SPEECH RECOGNITION ENHANCED BY LIGHTLY-SUPERVISED AND SEMI-SUPERVISED ACOUSTIC MODEL TRAINING

Sheng LI

Graduate School of Informatics

Kyoto University

ABSTRACT

Automatic transcription of lectures is one of the promising applications of automatic speech recognition (ASR), since captions to the lectures are needed not only for hearing-impaired persons but also for non-native viewers and elderly people. ASR is also useful for indexing the content. This work addresses effective acoustic model training targeted on Chinese spoken lectures. ASR of lectures has been investigated for almost a decade in many institutions world-wide, but there are still technically challenging issues for the system to reach a practical level. The biggest challenge is the limitation of training data.

In this work, a relatively small-sized database for Chinese spoken lectures with faithful transcripts is first compiled, but it is not sufficient for supervised training. On the other hand, there is huge amount of audio and video data of lectures with closed caption texts or without any related texts, which should be exploited to increase the training data. This thesis presents a progressive framework for acoustic model training by effectively incorporating speech data without faithful transcripts. Since the automatically generated label with a seed model will have a low accuracy, lightly-supervised training is introduced by leveraging closed caption texts. Then, semi-supervised training is reasonably adopted by incorporating unlabelled data. A novel discriminative approach is proposed to select reliable data in this framework. A dedicated set of classifiers are designed to select or verify the hypothesis from multiple ASR systems or the closed caption text.

Chapter 1 introduces the background, the problem, and the approaches addressed in the thesis. In Chapter 2, a review of speech recognition and deep neural network (DNN)-based acoustic model training is presented, and then the basic concept of lightly-supervised and semi-supervised training in the machine learning paradigm is introduced with related work.

Chapter 3 describes the corpus and the baseline system. For a comprehensive study on ASR of spontaneous Chinese, a corpus of Chinese Lecture Room (CCLR) is compiled. An overview of this corpus and some linguistic analysis are presented. Then, a baseline ASR system is developed based on GMM (Gaussian Mixture Model) and DNN using this corpus.

In Chapter 4, the proposed lightly-supervised acoustic model training with discriminative data selection from closed caption texts is explained. In the proposed method, a sequence of the closed caption text and that of the ASR hypothesis by the baseline system are aligned. Then, a set of dedicated classifiers based on CRF (Conditional Random Fields) is designed and trained to select the correct one among them or reject both. It is demonstrated that the classifiers can effectively filter the usable data for acoustic model training without tuning any threshold parameters. A significant improvement in the ASR accuracy is achieved from the baseline system and also in comparison with the conventional method of lightly-supervised training based on simple matching and confidence measure score (CMS).

In Chapter 5, the proposed semi-supervised acoustic model training with discriminative data selection from multiple ASR systems' hypotheses is described. In the proposed method, ASR hypotheses are obtained from complementary GMM and DNN based ASR systems. Then, a set of CRF-based classifiers are trained to select the better hypothesis and verify the selected data. The combined hypothesis for acoustic model training shows higher quality compared with the conventional system combination method (ROVER). Moreover, compared with the conventional data selection based on CMS, the method is demonstrated more effective for filtering usable data. A significant improvement in the ASR accuracy is achieved over the baseline system and in comparison with the models trained with the conventional system combination and data selection methods.

Chapter 6 concludes the thesis with a brief outlook of future work.

ACKNOWLEDGEMENTS

I firstly would like to acknowledge Professor Tatsuya Kawahara for his continued support of my research and life in Japan, without whom I would not have any chances to undertake research in spoken language processing. Before that, I only knew some old fashion of acoustic modelling. As a mentor, he has been a continual source of guidance, advice and inspiration and I could not have asked for a better supervisor.

I would also like to thank Professor Sadao Kurohashi and Professor Hisashi Kashima for giving their valuable time to reviewing this work and providing insightful comments.

Of course this work could not have been achieved without the efforts of the Group of Media Archiving Lab, in particular the senior members: Associate Professor Shinsuke Mori, previous Assistant Professor Yuya Akita, visiting Professor Nigel Ward and Mr. Masato Mimura and Dr. Shinsuke Sakai. Furthermore, I would like to acknowledge all past and current members of the lab.

I especially thank my research co-operators in NICT: Dr. Xugang Lu, Dr. Xinhui Hu, Dr. Naoyuki Kanda, Dr. Pen Shen, Dr. Jinfu Ni and former researcher Dr. Shigeki Matsuda and former director Chiori Hori. I especially thank Dr. Xugang Lu, an excellent mentor and researcher, for kindly sharing his precious ideas and opening his private time to me for discussion almost anytime, and also Dr. Xinhui Hu for all of his kind help to me on both research and engineering.

I should acknowledge the institutions, which made this work possible. The Ministry of Education, Culture, Sports, Science and Technology of Japan provided me with a full scholarship to undertake this research. Of course, I should also thank Kyoto University for all their administrative support.

And I cannot forget the reference letters and/or kind helps from Dr. Yu Qiao, Dr. Lan Wang, Dr. Dean Luo, Dr. Yongchuan Li of Chinese Academic of Sciences and Professor Haoran Wang and Mr. Timing Han of Nanjing University when I applying this top university in Japan three years ago. And I also thank department of Computer Science of Nanjing University and my former laboratory (AIMSL) in Chinese Academic of Sciences for giving me enough academic training in computer science and speech recognition.

At last, I would like to dedicate this work to my parents. They are always my strongest support. Finally, I want to say to my grandparents, who passed away during my PhD program: Wish you peace in Heaven.

CONTENTS

1 INTRODUCTION 1

1.1 RESEARCH BACKGROUND 1

1.2 CHALLENGES OF THESIS WORK 2

1.2.1 Spoken Lectures as the Task Domain 2

1.2.2 Training Data is Key for Acoustic Model Training 2

1.3 PROPOSED FRAMEWORK 3

1.3.1 Resources and Related Acoustic Model Training Scheme 3

1.3.2 Progressive Framework for Acoustic Model Training 4

1.3.3 Discriminative Data Selection Approach 5

1.3.4 Overview of Proposed Training Flow 5

1.4 ORGANIZATION OF PROPOSED THESIS 7

2 REVIEW OF ACOUSTIC MODEL TRAINING FOR SPEECH RECOGNITION 9

2.1 SPEECH RECOGNITION 9

2.1.1 ASR Framework 9

2.1.2 ASR with DNN-HMM Hybrid Architecture 14

2.1.3 ASR with Other Architectures 15

2.2 DNN-BASED ACOUSTIC MODEL TRAINING 16

2.2.1 Feed Forward Computation of DNN 17

2.2.2 Back Propagation Training of DNN 18

2.2.3 Enhanced DNN-based Acoustic Model Refinement 20

2.3 ACOUSTIC MODEL TRAINING WITH LIGHTLY-SUPERVISED AND SEMI-SUPERVISED TRAINING 21

2.3.1 Lightly-supervised Training 21

2.3.2 Related Work of Lightly-supervised Acoustic Model Training 22

2.3.3 Semi-supervised Training 22

2.3.4 Related Work of Semi-supervised Acoustic Model Training 23

2.3.5 Conventional ROVER Combination 25

3 CORPUS AND TRANSCRIPTION SYSTEM OF CHINESE LECTURES 27

3.1 REVIEW ON LECTURE TRANSCRIPTION PROJECTS 27

3.2 CONSTRUCTION OF CHINESE CORPUS OF LECTURE ROOM (CCLR) 30

3.2.1 Background Knowledge for Chinese Acoustic Modelling 30

3.2.2 Corpus Description *34*
3.2.3 Annotation Scheme *34*
3.2.4 Statistics on CCLR *34*
3.3 BASELINE ASR SYSTEM 39
3.3.1 Data Sets *39*
3.3.2 Baseline System *40*
3.3.3 Speaker Adaptation on DNN model *43*
3.3.4 Speaker Adaptive Training on DNN model *43*
3.4 CONCLUSION 43
4 LIGHTLY-SUPERVISED ACOUSTIC MODEL TRAINING 45
4.1 INTRODUCTION 45
4.2 A TWO-STEP CRF-BASED CLASSIFICATION SCHEME FOR DATA SELECTION 46
4.2.1 Proposed Lightly-supervised Training Framework *46*
4.2.2 Category of Word Alignment Patterns *47*
4.2.3 Cascaded Classifiers for Word-level Data Selection *48*
4.2.4 Feature Set Design for Classifiers *49*
4.2.5 Utterance Selection for Acoustic Model Training *51*
4.3 EXPERIMENTAL EVALUATIONS 51
4.3.1 Classifier Implementation and Performance *51*
4.3.2 Utterance Selection for Model Training *53*
4.3.3 ASR Performance with Enhanced Model Training *55*
4.4 CONCLUSION 56
5 SEMI-SUPERVISED ACOUSTIC MODEL TRAINING 59
5.1 INTRODUCTION 59
5.2 COMPARATIVE ANALYSIS OF DIFFERENT BASELINE SYSTEMS 60
5.3 CRF-BASED HYPOTHESIS COMBINATION AND DATA SELECTION FOR DNN MODEL TRAINING 62
5.3.1 Proposed Framework *62*
5.3.2 Categories of Alignment Patterns *63*
5.3.3 Classifier Design *63*
5.3.4 Feature Design *65*
5.3.5 Data Selection for Acoustic Model Training *67*
5.4 EXPERIMENTAL EVALUATIONS 69
5.4.1 Classifier Implementations *69*
5.4.2 Classification Accuracy of CRF Classifiers *72*

5.4.3 Performance of Hypothesis Selection and Verification *73*
5.4.4 Performance of DNN Acoustic Model Enhanced by Selected Data *75*
5.5 CONCLUSION 78
6 CONCLUSIONS 79
6.1 CONTRIBUTION OF THESIS WORK 79
6.2 SUMMARY OF PERFORMANCE IMPROVEMENT 80
6.3 FUTURE WORK 81
7 REFERENCES 83
8 APPENDICES 97
APPENDIX 1: SPEAKER ADAPTIVE TRAINING DNN 98
A1.1. ENSEMBLE SPEAKER MODELLING USING SPEAKER ADAPTIVE TRAINING DNN ... 98
A1.1.1 Multi-task Learning Architecture for SAT-DNN *99*
A1.1.2 Ensemble Speaker Matrix Factorization *99*
A1.1.3 Low-rank Matrix Approximation *101*
A1.1.4 Adaptation on SAT-DNN Ensemble Models *101*
A1.2 IMPLEMENTATION AND EVALUATIONS 104
APPENDIX 2: TRAINING CRF WITH PARTIALLY ANNOTATION 107
A2.1 TRAINING A CRF WITH PARTIALLY ANNOTATED DATA 107
A2.1.1 Full and Partial Annotations *107*
A2.1.2. CRF-based Confidence Estimation with Full Annotation *108*
A2.1.3. Train CRF Model with Partial Annotations *109*
LIST OF PUBLICATIONS BY THE AUTHOR 111

LIST OF TABLES

TABLE 1.1 LIGHTLY-SUPERVISED AND SEMI-SUPERVISED TRAINING. 3

TABLE 3.1 SOME BENCHMARK SYSTEMS BEFORE 2014 29

TABLE 3.2: MAPPING FROM ENGLISH CONSONANTS MISSING IN MANDARIN TO MANDARIN EXISTING EQUIVALENTS (CITED FROM [110]) 33

TABLE 3.3: MAPPING FROM ENGLISH VOWELS MISSING IN MANDARIN TO MANDARIN EXISTING EQUIVALENTS (CITED FROM [110]) 33

TABLE 3.4 BASIC CORPUS DESCRIPTION. 34

TABLE 3.5 MAJOR SPONTANEOUS PHENOMENA. 35

TABLE 3.6 ACOUSTIC CONDITIONS FOR LECTURE RECORDINGS. 35

TABLE 3.7 DISTRIBUTIONS OF TOPICS. 36

TABLE 3.8 DISTRIBUTIONS OF SPEAKERS' AGES AND ACCENTS. 36

TABLE 3.9 SPEECH RATE AND FILLER RATE IN DIFFERENT CORPORA. 37

TABLE 3.10 AVERAGE DISFLUENCY EDIT RATE (WORD LEVEL). 37

TABLE 3.11 ANALYSIS OF INSERTION AND FILLER WORDS. 38

TABLE 3.12 TOP 50 MOST FREQUENT EDIT DISFLUENCY RELATED WORDS. (THE NUMBER IN THE PARENTHESIS IS TOTAL OCCURRENCE IN THE LABELED DATA) 38

