Central Television (CCTV) Channel 10. This program was designed for the purpose of delivering national flagship to all citizens. Since 2001, a series of lectures have been given by luminary figures from a variety of areas almost every week. By the end of the year 2013, we have finished annotation on 98 lectures. The total size of the corpus is 61.6 hours in speech and 1.2 M characters in text (Table 1). The corpus is named as the Corpus of Chinese Lecture Room (CCLR).

Table 1. *Basic corpus description*.

| #lectures | #speakers | durations | #characters |
|---|---|---|---|
| 98 | 21 female/69 male | 61.6 hours | 1.2 M |

### 2.2. Annotation scheme

A part of the annotated corpus (68 lectures) includes both faithful transcripts and caption texts. They are regarded as a parallel corpus between written style and spoken style. Based on previous studies on Chinese spontaneous phenomena [6, 7, 8, 9, 10, 11] and already existing corpora such as CSJ [4] and CASIA-863 [12], we figure out the most frequent and basic spontaneous phenomena including: fillers, grammatical particles, discourse markers, repairs, reorders, substitutions, and deletions. Other complex patterns can be regarded as composition of these basic patterns. Since it is very difficult and costly to annotate these phenomena accurately, we simplified the annotations into four categories: insertion, deletion, substitution and fillers (interjections). For each of these categories, we list their characteristic patterns in Table 2.

Table 2. *Major spontaneous phenomena*.

| Annotation | Spontaneous Patterns |
|---|---|
| **Fillers** | **Interjections:**<br>(examples: 啊, 哦, 鹅 … ) |
| **Insertion** | **Grammatical Particles:**<br>1.auxiliary fragment (examples: 的…)<br>2.aspect marker (examples: 了…)<br>3.question marker (examples: 吗…)<br>4.structure particle (examples: 是, 把, 被…) |
| | **Discourse markers:**<br>this, that ,then, that is to say etc.<br>(examples: 这, 那, 那么, 那就是说 …) |
| | **Repairs:**<br>1.correcting or giving up earlier statements<br>(examples: 到了-阴历-啊不-农历的七月初七)<br>2.further explanation or emphasis<br>(examples: 用一望远镜-天文望远镜-来观察星空)<br>3.repetition or partly repetition for hesitation or uncertainty (examples: 如-如果) |
| **Substitution** | reordering for more flexible structures<br>(spoken: 颜色 -不对了→written:不对了- 颜色) |
| | replace nouns by pronouns or their short forms<br>(spoken:太阳系的-星球→ written:太阳系的-这些) |
| | Informal/undecorated expressions in oral language<br>(spoken:把银河消失掉了→ written:遮住了银河) |
| **Deletion** | skips according to context (spoken:空间拍的-传回地球→ written:空间拍的-照片-传回地球) |

978-1-4799-4219-0/14/$31.00 ©2014 IEEE

## 3. Statistics on CCLR

### 3.1. Topic-related words and code-mixing

The annotated 98 lectures can be categorized into three topics as shown in Table 3. We can see the topics of the total lectures are generally balanced. We also calculate the percentage of the topic-related words and code-mixing rates in different topic categorizes. The topic-related words are defined as named entities and technical terms.

The results show the lectures about science and technology include a higher proportion of professional terms. The largest difference between these three topic categories are reflected on the foreign word rates; the rate in the science and technology topic is more than twice as much as those in the other two topics.

Table 3. *Distributions of topics*.

| Topics | #lectures | %Topic related words (Chinese) | %Foreign words |
|---|---|---|---|
| history/culture/art | 38 | 13.17% | 0.14% |
| society/economy/politics | 29 | 13.32% | 0.17% |
| science/technology | 31 | 17.33% | 0.39% |

### 3.2. Speaker distribution

For the annotated lectures by 90 speakers (21 female, 69 male), distribution of speakers' age and accent is listed in Table 4. When we annotate the accent type for these speakers, we follow the pronunciation rules summarized in [13]. Although all speakers have a high education background, accent still exists in 45% of them, especially for male speakers. Since accent could be an important factor in spontaneous speech, this statistics may give us some cues for developing acoustic modeling and the speaker adaptation strategy.