TABLE 3.13 ORGANIZATION OF DATA SETS. 39

TABLE 3.14 COMPONENT LANGUAGE MODELS AND THEIR INTERPOLATED MODEL. 40

TABLE 3.15 ASR PERFORMANCE (CER%) OF GMM MODELS. 41

TABLE 3.16 ASR PERFORMANCE (CER%) OF DNN MODELS. 42

TABLE 3.17 PERFORMANCE (CER%) OF SPEAKER ADAPTATION. 43

TABLE 4.1 CATEGORY OF ALIGNMENT PATTERNS (WORD LEVEL). 47

TABLE 4.2 FEATURE SET FOR CLASSIFICATION. 50

TABLE 4.3 FEATURE EVALUATION OF CRF-1 BY 5-FOLD CROSS VALIDATION (ON CCLR-SV) 53

TABLE 4.4 FEATURE EVALUATION OF CRF-2 BY 5-FOLD CROSS VALIDATION (ON CCLR-SV). 53

TABLE 4.5 ASR PERFORMANCE (CER%) BY LIGHTLY-SUPERVISED TRAINED DNN ACOUSTIC MODEL (FBANK). 56

TABLE 4.6 ASR PERFORMANCE (CER%) BY LIGHTLY-SUPERVISED TRAINED DNN ACOUSTIC MODEL (PLP)..... 56

TABLE 5.1 COMPONENT LANGUAGE MODELS AND THEIR INTERPOLATED MODEL. 60

TABLE 5.2 ASR PERFORMANCE ON CCLR-DEV. 61

TABLE 5.3 PAIR-WISE EDIT DISTANCE OF ASR RESULTS ON CCLR-DEV (CHARACTER LEVEL) 61

TABLE 5.4 CATEGORY OF ALIGNMENT PATTERNS. 63

TABLE 5.5 FEATURE DESIGN FOR CRF-1 65

TABLE 5.6 FEATURE DESIGN FOR CRF-2 66

TABLE 5.7 TRAINING DATA OF CRF-1 71

TABLE 5.8 TRAINING DATA OF CRF-2 71

TABLE 5.9 FEATURE SET EVALUATION OF CRF-1 ON CCLR-DEV. 72

TABLE 5.10 FEATURE SET EVALUATION OF CRF-2 ON CCLR-DEV. 73

TABLE 5.11 EVALUATION OF THE DATA SELECTION AND VERIFICATION..... 75

TABLE 5.12 ASR PERFORMANCE (CER%) OF CROSS-ENTROPY DNN MODEL BY UTTERANCE-LEVEL SELECTION 76

TABLE 5.13 ASR PERFORMANCE (CER%) OF CROSS-ENTROPY DNN MODEL BY FRAME-LEVEL SELECTION. 77

TABLE 5.14 ASR PERFORMANCE (CER%) OF SMBR DNN MODEL. 77

TABLE 8.1 ADAPTATION PERFORMANCES (CER% ON TST). 105

LIST OF FIGURES

FIGURE 1.1: CLOSED CAPTION AND FAITHFUL TRANSCRIPT IN A LECTURE PROGRAM. 4

FIGURE 1.2: OVERVIEW OF PROPOSED FRAMEWORK. 5

FIGURE 2.1: TYPICAL FRAMEWORK OF ASR 10

FIGURE 2.2: HMM-GMM HYBRID ACOUSTIC MODEL 12

FIGURE 2.3: DNN-HMM HYBRID ARCHITECTURE IN ASR..... 14

FIGURE 2.4: A SINGLE NEURON IN DEEP NEURAL NETWORK..... 17

FIGURE 2.5: SEQUENCE DISCRIMINATIVE TRAINING DNN MODELS. 20

FIGURE 2.6: CONVENTIONAL LIGHTLY-SUPERVISED ACOUSTIC MODEL TRAINING 22

FIGURE 2.7: CONVENTIONAL SEMI-SUPERVISED ACOUSTIC MODEL TRAINING 23

FIGURE 3.1: FIVE TONAL PATTERNS OF STANDARD MANDARIN LANGUAGE 30

FIGURE 3.2: MAIN DIALECTS AND THEIR GEOGRAPHIC DISTRIBUTIONS 31

FIGURE 3.3: HOW TO DERIVE (C)+V(N)(R) STRUCTURE OF CHINESE SYLLABLES. 32

FIGURE 3.4: PART-OF-SPEECH (POS) STATISTICS IN PARALLEL TEXT. 39

FIGURE 4.1: PROCESS FLOW OF PROPOSED LIGHTLY-SUPERVISED TRAINING..... 47

FIGURE 4.2: CASCADED CLASSIFICATION SCHEME FOR DATA SELECTION. 49

FIGURE 4.3: ASR PERFORMANCE (GMM-HMM ON CCLR-DEV) FOR DIFFERENT PA THRESHOLD VALUES. 54

FIGURE 4.4: ASR PERFORMANCE (GMM-HMM ON CCLR-DEV) FOR DIFFERENT CMS THRESHOLD VALUE. 54

FIGURE 5.1: PROCESS FLOW OF THE PROPOSED SEMI-SUPERVISED TRAINING. 62

FIGURE 5.2: CASCADED CLASSIFICATION SCHEME FOR DATA SELECTION. 65

FIGURE 5.3: FRAME-LEVEL DATA SELECTION METHODS..... 69

FIGURE 5.4 EXTRACTION OF PARTIALLY ANNOTATED DATA BY VOTING 70

FIGURE 6.1: PERFORMANCE IMPROVEMENT BY THE PROPOSED TRAINING METHOD. 80

FIGURE 8.1: MULTI-TASK LEARNING ARCHITECTURE FOR SAT-DNN. 99

FIGURE 8.2: BEFORE (LEFT) AND AFTER (RIGHT) MATRIX FACTORIZATION IN ONE LAYER OF DNN..... 100

FIGURE 8.3: DECOMPOSITION TO $\mathbf{W}_{sd}^{best}$ FOR SPEAKER COEF-MATRIX ADAPTATION..... 102

FIGURE 8.4: DECOMPOSITION TO $\mathbf{W}_{sd}^{best}$ FOR SINGULAR VALUES ADAPTATION. 103

FIGURE 8.5 EXAMPLE OF FULLY ANNOTATED UTTERANCE. 108

FIGURE 8.6 EXAMPLE OF PARTIALLY ANNOTATED UTTERANCE. 108

1 INTRODUCTION

1.1 Research Background

In the recent big data era, there are a huge number of audio and video lectures such as TED and MOOCs accumulating day by day. They need to be transcribed for digital archiving. Moreover, the audio-visual material is not easy to search and browse without time-aligned caption texts. The caption texts are necessary not only for hearing-impaired persons but also for non-native viewers and elderly people.

For this reason, automatic transcription of lectures, as one of the promising applications of automatic speech recognition (ASR), has a great strategic value. Spoken lectures are a kind of spontaneous speech. One of the most fundamental problems in training an acoustic model for this kind of spontaneous speech is the insufficient amount of training data to cover wide variation of the acoustic and linguistic features. It is very difficult and costly to prepare a large speech corpus, because it involves manual transcription of utterances with many disfluencies, compared to the reading of prepared texts.

The thesis addresses this problem, and presents effective and efficient approaches of acoustic model training, which do not require faithful transcription for generating training labels. While the semi-supervised training approach exploits unlabelled data, the lightly-supervised training makes use of related texts such as closed captions often provided with video lectures. They are effectively combined for building a high-performance automatic transcription system of spoken lectures.

1.2 Challenges of Thesis Work

1.2.1 Spoken Lectures as the Task Domain

ASR of lectures has been investigated for almost a decade in many institutions world-wide [1][2][3][4][5][6][7][74][75], but there are still technically challenging issues for the system to reach a practical level. Unlike interactions of multiple speakers such as interview, dialogue, debates and meetings, the spoken lecture is monologue. Compared to extemporaneous presentations, the spoken lecture is a kind of formal presentation. However, the natural or spontaneous speech recognition is much different from conventional speech recognition tasks such as dictations and broadcast news, because of the following aspects [71].

- Spontaneous speech is not reading prepared script and given to live audiences, so it has more colloquial expressions.
- Since speakers are not professional narrators, there are more frequent fillers and self-repairs in their speech. Therefore, the spontaneous speech has more disfluencies than the read speech.
- Spontaneous speech also involves larger variation from orthodox pronunciation including accents. These lead to more acoustic variations.

There are also differences in recording quality depending on whether the speaker wore a close-talking headset or a lapel microphone. The academic lectures are usually recorded with a lapel microphone worn by a lecturer.

Report of the DARPA project [16] showed the ASR performance gaps of different (NIST) benchmarks. Although there is no speech recognition tasks on spoken lectures, the spontaneous speech (conversation) has always a higher error rate than the read speech (broadcasting news) indicating the difficulty of this task empirically.

1.2.2 Training Data is Key for Acoustic Model Training

Training data is a key for acoustic model training in both quantity and quality.

Acoustic model training requires large quantities of data. For this thesis task, moreover, works on Chinese lectures or spontaneous speech in general are very limited [74][75]. There is no large corpus publicly available in this category. So the most serious problem is the limitation of training data. It is easy to collect data, but very costly to transcribe them. Since the spontaneous speech used in this thesis work

includes many disfluencies, fillers and colloquial expressions, the annotators will spend much time in transcribing them. For some technical topics and accented speeches, only the specialists can annotate. These reasons make the transcribed data very limited for the high expense.

Training acoustic model also requires high quality data. Discriminative training of GMM model is reported in [8] [124] very sensitive to the accuracy of the transcriptions. The accuracy of the lattices will become very poor when using unfaithful transcriptions. The gains obtained from discriminative training then decrease dramatically. Training DNN acoustic model is also sensitive to the noise in the training label, especially with sequence discriminative training [11][41][42].

1.3 Proposed Framework

1.3.1 Resources and Related Acoustic Model Training Scheme

It is feasible and necessary to compile a relatively small-sized database, but it will not be enough for a high performance ASR system. On the other hand, there is huge amount of audio and video data of lectures with closed caption texts or without any related texts. It is critical to exploit these resources to increase the training data for an acoustic model.

Table 1.1 Lightly-supervised and Semi-supervised Training.

	Supervised Training	Lightly-supervised Training	Semi-supervised Training
Closed Captions	×	√	×
Manual Transcriptions	√	×	limited
Audio	√	√	√

Table 1.1 summarizes the training schemes using a variety of resources.

Lightly-supervised acoustic model training, which does not require faithful transcripts but exploits available verbatim texts, has been explored for broadcast news [32][119][120] and parliamentary meetings [73]. This work focuses on lightly-supervised training which uses closed caption texts.

It is observed that, even given a caption text, a lot of work is needed to make a faithful transcript because the caption text is much different from what is actually spoken, and phenomena of spontaneous speech such as fillers and repairs need to be edited (see an example in Figure 1.1).

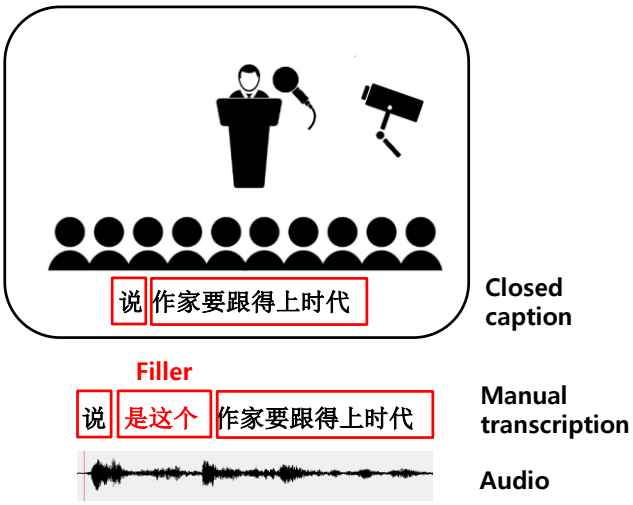


Figure 1.1: Closed Caption and Faithful Transcript in a Lecture Program.

Semi-supervised acoustic model training [8][9][10][11][12][13] is also designed by enhancing the limited labelled data with huge quantity of unlabelled data on holding. The automatically generated transcripts (ASR result) can be used as label for the latter. This scheme solves the quantity issue of the training data, but the quality of automatically generated transcripts is important.

1.3.2 Progressive Framework for Acoustic Model Training

In order to enhance the quantity of the quantity of the training data efficiently, this thesis presents a framework for acoustic model training by progressively applying the above-mentioned training schemes. The limited annotated data are used to train a seed acoustic model by the conventional supervised training. Considering the automatically generated label will have a low accuracy when directly using the seed acoustic model, a lightly-supervised training is introduced to derive an enhanced seed acoustic model. Finally, semi-supervised training is reasonably adopted for further improvement. Since the conventional lightly-supervised and semi-supervised acoustic model training have their limitations, this thesis investigates novel approaches for improving these two training schemes.

1.3.3 Discriminative Data Selection Approach

In order to enhance the quality of training data, this work integrates diverse resources such as closed captions with an ASR hypothesis in lightly-supervised training or multiple ASR hypotheses in semi-supervised training. Discriminative classifiers are designed to select one of them or reject both. The data selection problem is solved by using a cascade of CRF (conditional random fields) classifications. The CRF-based classifiers are prepared for two sub-tasks: selector CRF and verifier CRF. The selector CRF is trained to select a correct (or better) hypothesis from the aligned hypotheses. The verifier CRF is then used to determine whether the selected result is reliable or not. Data selection for acoustic model training is conducted according to the verification result.

1.3.4 Overview of Proposed Training Flow

Figure 1.2 illustrates an overview of the proposed training process flow.

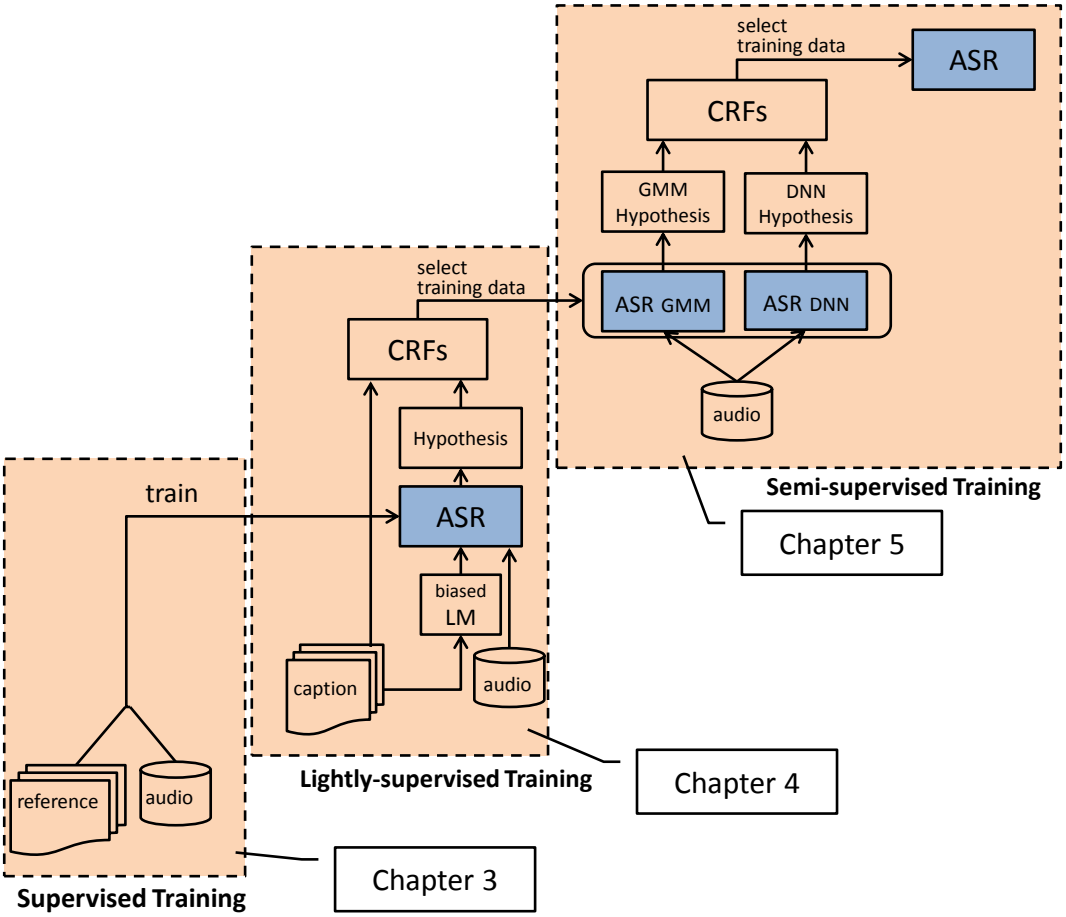


Figure 1.2: Overview of Proposed Framework.

A seed acoustic model is supervised trained with limited faithful data. Then in the proposed lightly-supervised training, ASR hypotheses are generated using this seed model. CRF classifiers are prepared to select training data to enhance the seed model given the alignment patterns of the ASR hypotheses and closed caption texts. In the next step of the proposed semi-supervised training, similar CRF classifiers are prepared to select training data given the alignment patterns of the GMM-based and DNN-based complementary hypotheses. The acoustic model is further enhanced.