Table 4. *Distributions of speakers' ages and accents*.

| | 30≤age≤49 | | 50≤age≤69 | | age>70 | |
|---|---|---|---|---|---|---|
| | #female | #male | #female | #male | #female | #male |
| No accent | 4 | 15 | 10 | 15 | 1 | 4 |
| South accent | 2 | 9 | 3 | 16 | 0 | 4 |
| North accent | 0 | 3 | 1 | 3 | 0 | 0 |
| Total | 6 | 27 | 14 | 34 | 1 | 8 |

### 3.3. Speech rate and filler rate

Speech rate and filler (interjection) rate are two major factors closely related to the speaking style. We compared CCLR with other corpora such as Hub4 (broadcast news) and GALE (broadcast conversation) of Chinese as shown in Table 5. We can figure out the broadcast news (Hub4) has the lowest filler rate and moderate speech rate, because the speakers are professional narrators. And the broadcast conversation (GALE) has the highest filler rate and highest speech rate. This is probably because broadcast conversations are highly extemporaneous and less formal. The academic spoken lectures show an intermediate tendency, while the speech rate is comparable to that of the broadcast news (Hub4), the filler rate is comparable to that of the broadcast conversations (GALE). This suggests the speech is formal but spontaneous.

Table 5. *Speech rate and filler rate in different corpora.*

| Corpus | filler rate (interjection) | speech rate (words/minute) |
|---|---|---|
| Hub4 (broadcast news) | 1.33% | 159 |
| GALE (broadcast conversation) | 4.19% | 179 |
| CCLR (academic lectures) | 3.95% | 153 |

### 3.4. Disfluency edit

For 68 lectures that have both faithful transcripts and caption texts, an alignment of these parallel texts is conducted to get a detailed statistics on the disfluency edits as shown in Table 6.

In Table 7, we list percentages of the insertion and filler cases. We further break down the insertion case into discourse markers, grammatical particles and others. We find the interjections, discourse markers and grammatical particles together amount to approximately 8% in the transcripts. They are all irrelevant to the lecture's contents, and are omitted in the caption.

This result can be useful for developing specific language modeling and lightly-supervised acoustic model training [14].

Table 6. *Average disfluency edit rate% (word level)*.

| #Word | Substitution | Insertion | Filler | Deletion |
|---|---|---|---|---|
| 5677 | 1.73% | 6.81% | 3.60% | 1.07% |

Table 7. *Analysis of insertion and filler words*.

| Type | Percentage | Most frequent edit words (example) |
|---|---|---|
| Interjections | 3.60% | 啊, 呢, 呃, 吧, 嗯, 呀, 哎… |
| Discourse markers | 2.40% | 这个, 那么, 就是, 就, 那个… |
| Grammatical particles | 1.70% | 的, 是, 了, 在, 有, 也… |
| Others | 2.71% | 你, 我, 他, 我们, 这种, 什么… |

### 3.5. Part-of-Speech (POS) statistics in parallel text

We segment the text and analyze the Part-of-Speech (POS) frequency in the faithful transcripts and the caption texts. In Figure 1, we observe the difference in frequency of nouns and verbs. It suggests the majority of edits are related to these POS.

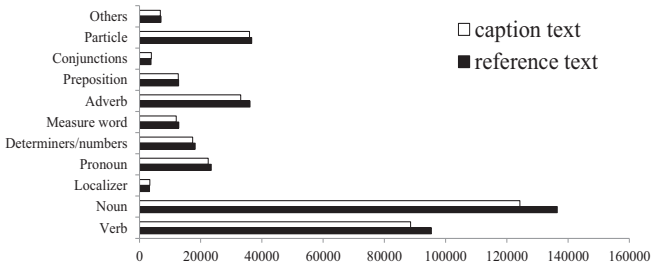


Figure 1: *Part-of-Speech (POS) statistics in parallel text*.

# 4. Transcription system for CCLR

## 4.1. Data sets

For the experimental purpose, we use 58 annotated lectures as the training set (CCLR-TRN), and 19 annotated lectures as the test set (CCLR-TST). We also have a large lecture data set (CCLR-LSV), which does not have faithful transcripts but has caption texts collected from the Internet. This data set can be used for enhancing the acoustic model. These data sets are listed in Table 8.

For all audio files, we conducted speech segmentation to the utterance unit (each one is less than 10sec) based on the BIC method [15] and speech clustering to remove non-speech segments and speech other than the main lecturer.

Table 8. *Organization of data sets*.

| | #Speaker | Duration |
|---|---|---|
| CCLR-TRN | 51 | 35.2 hours |
| CCLR-TST | 19 | 11.9 hours |
| CCLR-LSV | 126 | 62.0 hours |

## 4.2. Baseline system

### *4.2.1 Acoustic model*

The typical structure of Chinese syllables is: (C)+V(N)(R) [16], where C is an optional consonant, V is a vowel with 5 tones, N is an optional final nasal consonant, and R is an optional rhotic coda /r/ (we also consider it as a consonant). Usually a complete pinyin unit can be regarded as a syllable. According to this, we can separate each syllable (pinyin) to over 100 phoneme-like units [17] (87 tonal vowels and 24 consonants), and this method can make the monophone list compact.

We use 39-dimensional PLP features with CMN/CVN applied to each speaker for the baseline context-dependent GMM-HMM. The total number of the tied triphone states is fixed to 3000 and each state has 16 mixture components. Both MLE and MPE model are trained and compared.

### *4.2.2 Lexicon and language model*

From CCLR-TRN together with Hub4 and TDT4, we define a 53k dictionary and the OOV rate on CCLR-TST is 0.368%. Word pronunciations are derived from CEDICT open-source dictionary and HKUST dictionary. There are 1.7k English word entries and most of them are technical terms and persons' names. We prepared a set of phone-to-phone mapping from English phonemes to Mandarin phoneme-like units based on the approximate language transfer rules described in [18].

A word N-gram (3-gram) language model was trained. Since the training data size of CCLR is small, we incorporate other three corpora (Hub4, GALE, TDT4) distributed through LDC. They were interpolated to get the lowest perplexity on a development set, as shown in Table 9. The lecture corpus includes faithful transcriptions of CCLR-TRN and phoenix lecture texts[1].

[1]talk.ifeng.com

Table 9. *Component language models and their interpolated model.*

| | Corpora | #Words | PPlex. | Weights |
|---|---|---|---|---|
| Component Language Models | TDT4 | 4.75M | 1208 | 0.07 |
| | HUB4 | 0.34M | 1254 | 0.01 |
| | GALE | 1.03M | 519 | 0.36 |
| | Lecture | 1.07M | 451 | 0.56 |
| Interpolated Language Model | | 7.19M | 371 | / |

### *4.2.3 Baseline evaluation*

We use our own decoder Julius 4.3.1 [19]. ASR performance is evaluated on CCLR-TST. In preliminary experiments, we found that mismatch between the training and testing data seriously deteriorate the ASR performance and using other corpora do not have any effect. Therefore, we only use the in-domain data (CCLR-TRN) for acoustic model training. We trained both MLE and MPE models and they are compared in Table 10. We can see that CER is high (39.31%) with the basic MLE model. The MPE model gets a significant CER reduction of absolute 2.65%.

Table 10. *ASR performance (CER%)*.

| Data set | Durations (Hours) | MLE | MPE |
|---|---|---|---|
| CCLR-TRN | 35.2 | 39.31% | 36.66% |

## 4.3. Acoustic model refinement with deep neural network (DNN)

### *4.3.1 DNN acoustic model*

Since deep neural network (DNN) becomes a state-of-the-art acoustic modeling technique [20], we also trained DNN-HMM hybrid model using CCLR-TRN.

A DNN-HMM hybrid model is trained with the same PLP features as the baseline GMM-HMM, except that the features are globally normalized to have a zero mean and a unit variance. The input to DNN is 11 frames (5 frames on each side of the current frame) of the 39 dimensional features. The baseline MPE model is used to generate the state alignment label sequences. The network has 429 nodes as input, 3000 nodes as output and 6 hidden layers with 1024 nodes per layer.