1) Lightly-supervised Acoustic Model Training with Discriminative Data Selection from Closed Caption Texts

In the conventional lightly-supervised method, a biased language model is constructed based on the closed caption of the relevant program to guide the baseline ASR system to decode the audio content. Then, reliable segments of the ASR output are filtered for acoustic model training, usually by matching it against the closed caption.

The conventional filtering method of lightly-supervised training has a drawback that it significantly reduces the amount of usable training data. Moreover, it is presumed that the unmatched or less confident segments of the data are more useful than the matched segments because the baseline system failed to recognize them and may be improved with additional training [8].

A novel method is presented in this thesis to solve this problem. A sequence of the closed caption text and that of the ASR hypothesis by the baseline system are aligned. Then, a set of dedicated classifiers is designed and trained to select the correct one among them or reject both.

2) Semi-supervised Acoustic Model Training by Discriminative Data Selection from Multiple ASR Systems' Hypotheses

For the semi-supervised acoustic model training, unlabelled data are first transcribed with a seed model. The automatically generated transcripts (or combined ASR result) are used as a label. Then the seed model is retrained by adding the newly transcribed data to the existing labelled data.

Taking use of the unlabelled data without data selection will make the model training less effective. As mentioned in Subsection 1.2.2, the DNN model training is sensitive to the noise in the training label, especially in the sequence discriminative

training. The most commonly used method is based on the word-level confidence measure score (CMS) which is averaged over the utterance unit for data selection. However, the accuracy of the automatically generated transcript is low and data selection based on CMS is not so effective.

To solve the problem, a novel method is proposed. The ASR hypotheses by complementary GMM and DNN based ASR systems are first obtained. Then, a set of CRF-based classifiers are trained to select the correct hypotheses and verify the selected data.

1.4 Organization of Proposed Thesis

This thesis is organized as follows: Chapter 2 first gives a review of speech recognition and DNN-based acoustic model training, and then the basic concept of lightly-supervised and semi-supervised training. In Chapter 3, the background knowledge of Chinese acoustic modelling, the Corpus of Chinese Lecture Room and the baseline system are described. In Chapter 4, the proposed lightly-supervised acoustic model training with discriminative data selection from closed caption texts is explained in detail. In Chapter 5, the proposed semi-supervised acoustic model training with discriminative data selection from multiple ASR systems' hypotheses is explained. Finally, this thesis is concluded with a brief outlook of future work in Chapter 6.

2 REVIEW OF ACOUSTIC MODEL TRAINING FOR SPEECH RECOGNITION

In this chapter, a review of speech recognition and DNN-based acoustic model training is presented, and then the basic concept of lightly-supervised and semi-supervised training in the machine learning paradigm is introduced. Finally, the related work on their application to acoustic model training is described.

2.1 Speech Recognition

Automatic speech recognition (ASR) is a dynamic classification problem, which attempts to find a sequence of words corresponding to a given sequence of speech signal. A series of technical elements are needed to achieve this goal.

2.1.1 ASR Framework

In the fundamental principle, a word sequence (w) is determined with a maximized posterior probability $P(w|x)$ given the input speech feature (x). As it is difficult to compute the posterior probability $P(w|x)$ directly, Equation 2.1 can be derived according to Bayes's rule.

$$w' = \arg\max P(w \mid x) = \arg\max_{w} P(x \mid w) P(w) \tag{2.1}$$

where $P(x|w)$ is a likelihood of x for a given w and $P(w)$ is a prior probability of w.

If there are multiple pronunciations p of sentence w, $P(x|w)$ can be calculated with Equation 2.2.

$$P(x \mid w) = \sum_{p} P(x \mid p) P(p \mid w) \tag{2.2}$$

where $P(p|w)$ is a probability that w is pronounced as p and $P(x|p)$ is a likelihood of x for a given p. When only the most likely pronunciation is considered, Equation 2.1 is rewritten into Equation 2.3.

$$w' = \arg\max_{w,p} P(x \mid p) P(p \mid w) P(w) \tag{2.3}$$

In many cases or when there is only one pronunciation entry for w, we regard $P(p|w)=1$.

The components of the ASR system are illustrated in Figure 2.1.

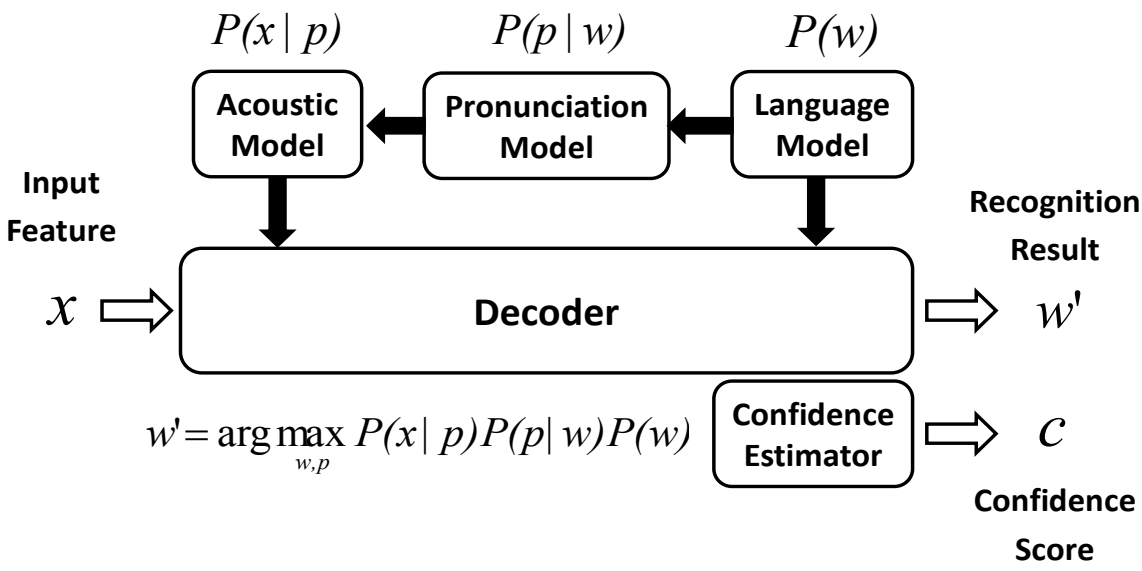


Figure 2.1: Typical Framework of ASR

According to Equation 2.3 and Figure 2.1, a typical ASR system consists of following components:

- **Input Speech Feature:** x.
- **Acoustic Model:** compute the likelihood $P(x|p)$ based on pattern recognition.
- **Language Model:** compute the prior probability $P(w)$ based on generative model.
- **Pronunciation Model (Lexicon)**.
- **Decoder:** searches sentence for the best hypotheses using Equation 2.3.
- **Confidence Estimator:** compute the score (c) within the range of [0,1] indicating the reliability of the recognition result (w').

2.1.1.1 Acoustic Features for Speech Recognition

The acoustic features most commonly used in GMM-based speech recognition system are Mel-frequency cepstrum coefficients (MFCC) [85] and peceptal linear prediction (PLP) [86] features. MFCC is extracted from the frequency domain using the Mel scale, which is based on the human perception. Therefore, it can extract parameters from the speech similar to ones that are used by humans for hearing speech, while at the same time, it deemphasizes all other information. PLP is determined by auto-regression model based on the concept of psychophysics of hearing [86].

Although PLP is derived independently of the MFCC technique, there are many similarities between the two methods. PLP features are reported [87] to be more robust when there is an acoustic mismatch between training and test data. Under clean conditions and when there is no significant mismatch, MFCC features lead to a performance that is slightly superior to PLP [88].

In the DNN-based speech recognition system, the entire DNN model can be regarded as a feature extractor. The filterbank feature [142] is most commonly used as input of the DNN model. It is feed-forwarded through DNN model and then transformed into the log-posterior features for decoding.

2.1.1.2 Acoustic Model

An acoustic model evaluates an emission probability $P(x|p)$ of a sequence of features x for a certain symbol p. Subword units of speech such as phones, initial-finals and syllables are normally used for symbols.

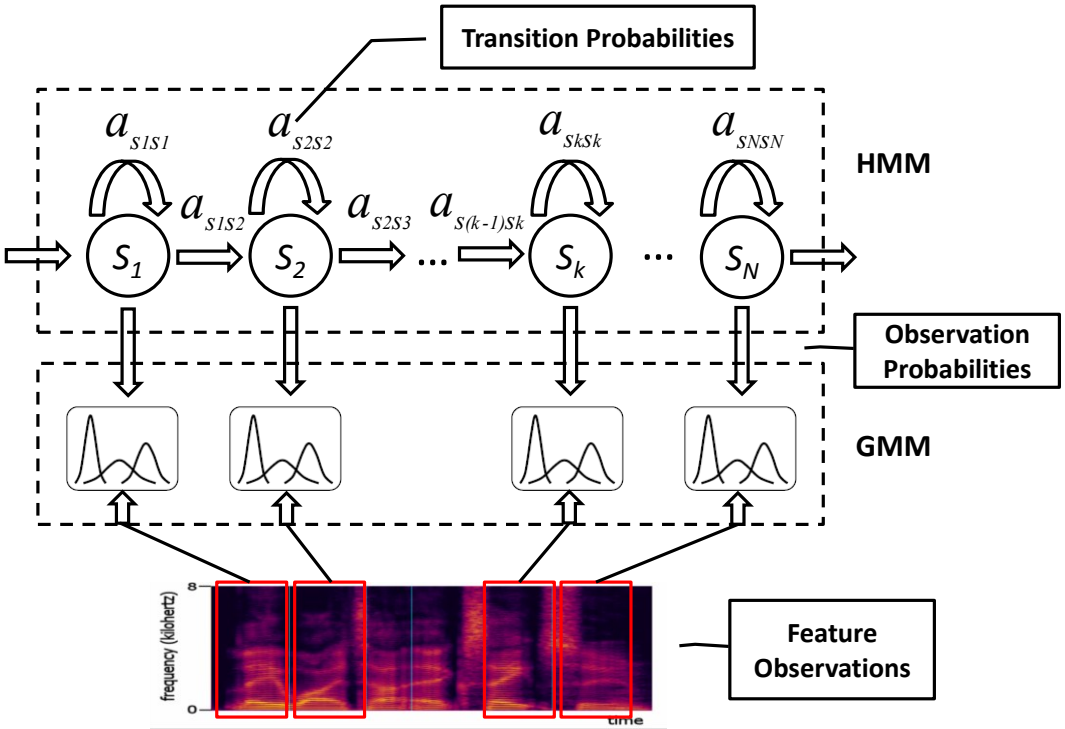


Figure 2.2: HMM-GMM Hybrid Acoustic Model

Conventional GMM-HMM hybrid acoustic model is shown in Figure 2.2. Hidden Markov model (HMM) has the several states with the typical left-to-right structure including self-transitions. Transition probabilities between states are defined and observation probabilities are computed for input features in each state. The observation probability desity is usually represented by Gaussian mixture models (GMM). For a sequence of input features, the resulting probability is calculated by multiplying transition probabilities and observation probabilities.

When phone is used as the subword unit, each phone is modelled with a HMM. By considering its contexts such as preceding and following phones, the context-dependent (CD) HMM can be defined. Triphone is commonly adopted. Moreover, the state of HMMs can be shared to reduce the parameter space of CD triphones. Thus, the emission probability is defined the tied triphone for states.

2.1.1.3 Language Model

A language model provides a linguistic constraint for a word sequence. Framework of language model is roughly classified into several categories: grammar-based language model [89], statistical n-gram language model [90] and neural network based language model [91]. Only the grammar-based language model and statistical n-gram language model can be used in the decoding process. The neural network based language model can only be used for re-scoring of the initial decoding result.

2.1.1.4 Pronunciation Model

A pronunciation model defines pronunciation of words as sequences of subword units. Take the most commonly used subword unit phone as an example, the pronunciation model gives each word entry a sequences of phones. Some word entries have more than one pronunciation entries. As shown in Equation 2.2, the actual pronunciation will be automatically determined by Viterbi algorithm [92]. The probability is calculated from the corpus annotated with pronunciation.

2.1.1.5 Decoder

Decoder is the most complex component of speech recognition. There are many algorithms and features. Take "Julius" decoder [47] as an example. It is a two-pass large vocabulary continuous speech recognition (LVCSR) decoder. Based on word N-gram and context-dependent HMM, it can perform almost real-time decoding on most current PCs in 60k word dictation task. Major search techniques are fully incorporated such as tree lexicon, N-gram factoring, cross-word context dependency handling, enveloped beam search, Gaussian pruning, Gaussian selection. Besides search efficiency, it is also modularized carefully to be independent from model structures, and various HMM types are supported such as shared-state triphones and tied-mixture models.

Other famous decoders are HTK [95], Sphinx [93], RASR [94], Jucier [100], Kaldi [46], T^3 [101] and SprinTra [102]. The last four decoder implements WFST [96][97] based decoding scheme which is now state-of-the-art for LVCSR systems.

2.1.1.6 Confidence Estimator

Together with the recognition result (w'), the ASR system can also obtain a confidence score for each word with a value [0,1] indicating the reliability of the recognition result. The most commonly used method for generating the confidence measure scores (CMS) is approximation of posterior probabilities over a lattice and its

variants [22][23][24][25][26][27][28]. In recent years, discriminative models such as conditional random fields (CRF) models [29], which can combine multiple sources such as acoustic, lexical and linguistic features with contextual information, are used for confidence estimation [30][31]. DNN-based method is also introduced to estimate the CMS [125].

2.1.2 ASR with DNN-HMM Hybrid Architecture

There are several methods to apply the neural networks to speech recognition. Now the most widely used method is called "DNN-HMM hybrid architecture", which combines a DNN with HMMs in a framework as illustrated in Figure 2.3. In this framework, the HMMs capture the temporal dynamics of the speech signal and the DNN estimates the observation probabilities given the acoustic observations. Each output neuron of the DNN corresponds to the tied triphone state.

The architecture in Figure 2.3 makes it possible to reuse the mechanism of decoding GMM-HMM.

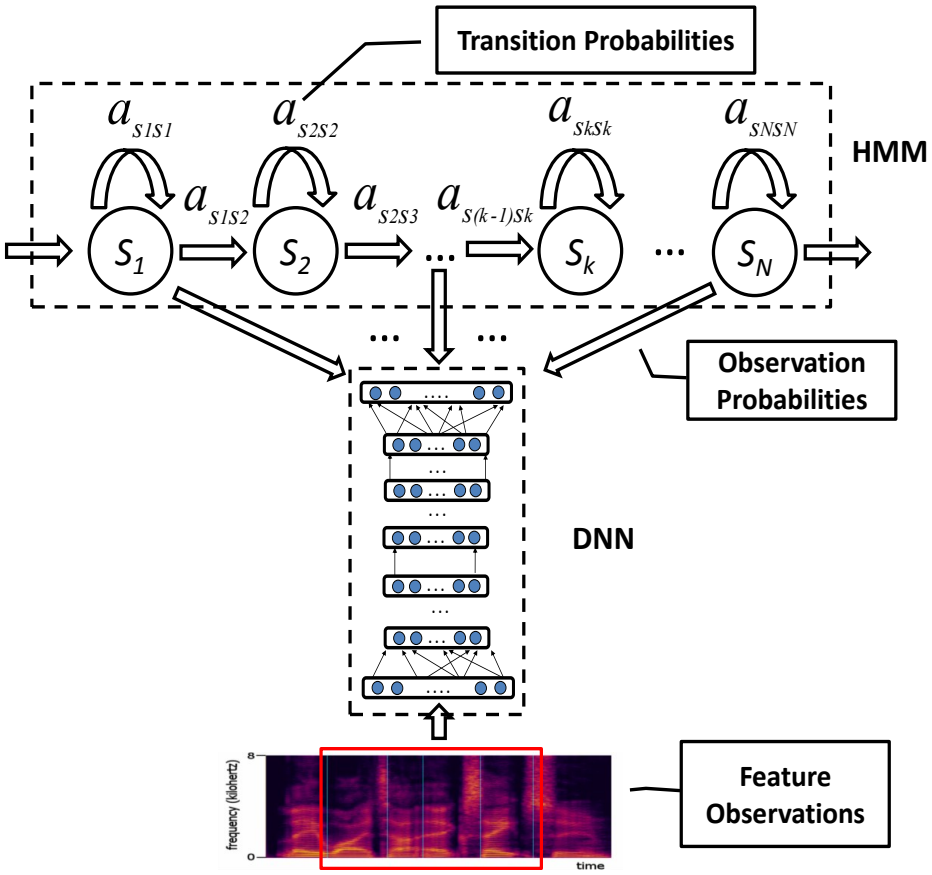


Figure 2.3: DNN-HMM Hybrid Architecture in ASR

The DNN can be regarded as a feature extractor. The speech feature (PLP, MFCC or filterbank) is feed-forwarded through DNN and is transformed into the posterior probabilities. Since the HMM requires the likelihood instead of the posterior probability during the decoding process, it is necessary to convert the posterior probability to the likelihood by dividing a prior probability of each tied-state estimated from the training set. The prior probability of each tied-state can be calculated by counting the number of frames based on Viterbi algorithm.

The state transition probabilities can be copied from a well-trained GMM-HMM. This work follows the implementation in [37] by replacing each GMM with a (pseudo) single one-dimensional Gaussian. The variance of the Gaussian is set to any positive value (e.g., always set to 1). The mean value of each tied-state is set to the corresponding state ID. In this way, evaluating each state is equivalent to a table lookup providing the log-likelihood features with the index indicated by the state ID.

2.1.3 ASR with Other Architectures

There also some other architectures of speech recognition, e.g. end-to-end [103][104] and waveform-based [151][152].

In the end-to-end architecture, the connectionist temporal classification (CTC) objective function is used to infer speech-label alignments instead of conventional HMM. This makes the model architectures and the decoding mechanism very different from the conventional DNN-HMM hybrid architecture. Now the performance of the state-of-the-art end-to-end ASR system shows significant improvement over the conventional DNN-HMM hybrid architecture.

Learning an acoustic model directly from the raw waveform has been an active area of research. Designing an appropriate feature representation and designing an appropriate classifier for these features have always been treated as separate problems in the speech recognition. This suggests the designed features might not be best for the classification objective. DNNs and their variants can be regarded as performing feature extraction jointly with classification. Current state-of-the-art waveform-based models achieve comparable performance of filterbank-based DNNs.

2.2 DNN-based Acoustic Model Training

GMM-based acoustic model and DNN-based acoustic model are two ways to classify every frame in the speech, and they are both used together with HMM model and Viterbi algorithm to decode frame sequences of speech. Generally speaking, GMM-based acoustic model is faster to compute and easier to learn. A GMM-based system could be bootstrapped from flat data.

DNN is a more complex classifier. And DNN-based acoustic model is very slow to train compared to GMM. They usually require a GPU or lots of CPUs. Training DNN-based acoustic model is usually bootstrapped from GMM-based system, since DNN requires a good initial model. Moreover, DNN training is more complicated. Following subsections will present detailed training steps of DNN-based acoustic model:

1. Train a GMM-HMM Acoustic Model

As mentioned above, the DNN-HMM hybrid model shares the tied-triphone-states and the state transition probabilities with the GMM-HMM system. Moreover, the first step of the DNN-HMM model training is to train a GMM-HMM system using the training data. The MPE training [124] is used.

2. Generate the DNN Training Labels

With the GMM-HMM model, forced alignment can be conducted at the state level using the Viterbi algorithm on the training set. The state alignment can be converted to the DNN training labels.

3. Train DNN-HMM Acoustic Model

Typical training of DNN acoustic model [37] includes pretraining and fine-tuning.

(a) Pretraining a DNN-based Acoustic Model

A number of methods can be used for pretraining. The most commonly used method is unsupervised training based on the restricted Boltzmann machines (RBMs) using contrastive divergence (CD) algorithm [38][39], and then stacking the RBMs layer by layer. Some researches alternatively use discriminative pretraining with realignment [141]. When training data is large enough, it is not necessary to conduct the pretraining [142].

(b) Fine-tuning a DNN-based Acoustic Model

In the fine-tuning stage, the pretrained network is supervised trained by error back-propagation (BP) algorithm [40]. For each frame, the gradients are used to update output softmax layer and other hidden layers (both weight matrix and bias). Moreover, the gradients need to be averaged over the mini-batch to prevent over-fitting.

2.2.1 Feed Forward Computation of DNN

Computing the outputs of the DNN is referred to as feed-forward (or forward-propagation), since it involves using each layer's output to compute the next layer's output. During forward-propagation, the training examples pass through the network layer-by-layer, with each node computing a dot product for each training sample. The dot product can be computed in parallel as one matrix multiplication per layer.

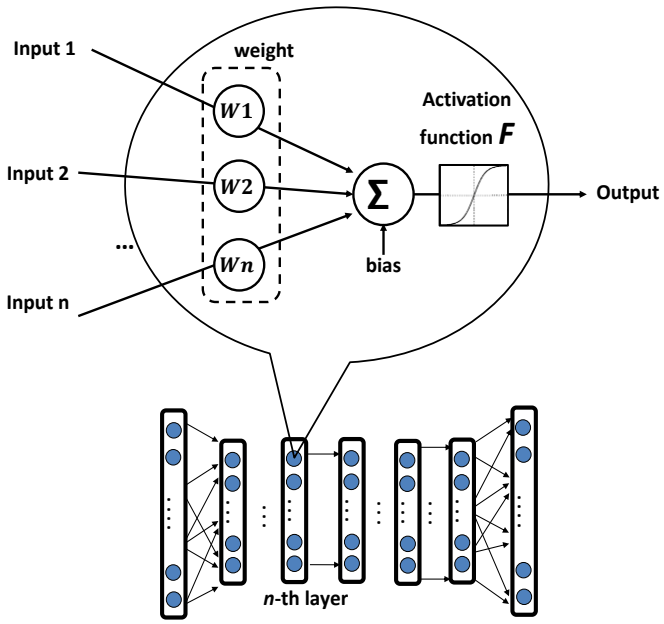


Figure 2.4: A Single Neuron in Deep Neural Network

Figure 2.4 is an example individual neuron. Suppose this neuron is in the n-th layer. The inputs of this neuron are weighted summed up with a bias value and then passed through an activation function F to get the output signal. There are many activation functions such as sigmoid, hyperbolic tangent tanh(x), rectified linear units (ReLU) [155][156], maxout [157], p-norm [158] and etc. This thesis only uses sigmoidal function for the hidden layers.

Denote the output of the n-th layer as y_n, where n is the layer number. Let the first hidden layer starts from n=1. The output layer has n=N, where N is the number of layers

in network. The input can be treated as a layer: $y_0=x$. Each neuron has a weighted connection from every output in the previous layer; there are no intra-layer connections. The output of this neuron $y_{n,j}$ can be formulated as follows:

$$z_{n,j} = \sum_i (x_{n,i} w_{n,i,j}) + b_{n,j} \tag{2.4}$$

$$y_{n,j} = F(z_{n,j}) \tag{2.5}$$

where the neurons in the n-th layer are indexed by j and neurons in layer (n-1)-th layer are indexed by i. The weight from i to j is denoted by $w_{n,i,j}$, and the bias of neuron j is denoted by $b_{n,j}$. F is the activation function. For the non-output layers, the sigmoidal activation function is used. For the output layer, softmax activation function is used, which normalizes the output z_j and makes all the output sums to 1.

$$F(z_{n,j}) = \begin{cases} \dfrac{1}{1+e^{-z_{n,j}}} & (1 < n < N) \\ \dfrac{e^{z_{n,j}}}{\sum_k (e^{z_{n,k}})} & (n = N) \end{cases} \tag{2.6}$$

2.2.2 Back Propagation Training of DNN

Training of DNN is referred to as back-propagation, since it involves propagating the error from the last layer back to the first layer.

The most popular method for training DNN is Stochastic Gradient Descent (SGD). Using SGD, the DNN is trained to minimize an objective function. Possible objective functions include cross-entropy (CE) and mean squared error (MSE), usually accompanied with regularization terms of L1, L2 or both. The cross-entropy objective function is used here:

$$\mathcal{L} = -\sum_t \sum_c \hat{y}_{c,t} ln(y_{N,c,t}) \tag{2.7}$$

where $y_{N,c,t}$ is the final layer's output vector component at time t for the class c, and $\hat{y}_{c,t}$ is a vector component of the target label after binarization ($\hat{y}_{c,t}$=1 for class c, otherwise $\hat{y}_{c,t}$=0).

SGD is a supervised training algorithm and the parameters (weights and biases) of the network are first initialized. SGD then proceeds as following pseudo-code:

Algorithm 1: Stochastic Gradient Descent (one epoch)

Shuffle the training dataset, and divide it into mini-batches (each with *m* frames).

for each mini-batch

Compute $y_{N,c,k}$ and cross-entropy $\mathcal{L}_k = -\sum_c \hat{y}_{c,k} ln(y_{N,c,k})$ for each frame *k*

$$\frac{d\mathcal{L}_k}{dw_{N,i,j}} = -y_{N-1,c,k}(\hat{y}_{c,k} - y_{N,c,k}),\ \frac{d\mathcal{L}_k}{db_{N,j}} = y_{N,c,k} - \hat{y}_{c,k}$$

for each layer *n* from *N* to *1*

Sum the gradients across the examples to get one gradient per weight and bias:

$$\frac{d\mathcal{L}}{dw_{n,i,j}} = \sum_k \frac{d\mathcal{L}_k}{dw_{n,i,j}}$$

$$\frac{d\mathcal{L}}{db_{n,j}} = \sum_k \frac{d\mathcal{L}_k}{db_{n,j}}$$

Then update the weight and bias according to follows:

$$w'_{n,i,j} = w_{n,i,j} - \frac{\alpha}{m} * \frac{d\mathcal{L}}{dw_{n,i,j}}$$

$$b'_{n,j} = b_{n,j} - \frac{\alpha}{m} * \frac{d\mathcal{L}}{db_{n,j}}$$

end for

end for

The process described above is repeated until the training converges. Each pass through the whole training set is called an epoch.

The parameter α shown in Algorithm above is called learning rate, which controls how much the network learns in each training epoch. The learning rate α is decreased as the convergence slows. The method for setting and decreasing α is called "learning rate schedule". The convergence of the network is usually monitored on a held-out validation set.

The training starts with an initial learning rate and halves the rate when the improvement in the training objective (cross-entropy) on a cross-validation set between two successive epochs falls below a threshold.

Back-propagation is also computed with one matrix multiplication per layer. Each batch of training sample causes *2N* matrix multiplications. To accelerate the matrix multiplication, GPU can be used.

2.2.3 Enhanced DNN-based Acoustic Model Refinement

Then DNN-based acoustic model can be refined by realignment of the labels, and the sequence discriminative training [41][42].

The cross-entropy criterion treats each frame independently. However, speech recognition is a sequence classification problem. In speech recognition, the most popular criterion is sequential MBR (sMBR) [42] which minimizes the expected Bayes risk. Just like MPE training for GMM-HMM systems, sequential discriminative training of DNNs starts from a set of alignments (numerator lattices) and lattices (denominator lattices) that are generated by decoding the training data with a weak language model.

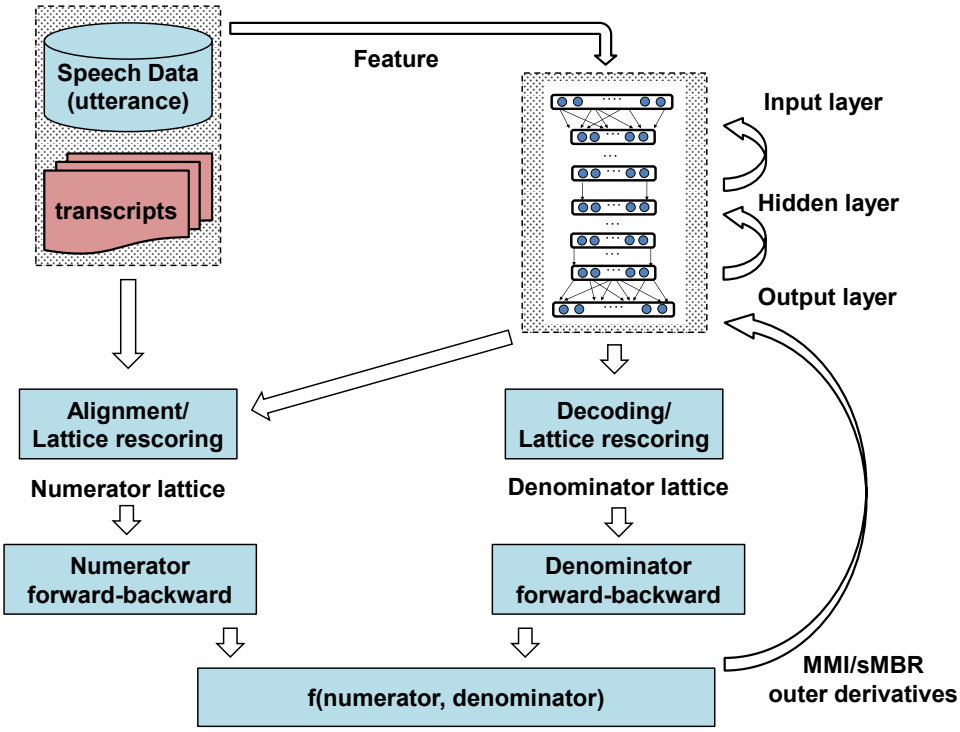


Figure 2.5: Sequence Discriminative Training DNN Models.