In the unsupervised pre-training stage, we pool all of the data as the training data. And the network is initialized with stacked restricted Boltzmann machines (RBMs) that are pre-trained in a greedy layer-wise fashion. The Gaussian-Bernoulli RBM is trained with an initial learning rate of 0.01 and the Bernoulli-Bernoulli RBMs with a rate of 0.4. During pre-training, the momentum *m* is linearly increased from 0.5 to 0.9, which is accompanied by a rescaling of the learning rate using *1-m*. Also the L2 regularization is applied to the weights, with a penalty factor of 0.0002.

Then in the fine-tuning stage of frame-level cross-entropy training, we hold 1/8 of total utterances for cross validation and the other 7/8 of total utterances for supervised training. The utterance frames are presented in a randomized order while using SGD to minimize the cross-entropy between the supervision labels and network output. The SGD uses mini-batches of 256 frames, and an exponentially decaying schedule that starts with an initial learning rate of 0.01 and halves the rate when the improvement in the frame accuracy on the held-out set between two successive epochs falls below 0.5%. The

training terminates when the frame accuracy increases by less than 0.1%. We use single GPU (Tesla K20m) to accelerate the training time.

For decoding, we use Julius 4.3.1 (DNN version). It performs fast decoding with a pseudo-HTK format model and 3000-dimensional likelihood feature vectors generated by the DNN. The ASR performance (CER%) is 31.59% and an absolute reduction of 5.07% is achieved from the best MPE model in Sub-section 4.2.

### *4.3.2 Training with data with caption texts*

For further model improvement, we exploit the lecture data that do not have faithful transcripts but caption texts (CCLR-LSV) of 62.0 hours.

The pre-training step works in a totally unsupervised way, and only the back-propagation step needs supervision labels. Since the caption texts are not faithful, we use an ASR hypothesis generated from the baseline MPE model and a biased language model for each lecture as the training transcript [21]. The biased language model for each lecture is created by interpolating its closed-caption language model and the baseline language model with the weights 0.9 and 0.1. The experimental result in Table 11 shows the effectiveness of the data increase, although the transcripts are not faithful.

Table 11. *ASR performance (CER%) of DNN models*.

| Data set | Durations (Hours) | Ave. CER |
|---|---|---|
| CCLR-TRN | 35.2 | 31.59% |
| CCLR-TRN +CCLR-LSV (only with captions) | 97.2 | 28.80% |

### *4.3.3 Speaker adaptation on DNN model*

To further enhance the DNN model, we conduct unsupervised speaker adaptation by retraining the DNN for every speaker of the test set with its initial recognition hypothesis [22, 23]. Fine-tuning of DNN is conducted using the initial ASR hypothesis with a small learning rate. Once the model is updated, ASR is conducted again, and the adaptation process can be iterated. The experiment result in Table 12 shows that the iterative adaptation consistently gets small improvement on the CER. The best CER after three iterations of adaptation is 26.58%.

Table 12. *ASR performance (CER%) of adapted the lightly-supervised trained DNN model*.

| Without Adaptation | Unsupervised Adaptation | | |
|---|---|---|---|
| | 1st iter. | 2nd iter. | 3rd iter. |
| 28.80 | 26.99% | 26.70% | 26.58% |

## 5. Conclusions

The paper introduces our project on the spoken lecture corpus and the ASR system. We compile a corpus of Chinese Lecture Room (CCLR) of approximately 100 lectures. They are used to train a baseline ASR system. We also investigate lightly-supervised training that exploits data with caption texts. By using DNN-HMM acoustic modeling and the adaptation technique, we achieved CER of 26.58%.

## 6. References


[1] S. Furui and T. Kawahara. "Transcription and distillation of spontaneous speech," Springer Handbook on Speech Processing and Speech Communication, pp.627-651, 2008.

[2] J. Glass, T. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay. "Recent Progress in the MIT Spoken Lecture Processing Project," In Proc. INTERSPEECH, Antwerp, 2007.

[3] T. Kawahara. "Transcription system using automatic speech recognition for the Japanese Parliament (Diet)," In Proc. AAAI/IAAI, pp.2224-2228, 2012.