2.3 Acoustic Model Training with Lightly-supervised and Semi-supervised Training

The training of GMM-HMM and DNN-HMM described in the previous section depends on supervised learning. However, preparing a large scale of labelled data is very time-consuming and expensive. Moreover, usually the DNN training is more sensitive to the noise in the state label compared to GMM model training, especially in sequence discriminative training. When labelled training data is limited, the vast parameters of the DNN model cannot be fully trained. For these reasons, training paradigms other than the conventional supervised training are investigated.

2.3.1 Lightly-supervised Training

Lightly-supervised training is also called weakly-supervised training. In this method, we use only a limited amount of labelled data and a wealth of prior knowledge. Compared to other unsupervised or semi-supervised training method, this lightly-supervised training method can give better performance. And it has already been widely used in many different researches areas, e.g. for parsing and POS tagging [60], for dialog system [61], for slot tagging [62], for text normalization [63], for machine translation [64][65], speech synthesis [66]. Prior knowledge may be not straightforward. How to incorporate the prior knowledge depends on the nature of the task.

In speech recognition, we assume relevant text labels, which are not faithful transcript, such as closed caption. As shown in Figure 2.6, a typical lightly-supervised acoustic model training [32] consists of two steps. In the first step, a biased language model is constructed based on the closed caption of the relevant program to guide the baseline ASR system to decode the audio content. The second step is to filter the reliable segments of the ASR output, usually by matching it against the closed caption. In this simple method, only matched segments are selected.

The conventional filtering method, however, has a drawback that it significantly reduces the amount of usable training data. Moreover, it is presumed that the unmatched or less confident segments of the data are more useful than the matched segments because the baseline system failed to recognize them and may be improved with additional training [8].

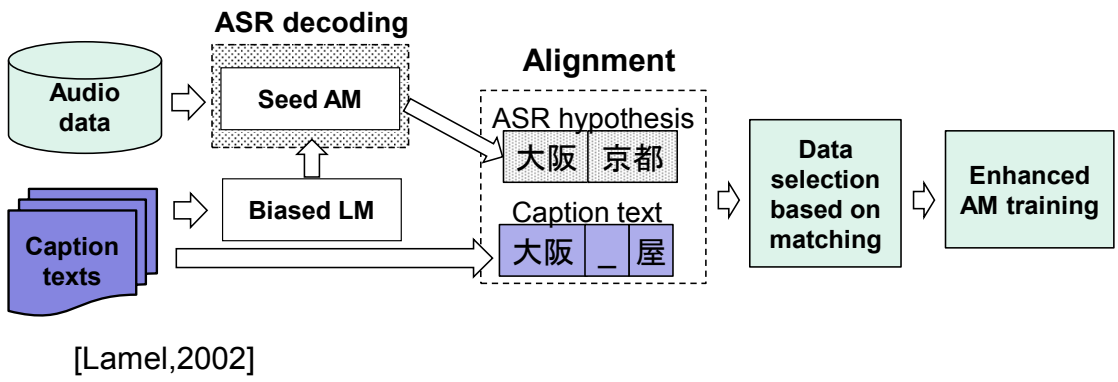


Figure 2.6: Conventional Lightly-supervised Acoustic Model Training

2.3.2 Related Work of Lightly-supervised Acoustic Model Training

Since the conventional method has a drawback that it significantly reduces the amount of usable training data, many researches focus on improving the conventional method. Recent work by Long et al. [108] proposed methods to improve the filtering by considering the phone error rate and confidence measures. Other studies, e.g. [109] introduced an improved alignment method for lightly-supervised training and [107] transformed the speaking style using SMT techniques.

Lightly-supervised training is very effective and it has already become a standard method for acoustic model training. It can be found in many state-of-the-art DNN-based ASR transcription systems, e.g. the systems for the IWSLT challenges [67][68][69][71].

2.3.3 Semi-supervised Training

Semi-supervised learning makes use of unlabelled data for training, typically enhancing a small amount of labelled data with a large amount of unlabelled data. Semi-supervised learning falls between unsupervised learning (without any labelled training data) and supervised learning (with completely labelled training data). Many machine-learning researchers have found that unlabelled data, when used in conjunction with a small amount of labelled data, can produce considerable improvement in learning accuracy.

In natural language processing (NLP) field, semi-supervised training can be addressed by adopting partial annotation, e.g. CRF-based part-of-speech (POS) tagging, word segmentation [128][129], named entity recognition tasks [130]. In these works, conditional probabilities over partially annotated data are formulated. Training is

achieved by modification to the learning objective function, incorporating partial annotation likelihood, so that a single model can be trained consistently with a mixture of full and partial annotation [131] (see Appendix 2).

In speech recognition field, semi-supervised training of acoustic model [8][9][10][11][12][13] is also developed, when labelled training data is limited and huge quantities of unlabelled data on holding. It usually takes following steps:

(1) transcribe unlabelled data with a seed model or a set of seed models trained with the labelled data.

(2) use the automatically generated transcript (or combined ASR result) as label.

(3) retrain the model by adding the newly transcribed data to the existing labelled data.

However, taking use of the unlabelled data without data selection will make the model training less effective. Because the DNN model training is sensitive to the noise in the state label, especially in sequence discriminative training [11][41][42].

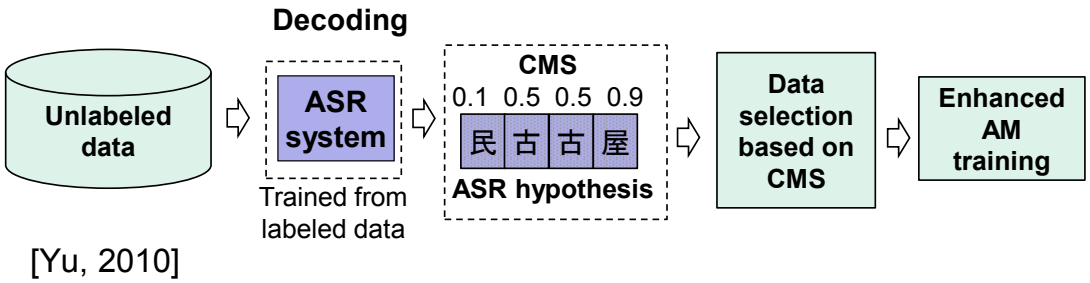


Figure 2.7: Conventional Semi-supervised Acoustic Model Training

Yu et al. [8] described the most commonly used data selection method as shown in Figure 2.7, in which utterance-level CMS is adopted in semi-supervised training of GMM-based acoustic models from unlabelled data. By sorting the utterances according to utterance-level CMS, a certain percentage of top utterances can be selected to be used for model training.

2.3.4 Related Work of Semi-supervised Acoustic Model Training

Compared to the lightly-supervised training, semi-supervised acoustic model training involves more techniques. Here are methods investigated so far.

For semi-supervised training of DNN-based acoustic models, different kinds of data filtering method has been investigated [9][10][13].

Liao et al. [9] showed that the high-confidence data are always clustered like “island of confidence”, by alternatively adopting binary word confidence scores. Applying an “island of confidence” filtering heuristic to select useful training segments, they achieved significantly improved performance for transcribing YouTube videos.

Zhang et al. [13] explored semi-supervised training of DNN in a meeting recognition task. They introduced improved DNN-based CMS estimators. Together with their error resolution, the CMS-based data selection achieved significant WER reduction.

Huang et al. [10] investigated semi-supervised GMM and DNN acoustic model training. They proposed a multi-system combination to improve the transcription accuracy and a confidence re-calibration approach to improve the data selection. Experiments showed significant improvement of retrained acoustic model on the mobile data.

Thomas et al. [19] selected the untranscribed data based on their utterance-level CMS, which was a log-linear combination of the ASR-based confidence and MLP posteriogram-based confidence. In their experiments, the method yielded a good result in a low-resource LVCSR setting.

In the fine-tuning step of DNN training, the gradients are used to update network parameters (of the weight matrix and bias) over frame-level mini-batches. It is possible to perform frame-level data selection since CMS is computed for frame-level.

Vesely et al. [11] found it beneficial to conduct frame selection based on per-frame CMS derived from confusion in a lattice, as well as to reduce the disproportion in the amounts of transcribed and untranscribed data by including the transcribed data several times in a low-resourced setting.

Imseng et al. [12] exploited un-transcribed foreign data during semi-supervised DNN training in a well-resourced setting. Their studies also revealed that CMS-based frame selection effectively reduced the size of the training data without degrading the ASR performance.

When DNN is regarded as a log-linear classifier (softmax output layer) upon a feature extractor (lower layers), unreliable data may help boost the training of lower layers, but is harmful for training the output softmax layer. Some recent studies [14][15] introduced a multi-task training architecture for semi-supervised training without confidence filtering. In [16][17][18], multi-lingual training data shares the same hidden layers but uses different softmax layers for language-dependent senone classification. This architecture is used for semi-supervised training by viewing the transcribed and untranscribed data as different languages. After training, the softmax layer for unlabelled data is thrown away and only the softmax layer for labelled data is preserved.

In summary, the objective of these methods is to avoid the unfaithful label "polluting" the softmax layer of the network. This thesis focuses on more effective data selection. There are also other machine learning methods for effective semi-supervised training of acoustic model, e.g. graph-based method [20] and submodular-based method [21]. They are very interesting, but not the focus of this thesis.

2.3.5 Conventional ROVER Combination

In automatic speech recognition (ASR), the combination of transcription hypotheses produced by multiple systems usually leads to significant word error rate (WER) reductions compared to the output of each individual system. One commonly used form of hypothesis level combination is ROVER [36].

ROVER is a two-step procedure comprised of alignment and voting. First, hypotheses from a total of component systems are iteratively aligned to create word transition networks. Then, the resulting network is searched to select the best scoring word at each node. The score is an interpolation between voting counts and word confidence scores. The score for word w at position i is formulated in Equation 2.8.

$$\text{score}(w,i) = \frac{1}{S}[\alpha \sum_{s=1}^{S} \delta(w, w_{s,i}) + (1-\alpha) \sum_{s=1}^{S} \text{conf}_s(w,i)] \tag{2.8}$$

where S is the number of systems, and $\alpha \in [0,1]$ is the parameter interpolates majority vote and confidence scores. The term $\delta(w, w_{s,i})$ is the kronecker delta and it calculates the majority voting result. Moreover, the term $\text{conf}_s(w,i)$ is the CMS of the word w at position i.

In general, ROVER also require the error rate performance of the component sub-systems to be close in order to be effective in combination.

The alignment depends on the system permutation. The combination process starts from one of the input hypotheses, which is used as “skeleton” for the greedy alignment of the others. The order in which the hypotheses are used to feed the process can hence determine significant variations in the WER of the resulting combination. Exhaustive experiments have shown that best results are obtained when systems are ordered by increasing WER.

Since ROVER is kind of confidence weighted voting, the confidence scores from the complementary systems should have the same distribution.

Another issue specially for Mandarin, where different character to word segmentations are used. When component systems use different word segmentation schemes, a direct combination between their outputs is problematic, Hence, for the Mandarin speech recognition tasks considered here, the most successful approach is to perform a character level combination. The CMS of each word is assigned to each character it contains.

3 CORPUS AND TRANSCRIPTION SYSTEM OF CHINESE LECTURES

3.1 Review on Lecture Transcription Projects

There are several projects working on spoken lecture transcription.

- CSJ: The Corpus of Spontaneous Japanese[1] [1]. This is a large-scale annotated corpus of spontaneous Japanese. This corpus is an outcome of Japan's national priority-area research project known as Spontaneous Speech: Corpus and Processing Technology (1999-2003) supported by the Ministry of Education, Culture, Sports, Science and Technology. This is a collaborative work of the National Institute for Japanese Language (NIJLA), the Communications Research Laboratory (CRL), and the Tokyo Institute of Technology (TITech). The whole CSJ contains about 650 hours of spontaneous speech that correspond to about 7M words. All these speech materials are recorded using a head-worn close-talking microphone. POS (part-of-speech) analysis is also conducted for the whole corpus. The benchmark system in Kyoto University before 2013 achieved a word error rate (WER) of 17.6% using 300 hours training data [145]. Recent benchmark system is found in the technical report[2] of Tokyo Institute of Technology.

- MIT-OCW: MIT Open Course Ware Project[3]. Many universities are now providing free web-based access to video recordings of academic lectures. This project aims to make the audio-visual material easy to search and browse with time-aligned caption texts. For this purpose, MIT-OCW speech corpus [4] is compiled, which consists of university lectures given by MIT professors over 500 hours in total, over 200 hours of which have been transcribed. The duration of each lecture is approximately 50 min. NTT Communication Science Laboratories have attempted to create a lecture transcription system by utilizing 104 hours of transcribed data. They achieved WER of 28.2% by the GMM-based system and a 22.4% by the DNN-based system in 2013 [144].

- TED: TED is short for Technology, Entertainment and Design. It is a non-profit organization that invites the world's most fascinating thinkers and doers to give a talk of their lives. TED talks is a collection of rather short speeches (max 18 minutes each, roughly equivalent to 2,500 words) which cover a wide variety of topics. Each talk is well prepared and presented by a very skilled speaker. Its website[4] makes the video recordings of the talks available under the Creative Commons license. All talks with corresponding English captions are freely download from website. Automatic transcription of TED talks has been investigated in the IWSLT challenges. In [139], NICT reported WER of 19.7% by the GMM-based system and a 14.0% by the DNN-based system. It is trained from 167.8 hours of 760 TED talks only with caption texts instead of faithful transcripts. Together with the TED data, 81.1 hours of WSJ [98] and 62.9 hours of HUB4 English Broadcast news [99] distributed from the Linguistic Data Consortium (LDC)[5] are also used. The recent performance is reported in [71].

[1]http://pj.ninjal.ac.jp/corpus_center/csj/misc/preliminary/index_e.html
[2]https://github.com/kaldi-asr/kaldi/blob/master/egs/csj
[3]http://ocw.mit.edu/index.htm
[4]https://www.ted.com/talks
[5]https://www.ldc.upenn.edu/

- YouTube: YouTube Automatic Captioning project[6]. There is a significant amount of spoken content on YouTube. In order to improve accessibility for the hearing impaired and non-native speakers, and to improve the videos searching function, Google investigated providing automatic closed captions by ASR technology in 2012 [138][140]. Although their system used almost all the state-of-the-art acoustic modelling techniques at that time and a training set consisted of approximately 1400 hours of speech, the system performance was quite low with a WER of 52.3% by the GMM-based system and a WER of 47.6% by the DNN-based system. This shows the difficulty of the task and one of the reasons is the speech data are not annotated with faithful transcripts.

- Other projects: automatic transcription were conducted for lectures of Portuguese [3], lectures of Mandarin Chinese [74][75] and lectures of Japanese language [2][5][6].

Table 3.1 summaries the benchmark systems mentioned above.

Table 3.1 Some Benchmark Systems Before 2014

Project Name	Target Language	Data size (hours)	System Performance (WER%)	
			GMM	DNN
TED talk	English	167	19.7%	14.0%
CSJ	Japanese	300	20.0%	17.6%
MIT-OCW	English	104	28.2%	22.4%
YouTube	English	1400	52.3%	47.6%

[6]http://googleresearch.blogspot.com/2009/12/automaticcaptioning-in-youtube.html

3.2 Construction of Chinese Corpus of Lecture Room (CCLR)

For a comprehensive study on ASR of spontaneous Chinese, a corpus of Chinese spoken lectures is compiled. In this section, an overview of this corpus and some linguistic analysis are presented.

3.2.1 Background Knowledge for Chinese Acoustic Modelling

3.2.1.1 Mandarin as a Tonal Language

Tone is the use of pitch in language to distinguish lexical or grammatical meaning. Languages that have this feature are called tonal languages. Tonal languages are very common in Africa, East Asia, and Central America, but rare elsewhere in Asia and in Europe; as many as seventy percent of world languages may be tonal.

In the most widely spoken tonal language, Mandarin Chinese, tones are distinguished by their distinctive shape, known as contour, with each tone having a different internal pattern of rising and falling pitch (as shown in Figure 3.1).

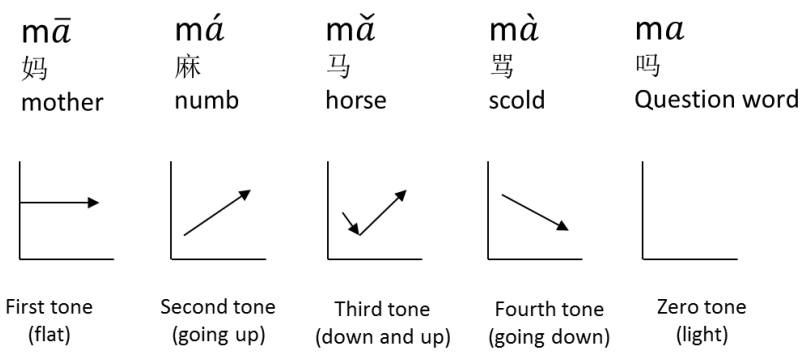


Figure 3.1: Five Tonal Patterns of Standard Mandarin Language

3.2.1.2 Pinyin

Pinyin, or Hanyu Pinyin, is the official phonetic system for transcribing the Chinese pronunciations of Chinese characters.

There are four kinds of Pinyin co-existing now.

- **Old Pinyin**: Created according to Japanese Katakana and used in old days. It still can be found in Taiwan.
- **Wade-Giles (WG) Pinyin**: now used in Hong Kong.
- **Taiwan/General/Standard/Universal Pinyin**: now used in Taiwan.
- **Mainland/Hanyu Pinyin**: now used in Chinese mainland and Singapore.

In these three annotation methods, WG was designed mainly for those people from western countries (especially English speaking countries) correctly pronouncing Cantonese. Both Pinyin of Taiwan and Chinese mainland are invented for standard Mandarin. They influence each other and are very similar[7].

For the task of this thesis, Pinyin of Chinese mainland is used for constructing the pronunciation dictionary. Chinese is a syllabic language and each character is a syllable. By using Pinyin, the syllable structure can be separated into the initial part and the final part or even smaller subword unit (see Subsection 3.2.1.4).

3.2.1.3 Accent/Dialect Problem

In current Chinese large vocabulary continuous speech recognition (LVCSR) tasks (Mandarin), variations in pronunciation of dialects have a great influence on the accuracy.

Figure 3.2: Main Dialects and Their Geographic Distributions

Traditionally, seven major groups of dialects have been recognized. Together with Mandarin Chinese, the other six are Wu Chinese, Hakka Chinese, Min Chinese, Xiang Chinese, Yue Chinese and Gan Chinese. There are great regional variations in pronunciation, vocabulary and grammar among these dialects. Figure 3.2 shows the main dialects and their geographic distributions.

[7]More detail at http://www.ntnu.edu.tw/tcsl/Teaching-Resources/pin-yin-contrast-3.htm

The regional accents of Chinese speakers show great differences. John Hopkins Summer Workshop[8] had a special report in this topic. This report [118] analysed the nature of accent through temporal, frequential and prosodic analysis. In this research here, this problem is fixed by using the dialect dictionary, which is included in CEDICT (see Chapter 3).

3.2.1.4 Acoustic Modelling Units

The typical structure of Chinese syllables is: (C)+V(N)(R) [50], where C is an optional consonant, V is a vowel explicitly annotated with 5 tones, N is an optional final nasal consonant, and R is an optional rhotic coda /r/ (it can be considered as a consonant).

An example is shown in Figure 3.3.

character	京	都	大	学
syllable	jīng	dū	dà	xué
initial+final	j+īng	d+ū	d+à	x+ué
phoneme	j+ī + ng	d+ū	d+à	x+ué

Figure 3.3: How to Derive (C)+V(N)(R) Structure of Chinese Syllables.

They derive over 100 phoneme-like units. The advantage of this method can make the monophone list compact compared with directly using Pinyin or Initial/Final as the acoustic modelling unit. The other advantage is the phoneme-like units of (C)+V(N)(R) structure is very similar to English phonemes. The Mandarin Chinese speech recognition system of this thesis can be built by using the existing English acoustic modelling techniques and solve the Mandarin-English code-mixing problem in the meantime [51].

[8] More detail available at http://old-site.clsp.jhu.edu/ws04/groups/ws04casr/

3.2.1.5 Code-mixing

Mandarin-English code-mixing is also a very challenging problem, since the spoken lectures include many English professional terms from science-technology and western literature topics. Moreover, more and more English words are used directly in the daily Chinese spoken language nowadays. From the contrastive analysis, a mapping from the English phonemes, which are missing in Mandarin, to their Mandarin equivalents (first tone by default) have been set up.

Table 3.2: Mapping from English Consonants Missing in Mandarin to Mandarin Existing Equivalents (cited from [110])

	English phonemes missing in Mandarin	Mandarin equivalents
Affricates	post-alveolar affricate /dʒ/	/tʃ/ (知)
Fricatives	inter-dental fricative /θ/	/s/ (思)
	inter-dental fricative /δ/	/z/ (资)
	voiced fricative /v/	/f/ (佛)
	voiced fricative /z/	/ts/ (资)
	voiced fricative /ʒ/	/tʃ/ (知)

Table 3.3: Mapping from English Vowels Missing in Mandarin to Mandarin Existing Equivalents (cited from [110])

	English phonemes missing in Mandarin	Mandarin equivalents
High vowels	lax vowels /ʊ/	/u/ (乌)
	lax vowels /ɪ/	/i/ (衣)
Mid vowels	mid-low front vowel /ɛ/	/ai/ (挨)
	rounded mid-low back vowel /ɔ/	/o/ (喔)
Low vowels	low front vowel /æ/	/ai/ (挨)
	low central vowel /ʌ /	/a/ (啊)

In Table 3.2 and Table 3.3, major phonemes in American English and Mandarin Chinese are all given in the form of International Phonetic Alphabet (IPA) symbols. In addition, Mandarin phonemes have additional notations of Chinese characters with the same pronunciation.

In addition, other diphthong missing in Mandarin, e.g. the /ɔɪ/, can be separated into two parts, e.g. /o/(喔)+/i/(衣).

3.2.2 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. Since 2001, a series of lectures have been given by luminary figures from a variety of areas almost every week. The total size of the corpus is 61.6 hours in speech and 1.2 M characters in text (as shown in Table 3.4). The corpus is named as the Corpus of Chinese Lecture Room (CCLR).

Table 3.4 Basic Corpus Description.

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

3.2.3 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 [111][112][113][114][115][116] and already existing corpora such as CSJ [1] and CASIA-863 [117], the most frequent and basic spontaneous phenomena can be figured out 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, the annotations have been simplified into four categories: insertion, deletion, substitution and fillers (interjections). For each of these categories, their characteristic patterns have been listed in Table 3.5.

3.2.4 Statistics on CCLR

3.2.4.1 Acoustic Conditions

The annotated 98 lectures can be categorized into five acoustic conditions as shown in Table 3.6. The lecture recordings came from conferences, TV studios, small seminar room, university classroom, and university lecture hall.

Table 3.5 Major Spontaneous Phenomena.

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

Table 3.6 Acoustic Conditions for Lecture Recordings.

Acoustic conditions	#lectures
Conferences	19
TV studio	25
Seminar room	21
University classroom	19
Lecture hall	14

3.2.4.2 Topic-related Words and Code-mixing

The annotated 98 lectures can be categorized into three topic categories as shown in Table 3.7. The topics of the total lectures are generally balanced. The percentage of the topic-related words and code-mixing rates in different topic categorizes can also be calculated. 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 is 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 3.7 Distributions of Topics.

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

3.2.4.3 Speaker Distribution

For the annotated 98 lectures by 90 speakers (21 female, 69 male), distribution of speakers' age and accent is listed in Table 3.8. When the accent types for these speakers are annotated, this thesis follows the pronunciation rules summarized in [118]. 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 provide some cues for developing acoustic modelling and the speaker adaptation strategy.

Table 3.8 Distributions of Speakers' Ages and Accents.

	30≤age≤49		50≤age≤69		age>70		
	#female	#male	#female	#male	#female	#male	total
No accent	4	15	10	15	1	4	39
South accent	2	9	3	16	0	4	34
North accent	0	3	1	3	0	0	7
Total	6	27	14	34	1	8	90

3.2.4.4 Speech Rate and Filler Rate

Speech rate and filler (interjection) rate are two major factors closely related to the speaking style. CCLR is compared with other corpora such as HUB4 (broadcast news) of Chinese [99] and GALE (broadcast conversation) of Chinese [126] as shown in Table 3.9. It is not difficult to figure out the broadcast news (HUB4) has the lowest filler rate and moderate speech rate, because the speakers are professional narrators. Moreover, 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 3.9 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

3.2.4.5 Disfluency Edit

For 68 lectures that have both faithful transcripts and caption texts, alignment of these parallel texts is conducted to get a detailed statistics on the disfluency edits as shown in Table 3.10.

Table 3.10 Average Disfluency Edit Rate (Word Level).

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

Table 3.11 lists percentages of the insertion and filler cases, and the insertion case is further broken down into discourse markers, grammatical particles and others. The interjections, discourse markers and grammatical particles together sum up to approximately 8% in the transcripts. They are all irrelevant to the lecture's content, and are omitted in the caption.

Table 3.11 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…

Table 3.12 lists the top 50 most frequent disfluency-related edit words. This result can be useful for developing specific language modelling and lightly-supervised acoustic model training [107].

Table 3.12 Top 50 Most Frequent Edit Disfluency Related Words.
(the number in the parenthesis is total occurrence in the labeled data)

1 啊 (5674)	11 就 (678)	21 在 (369)	31 要 (223)	41 和 (118)
2 呢 (5017)	12 吧 (573)	22 你 (362)	32 都 (199)	42 然后 (117)
3 这个(4028)	13 了 (554)	23 我 (345)	33 很 (196)	43 地 (117)
4 的 (2119)	14 那 (538)	24 有 (320)	34 儿 (185)	44 对 (116)
5 呃 (1721)	15 这 (520)	25 我们 (284)	35 就是说 (172)	45 到 (116)
6 是 (1513)	16 那个(489)	26 这种 (282)	36 像 (171)	46 一种 (115)
7 那么(1004)	17 它 (483)	27 呀 (254)	37 去 (147)	47 什么 (113)
8 一 (846)	18 说 (390)	28 的话 (247)	38 呐 (141)	48 所以 (110)
9 就是 (688)	19 他 (384)	29 哎 (241)	39 哈 (139)	49 把 (94)
10 个 (682)	20 嗯 (378)	30 也 (232)	40 这样 (123)	50 不 (90)

3.2.4.6 Part-of-Speech (POS) Statistics in Parallel Text

The text is segmented into words and the Part-of-Speech (POS) is analysed in the faithful transcripts and the caption texts. In Figure 3.4, it is observed that the difference in frequency of nouns and verbs. It suggests the majority of edits are related to these POS.

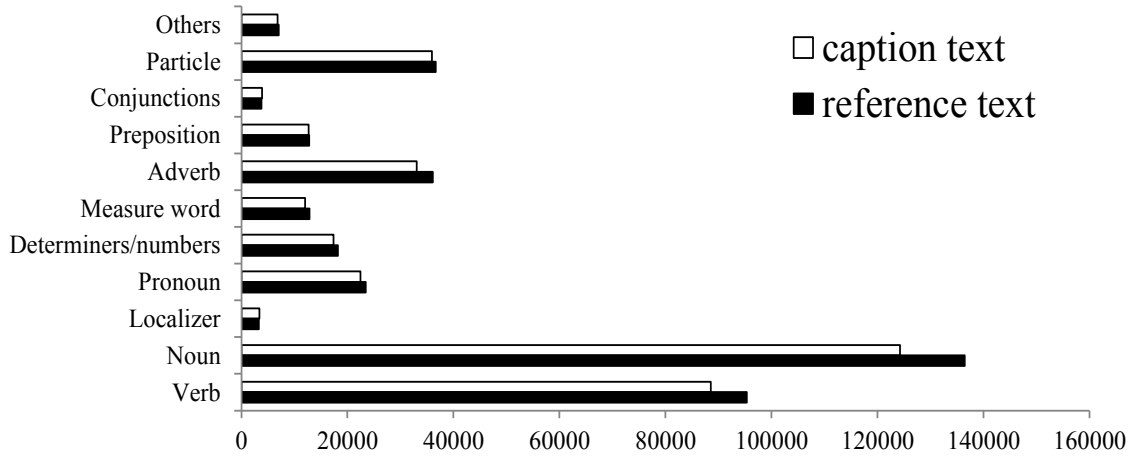


Figure 3.4: Part-of-Speech (POS) Statistics in Parallel Text.

3.3 Baseline ASR system

3.3.1 Data Sets

For the experimental purpose, 58 annotated lectures has been used as the training set (CCLR-SV), 19 annotated lectures as the test set (CCLR-TST), and 12 annotated lectures as the development set (CCLR-DEV). A large lecture data set (CCLR-LSV), which does not have faithful transcripts but has caption texts collected from the Internet, has been used together with a much larger data set (CCLR-USV), which is totally unlabelled. These two data sets can be used for lightly-supervised training and semi-supervised training, respectively. These data sets are listed in Table 3.13.