[4] K. Maekawa, "Corpus of Spontaneous Japanese: Its Design and Evaluation". In Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, pp. 7-12, 2003.

[5] I.Trancoso, R.Nunes, L.Neves, C.Viana, H.Moniz, D.Caseiro, and A.I.Mata, "Recognition of Classroom Lectures in European Portuguese". In Proc. INTERSPEECH, pp. 281-284, 2006.

[6] H. Chan, J. Zhang, P. Fung and L. Cao, "A Mandarin Lecture Speech Transcription System for Speech Summarization" in Proc. ASRU, 2007.

[7] C. Lin and L. Lee, "Improved Features and Models for Detecting Edit Disfluencies in Transcribing Spontaneous Mandarin Speech", IEEE Trans. on Audio, Speech and Language Processing, Vol 17, Issue 7, pp.1263-1278, September 2009.

[8] C. Chu, Y. Sung, Y. Zhao, and D. Jurafsky, "Detection of Word Fragments in Mandarin Telephone Conversation," In Proc. ICSLP, Pittsburg, 2006.

[9] S. Tseng, "Spoken Corpora and Analysis of Natural Speech," Taiwan Journal of Linguistics Vol. 6.2, pp.1-26, 2008.

[10] S. Tseng, "Repairs and repetitions in spontaneous Mandarin", in Proc DiSS'03, Disfluency in Spontaneous Speech Workshop, Göteborg University, Sweden, 5–8 September 2003.

[11] Y. Zhao, D. Jurafsky, "A preliminary study of Mandarin filled pauses", In Proc DiSS'05, Disfluency in Spontaneous Speech Workshop, Aix-en-Provence, France, 10-12 September 2005.

[12] A. Li, "Chinese prosody and prosodic labeling of spontaneous speech", In Proc. Speech Prosody. pp39-46. Aix-en-Provence, France, 2002.

[13] R. Sproat, T. F. Zheng, L. Gu, D. Jurafsky, I. Shanfran, J. Li, Y. Zheng, H. Zhou, Y. Su, S. Tsakalidis, P. Bramsen and D. Kirsch, "Dialectal Chinese speech recognition: Final technical report," Summer Workshop at CLSP/JHU Technical Report, 2004.

[14] T.Kawahara, M.Mimura, and Y.Akita, Language Model Transformation Applied to Lightly Supervised Training of Acoustic Model for Congress Meetings. In Proc. ICASSP, pp.3853-3856, 2009.

[15] M.Mimura, T.Kawahara, Fast Speaker Normalization and Adaptation Based on BIC for Meeting Speech Recognition, IEICE TRANSACTIONS on Information and Systems (Japanese Edition) vol.J95-D.

[16] X. Liu, J. L. Hieronymus, M. J. F. Gales and P. C. Woodland. "Syllable Language Models for Mandarin Speech Recognition: Exploiting Character Sequence Models, " Journal of the Acoustical Society of America, Vol. 133, Issue 1, pp.519-528, 2013.

[17] M. Hwang, W. Wang, X. Lei and et al. "Advances in Mandarin broadcast speech recognition". In Proc. INTERSPEECH, pp. 2613-2616, August 2007.

[18] S. Li, and L. Wang. "Cross Linguistic Comparison of Mandarin and English EMA Articulatory Data." In Proc. INTERSPEECH. 2012.

[19] A.Lee and T.Kawahara. Recent development of open-source speech recognition engine Julius. In Proc. APSIPA ASC, pp.131-137, 2009.

[20] G. Hinton, L. Deng, D. Yu, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". Signal Processing Magazine, IEEE, Vol 29, Issue 6, pp.82-97, 2012.

[21] L.Lamel, J.Gauvain, and G.Adda. "Investigating Lightly Supervised Acoustic Model Training". In Proc. ICASSP, pp. 477-480, 2001.

[22] Y.Xiao, Z. Zhang, S. Cai, J.Pan and Y.Yan, "A initial attempt on task-specific adaptation for deep neural network based large vocabulary continuous speech recognition," In Proc. INTERSPEECH, 2012.

[23] D.Yu, KYao, H.Su, G.Li, and F.Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," In Proc. ICASSP, pp.7893-7897, 2013.