For all audio files, speech segmentation is conducted to the utterance unit (each one is less than 10sec) based on the BIC method [48] and speech clustering to remove non-speech segments and speech other than the main lecturer is also applied.

Table 3.13 Organization of Data Sets.

Data Set	Corpora	#Speaker	#Lecture	Duration (hours)
Train	CCLR-SV	51	58	35.2
	CCLR-LSV	126	126	62.0
	CCLR-USV	184	184	114.7
Dev	CCLR-DEV	12	12	7.2
Test	CCLR-TST	19	19	11.9

3.3.2 Baseline System

3.3.2.1 Acoustic Model

The typical structure of Chinese syllables is: (C)+V(N)(R) [50], 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/. Each syllable (Pinyin) can be separated into over 100 phoneme-like units [51].

The baseline GMM system uses PLP features, consisting of 13 cepstral coefficients (including C0), plus their first and second derivatives, leading to a 39-dimensional feature vector. For each speaker, cepstral mean normalization (CMN) [135] and cepstral variance normalization (CVN) [135] are applied to the features. The total number of the tied triphone states is 3000 and each state has 16 mixture components. Both MLE [134] and MPE [124] model are trained and compared.

3.3.2.2 Lexicon and Language Model

From CCLR-SV together with HUB4 and TDT4 [133], a 53k dictionary is defined and the OOV rate on CCLR-TST is 0.368%. Word pronunciations are derived from CEDICT[9] open-source dictionary. There are 1.7k English word entries and most of them are technical terms and persons' names.

A word trigram language model (LM) was built for decoding. This thesis work complemented the small-sized text of CCLR-SV and CCLR-LSV with lecture texts collected from the web, whose size is 1.07M words. Then, this lecture corpus was interpolated with the corpora (**HUB4** of 0.34M, **TDT4** of 4.75M and **GALE** of 1.03M). The interpolated weights were determined to get the lowest perplexity on CCLR-DEV as shown in Table 3.14.

Table 3.14 Component Language Models and Their Interpolated Model.

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

[9] Available at http://cc-cedict.org/wiki/

3.3.2.3 Evaluation of Baseline GMM Model

For decoding, Julius 4.3.1 [47] is used here. ASR performance is evaluated on CCLR-TST. In preliminary experiments, mismatch between the training and testing data seriously deteriorated the ASR performance and using other corpora do not have any effect. Therefore, the in-domain data (CCLR-SV) is used for acoustic model training. Both of MLE and MPE models have been trained and they are compared in Table 3.15. We can see that CER is high (39.3%) with the basic MLE model. The MPE model gets a significant CER reduction of absolute 2.6%.

Table 3.15 ASR Performance (CER%) of GMM Models.

Data Set	Durations (Hours)	MLE	MPE
CCLR-SV	35.2	39.3%	36.7%

3.3.2.4 Acoustic Modelling with Deep Neural Network (DNN)

Since DNN becomes a state-of-the-art acoustic modelling technique [37], the DNN-HMM hybrid model using CCLR-SV is also trained.

The DNN system uses 40-dimensional filterbank features plus their first and second derivatives, and has 1320 nodes as input (5 frames on each side of the current frame), 3000 nodes as output, and 7 hidden layers with 1024 nodes per layer. Training of DNN consists of the unsupervised pretraining step and the supervised fine-tuning step. They are implemented with Kaldi toolkit (Karel's setup) [46].

In the unsupervised pretraining stage, all of the training data are pooled together. And the network is initialized with stacked restricted Boltzmann machines (RBMs) that are pretrained 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 pretraining, the momentum m is linearly increased from 0.5 to 0.9, which is accompanied by a rescaling of the learning rate using $1\text{-}m$. 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, 1/8 of total utterances are held out 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%. Single GPU (Tesla K20m) is used to accelerate the training time.

For decoding, Julius 4.3.1 (DNN version[10]) is used here. It performs fast decoding with a pseudo-HTK format model (see Subsection 2.1.2) and 3000-dimensional likelihood feature vectors generated by the DNN using the state transition probabilities of the GMM-HMM model. The ASR performance (CER%) is 30.2% and an absolute reduction of 6.5% is achieved from the best MPE model in Subsection 3.3.2.3.

3.3.2.5 Training with Caption Texts

For further model improvement, this thesis work exploits 62.0 hours of the lecture data without faithful transcripts but caption texts (CCLR-LSV).

The fine-tuning step of training DNN needs supervision labels. Since the caption texts are not faithful, this thesis uses the ASR hypothesis generated from the baseline MPE model and a biased language model for each lecture as the training transcript [32]. 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 3.16 shows the effectiveness of the data increase, although the transcripts are not faithful.

Table 3.16 ASR Performance (CER%) of DNN Models.

Model	Amount of data (hours)			ASR performance (CER%)
	CCLR-SV	CCLR-LSV	Total	
Baseline	35.2	0	35.2	30.2%
Lightly-supervised	35.2	62.0	97.2	27.5%

[10] Available at http://julius.osdn.jp/en_index.php#latest_version

3.3.3 Speaker Adaptation on DNN model

To further enhance the DNN model, the work of this thesis conducts unsupervised speaker adaptation by retraining the DNN for every speaker of the test set with its initial recognition hypothesis [83][84]. Fine-tuning of DNN is conducted using the ASR hypothesis with a small learning rate with 50 utterances, which are randomly selected with the averaged word confidence score larger than 0.8.

Table 3.17 Performance (CER%) of Speaker Adaptation.

Without Adaptation	Unsupervised Adaptation
27.5%	26.9%

3.3.4 Speaker Adaptive Training on DNN model

Inspired by speaker adaptive training (SAT) in GMM-HMM framework, speaker adaptive training deep neural network (SAT-DNN) with an ensemble speaker modelling framework [77] (see Appendix 1) is also investigated.

The lightly-supervised trained DNN model (without adaptation) is used for SAT-DNN training with the data from CCLR-SV and CCLR-LSV (97.2 hours and 184 speakers).

The SAT-DNN approach for speaker adaptation adopts the ideas from multi-task learning [78][79], in which one layer is specified as a speaker-dependent (SD) layer (second hidden layer) and all other layers are shared by all speakers [80]. To capture the large variations of the speaker acoustic space in real applications, all of the SD neural weight matrices are concatenated as a speaker ensemble (184 speakers). By applying singular value decomposition (SVD)-based low-rank (rank=300) matrix adaptation method [81][82] to the ensemble matrix, the system achieves better performance when testing on the CCLR-TST set (26.5%) compared the conventional speaker adaptation method (26.9%).

3.4 Conclusion

This baseline system achieved an average Character Error Rate (CER) of 39.31% with the GMM (MLE) model, 36.66% with the GMM (MPE) model, and 30.2% with the DNN model for CCLR-TST. The main reason of the relatively low performance of

the baseline system compared with the CSJ and TED talks is the small amount of faithful training data.

Moreover, the improvement from the conventional lightly-supervised DNN model (27.5%) is still limited. Therefore, this thesis will investigate how to improve conventional lightly-supervised training to exploit unfaithful data (see Chapter 4) and semi-supervised training to exploit unlabelled data (see Chapter 5).

Speaker adaptation and speaker adaptive training can improve the DNN based acoustic model. However, they are not the main focus in this thesis work.

4 LIGHTLY-SUPERVISED ACOUSTIC MODEL TRAINING

4.1 Introduction

In order to increase the training data for an acoustic model, a scheme of lightly-supervised training, which does not require faithful transcripts but exploits available verbatim texts, has been explored for broadcast news [32][119][120] and parliamentary meetings [73].

In the case of parliamentary meetings, verbatim texts are made by stenographers, and thus can be used to predict faithful transcripts. However, in the case of TV programs, closed caption texts are not so verbatim because of the space constraint, and thus can be used in an indirect manner for lightly-supervised training. A typical method [32][119] consists of two steps. In the first step, a biased language model is constructed based on the closed caption text of the relevant program to guide the baseline ASR system to decode the audio content. The second step is to filter the reliable segments of the ASR output, usually by matching it against the closed caption. In this simple method, only matched segments are selected.

The conventional filtering method, however, has a drawback that it significantly reduces the amount of usable training data. Moreover, it is presumed that the unmatched or less confident segments of the data are more useful than the matched segments

because the baseline system failed to recognize them and may be improved with additional training [120]. This chapter addresses an effective lightly-supervised training method using dedicated classifiers for data selection.

4.2 A Two-step CRF-based Classification Scheme for Data Selection

4.2.1 Proposed Lightly-supervised Training Framework

To perform lightly-supervised training, it is necessary to have a criterion to select data. The conventional lightly-supervised training relies on simple matching between the caption text and the ASR hypothesis, and thus discards so much data which could be useful.

In this thesis, a data selection framework is proposed based on dedicated classifiers to replace the simple method as shown in Figure 4.1. Training of the classifiers is conducted by using the training database of the baseline acoustic model (CCLR-SV).

First, an ASR hypothesis (1-best) is generated using the baseline acoustic model and a biased language model. A biased language model is made for each lecture by interpolating the baseline model with the language model generated by the caption text of the lecture. The weights of these language models are 0.1 and 0.9. Unsupervised maximum likelihood linear regression (MLLR) speaker adaptation [136] is conducted in decoding CCLR-LSV.

Then, the ASR hypothesis is aligned with the corresponding caption text by using dynamic programming. By referring to the annotation (faithful transcript) of CCLR-SV, both text-based and speech-based features are extracted from the alignment patterns between the ASR hypothesis and the caption text. They are used to train discriminative classifiers to select one of them or reject both.

Finally, for CCLR-LSV, an ASR hypothesis is also generated and aligned with the corresponding caption text in a similar manner. However, there is no faithful annotation for this data set, so the derived classifiers are applied to select and verify word by word either from the ASR hypothesis or the caption text.

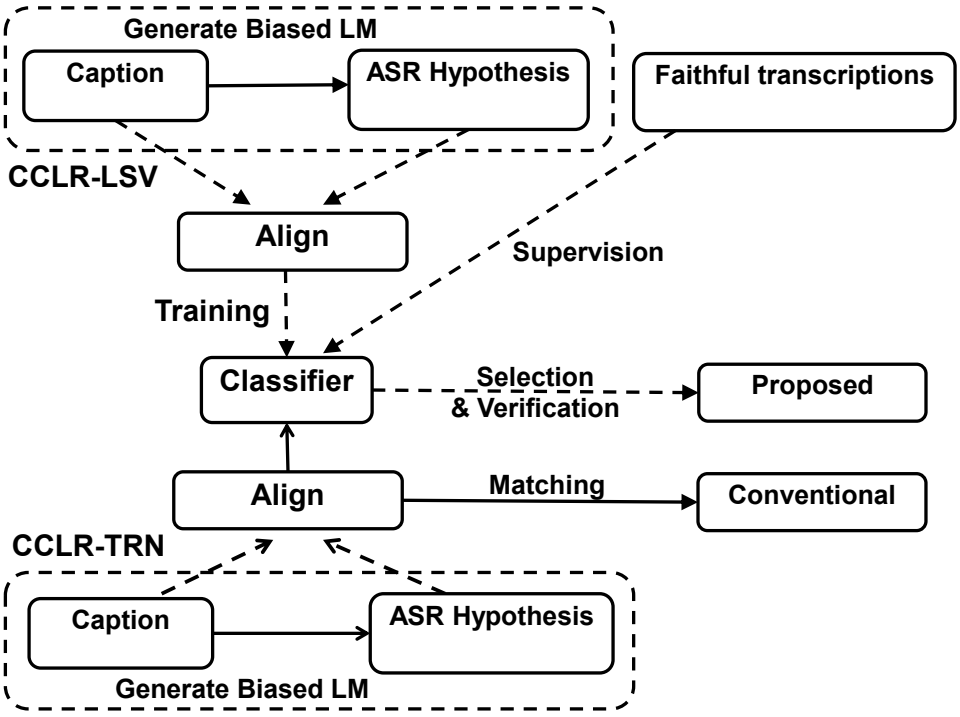


Figure 4.1: Process Flow of Proposed Lightly-supervised Training

4.2.2 Category of Word Alignment Patterns

By analysing the aligned word sequence between the ASR hypothesis and the caption text, alignment patterns can be categorized by referring to the faithful transcript, as listed in Table 4.1. Here, insertion and deletion cases are handled by introducing a null token.

Table 4.1 Category of Alignment Patterns (Word Level).

Category	Caption text		ASR hypothesis		Reference text	Percent %
C_1	论文	√	论文	√	论文	75.7%
C_2	沦亡	×	沦亡	×	论文	2.9%
C_3	论	×	沦亡	×	论文	3.9%
C_4	沦亡	×	论文	√	论文	13.5%
C_5	论文	√	沦亡	×	论文	4.0%

(× means mismatching with reference, √ means matching)

- C_1: the ASR hypothesis is matched with the caption and also the correct transcript. A majority of the samples falls in this category.
- C_2: although the ASR hypothesis is matched with the caption, it is not correct. This case is rare.
- C_3, C_4 and C_5: the ASR hypothesis is different from the caption. In C_3, neither of them is correct. In C_4, the ASR hypothesis is correct. In C_5, the caption is correct.

Note that the conventional method [32][119] is equivalent to simply using ***C1*** and ***C2***. The objective of this study is to incorporate more effective data (***C4*** and ***C5***) while removing erroneous data (***C2*** and ***C3***).

The distribution of these patterns in CCLR-SV is shown in Table 4.1. It is observed that 75.7% of them are categorized into ***C1***. Among others, ***C4*** is the largest because the caption text is often edited from the faithful transcript for readability.

Initially a classifier to conduct classification of these five categories was tried, but it turned to be difficult because of the complex decision and the data imbalance. Therefore, a cascaded approach is adopted.

4.2.3 Cascaded Classifiers for Word-level Data Selection

In the cascaded approach, two kinds of classifiers have been designed. One is for selection of the hypothesis and the other is for verification of the selected hypothesis.

C_1 and C_2 are the matching cases between the ASR hypothesis and the caption. In these cases, the data selection problem is reduced to whether to accept or discard the word hypothesis. On the other hand, C_3, C_4 and C_5 are the mismatching cases between the ASR hypothesis and the caption. A binary classifier is trained to make a choice between the ASR hypothesis and the caption word. Then, the other classifier is applied to verify it. This classifier can be the same as the one used for C_1 and C_2.

The classification is organized by the two binary classifiers in a cascaded structure as illustrated in Figure 4.2. The binary classifiers are focused on specific classification problems, so they are easily optimized. This design also mitigates the data imbalance problem. In Figure 4.2, one classifier is used for selection of the word hypothesis with highest credibility either from the ASR hypothesis or the caption text, and the other is used for verification of the selected (or matched) hypothesis.

To make binary classification, C_3 is merged into C_4, because it is observed the phone accuracy of the ASR hypothesis is higher than that of the caption text in C_3. The second classifier will reject erroneous patterns in C_3.

Note that the conventional method [32][119] simply accepts C_1 and C_2, but the proposed method can also incorporate more effective data (C_4 and C_5) and remove erroneous data (C_2).

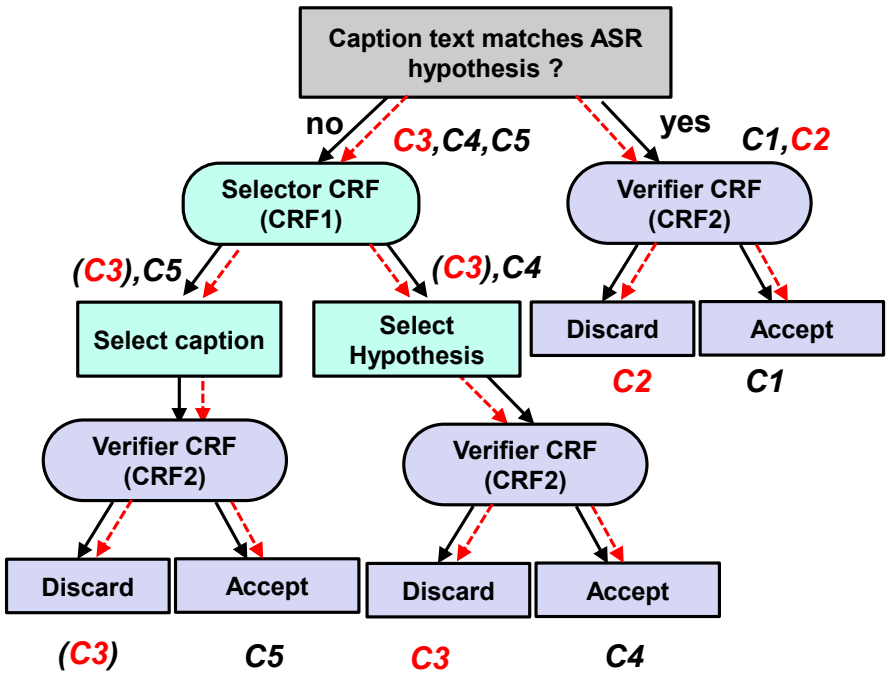


Figure 4.2: Cascaded Classification Scheme for Data Selection.

4.2.4 Feature Set Design for Classifiers

CRF [29] is used as the classifier for this task. It can model the relationship between the features and labels by considering sequential dependencies of contextual information. For this reason, it is used for many applications such as confidence measuring [31][30], ASR error detection [52][55], speech recognition [105][106], and automatic narrative retelling assessment [53].

When training the classifiers and conducting data selection, it is necessary to convert the alignment patterns into a feature vector. These features include both acoustic and linguistic information sources. They are selected by referring to the work on confidence measures and ASR error detection. The text-based features are defined

for both ASR hypothesis and caption text while the speech-based features are computed for the ASR hypothesis only.

Table 4.2 Feature Set for Classification.

Feature Type	Features
Text-based	1. Lexical feature (**LEX**) 2. Part-of-Speech (**POS**) 3. Language model probability (**LM**) 4. tf-idf (**TF**)
Speech-based	1. confidence measure by decoder (**CMS**) 2. word duration (**DUR**)

These features, listed in Table 4.2, are explained below.

- The lexical feature (**LEX**) is a lexical entry (ID) of the current word. It is a symbolic feature.
- The Part-of-Speech (**POS**) feature is obtained by a CRF classifier trained with Chinese-Tree-Bank (CTB) 4 [137]. Fifteen POS tag symbols have been defined according to the CTB's guideline. This feature is symbolic.
- The language model probability feature (**LM**) is a negative log probability of the current word by unigram, bigram and trigram models. Back-off is not considered here. This feature set is numeric.
- The tf-idf (**TF**) feature is computed by multiplying the tf-value and the log idf-value. The tf-value is calculated from the word frequency in the caption text of the current lecture. The idf-value is computed from the caption text of entire CCLR-SV and CCLR-LSV sets. This feature is numeric.
- The confidence measure score (**CMS**) is output by the Julius decoder [25] of the baseline ASR system. The value is between [0,1] approximating a posterior probability of the hypothesis word.
- The word duration (**DUR**) feature is the number of frames of the word.

Because most of the CRF implementations are designed to work with symbolic features, it is necessary to convert the numeric features into discrete features.

Moreover, for the symbolic features of **LEX** and **POS**, the contextual information of the current word is also incorporated by adding features of the preceding two words and the following two words.

4.2.5 Utterance Selection for Acoustic Model Training

For CCLR-LSV, the ASR hypothesis and the caption text are merged into a single word sequence after the matching and selection process, and every word in the sequence will have a label, either "accept" or "discard", based on the verification process according to Figure 4.2.

Then, it is necessary to make a decision whether or not this sequence of the data by the utterance unit is used for acoustic model training. Since the acoustic model is based on phone units, phone-based accuracy is a natural measure for selection of utterances [108]. In this work, it is possible to compute the phone acceptance rate (**PA**) for every utterance by distributing the "accept" and "discard" classification results to all phones. The "**PA**" actually means the ratio of "accept" phones over the total number of phones in an utterance. However, it is not easy to figure out the optimum point on the threshold of this measure between the growth of noise and the amount of training data [121]. It is affected by a number of factors and often determined a posteriori depending on the data set and the baseline performance. It will be shown that using only reliable utterances (**PA**=100%) is best for the proposed lightly-supervised acoustic model training.

4.3 Experimental Evaluations

4.3.1 Classifier Implementation and Performance

The proposed method is applied to CCLR-LSV (126 lectures, 62 hours) to make an enhanced acoustic model, which are tested on CCLR-TST.

Speech segmentation is first conducted to the utterance unit based on the BIC (Bayesian Information Criterion) method in Chapter 3 and speaker clustering to remove non-speech segments and speech from other than the main lecturer in CCLR-LSV.

In the implementation, the Wapiti[11] CRF classifier [122] is used here to train two classifiers using CCLR-SV: **CRF-2**, which is trained to discriminate C_1 vs. C_2, and **CRF-1**, which is trained to discriminate C_3+C_4 vs. C_5. Because of the sparse features with a high dimension, L1 regularization and the Orthant-Wise Limited-memory Quasi-Newton (OWL-QN) algorithm is used to train the CRF models [123].

[11] Available at http://wapiti.limsi.fr/

In the training dataset, there is serious imbalance between classes as observed in Table 4.1. This will bias the training of the classifiers. Thus, a re-sampling technique is introduced.

Specifically, the samples in C_2 are duplicated, and part of samples in C_1 and C_3+C_4 are discarded. As a result, the calibrated distributions are as follows: C_1: 44.1%, C_2: 24.1%, C_3+C_4: 19.5% and C_5: 12.3%.

Classification performance with various feature sets is compared by 5-fold cross validation on CCLR-SV, as shown in Table 4.3 and Table 4.4. Performance is measured by precision, recall and F-score:

$$Precision = TP / FP$$
$$Recall = TP / (FP + FN)$$
$$F - score = 2 \times Precision \times Recall / (Precision + Recall)$$

where TP is true positive (correct output), FP is false positive (false alarm), and FN is false negative (miss).

It is observed that the overall performance of **CRF-2** (Table 4.4) is higher than that of **CRF-1** (Table 4.3). It suggests selection of the hypothesis is more difficult than verification of the hypothesis. In **CRF-2** (Table 4.4), performance of C_1 (verification) is higher than that of C_2 (rejection), because the number of training samples of C_1 is much larger than that of C_2. The re-sampling technique does not essentially solve the problem of a smaller variety and coverage, though it mitigates it.

Among the set of features, the text-based features are generally more effective than the speech-based features, but combination of both feature sets shows further improvement. As an individual feature, the lexical feature is the most effective for **CRF-1**, while the **POS** feature is the most effective for **CRF-2**, since more variety is needed for selection than verification of the hypothesis. Note that the confidence measure score (**CMS**) is not so effective as expected. Its performance is comparable to that of the duration feature (**DUR**).

From these results, the complete feature set is adopted. Although errors by **CRF-1** in the first stage of the classification is inevitable, part of them are detected and discarded in the second stage of classification by **CRF-2**, as shown in Figure 4.2.

The classification rate (Recall) is ***C1***: 98.5%, ***C2***: 63.9%, ***C3+C4***: 84.5%, ***C5:*** 76.9%.

Table 4.3 Feature Evaluation of CRF-1 by 5-fold Cross Validation (on CCLR-SV).

	CRF-1					
	C3 + C4			***C5***		
Feature	Recall	Precision	F-score	Recall	Precision	F-score
LEX	0.831	0.818	0.825	0.709	0.727	0.718
POS	0.817	0.799	0.808	0.676	0.700	0.688
LM	0.773	0.815	0.794	0.724	0.669	0.695
TF	0.825	0.775	0.799	0.622	0.693	0.656
LEX+POS+LM+TF-IDF	0.828	0.834	0.831	0.740	0.732	0.736
CMS	0.789	0.783	0.786	0.655	0.663	0.699
DUR	0.785	0.810	0.797	0.709	0.676	0.692
CMS+DUR	0.810	0.807	0.808	0.694	0.698	0.696
All Features	**0.845**	**0.852**	**0.848**	**0.769**	**0.758**	**0.763**

Table 4.4 Feature Evaluation of CRF-2 by 5-fold Cross Validation (on CCLR-SV).

	CRF-2					
	C1			***C2***		
Feature	Recall	Precision	F-score	Recall	Precision	F-score
LEX	0.975	0.803	0.880	0.561	0.923	0.698
POS	0.960	0.828	0.889	0.634	0.897	0.743
LM	0.983	0.794	0.878	0.531	0.945	0.680
TF	0.906	0.762	0.828	0.480	0.737	0.581
LEX+POS+LM+TF-IDF	0.984	0.821	0.895	0.605	0.953	0.740
CMS	0.955	0.809	0.876	0.585	0.877	0.702
DUR	0.974	0.812	0.885	0.586	0.924	0.717
CMS+DUR	0.973	0.815	0.887	0.594	0.923	0.723
All Features	**0.985**	**0.833**	**0.903**	**0.639**	**0.958**	**0.766**

4.3.2 Utterance Selection for Model Training

For utterance selection for acoustic model training, the phone acceptance (**PA**) rate is defined for every utterance of CCLR-LSV by distributing the "accept" and "reject" classification results to all phones. The lower bound of **PA** can be set as a threshold for selecting utterances. However, it is not practical to tune the threshold by using the development set, as it would take so long to train the DNN model for each **PA** threshold value. Therefore, the tuning is conducted with GMM-HMM (MLE) by adding the selected data to CCLR-SV.

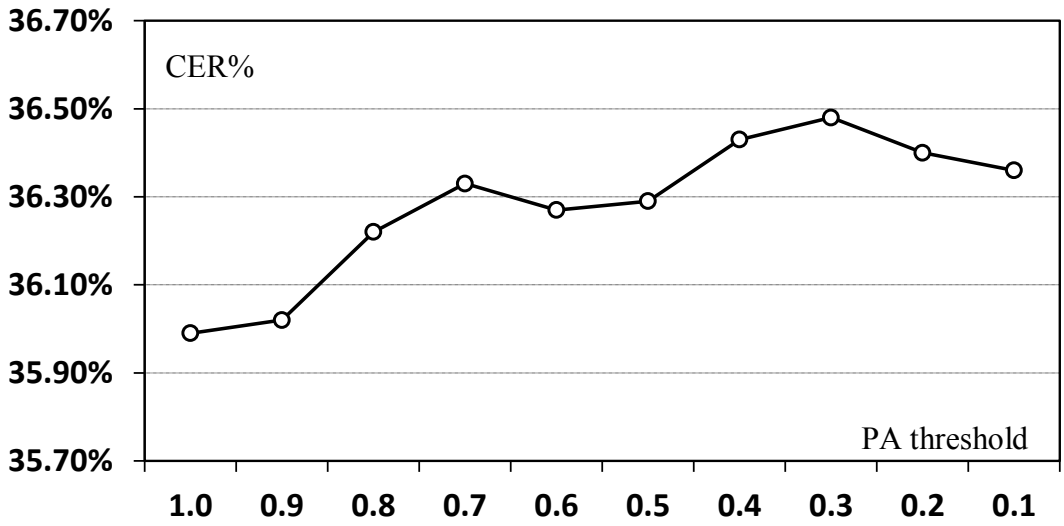


Figure 4.3: ASR Performance (GMM-HMM on CCLR-DEV) for Different PA Threshold Values.

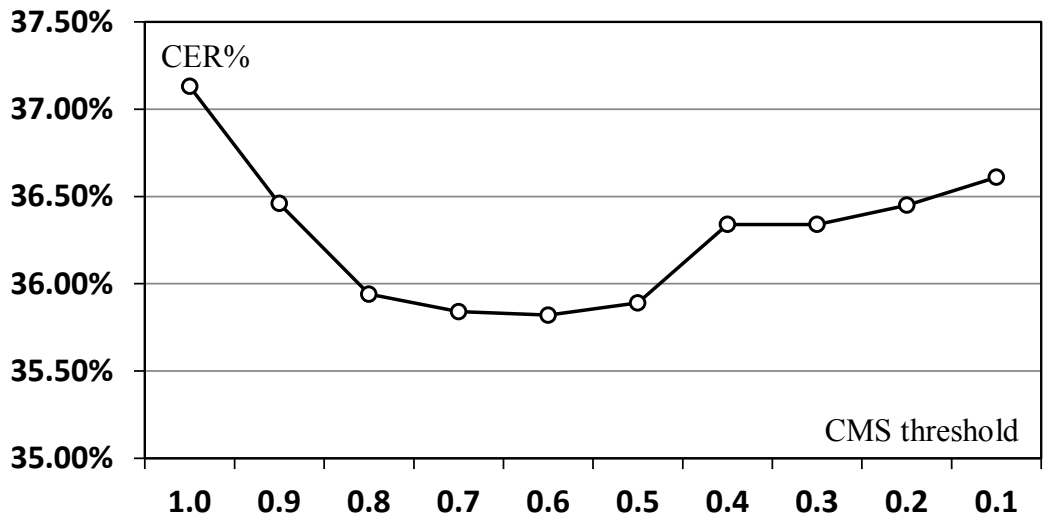


Figure 4.4: ASR Performance (GMM-HMM on CCLR-DEV) for Different CMS Threshold Value.

ASR performance (CER%) on CCLR-DEV is plotted in Figure 4.3. Note that adding more data by relaxing the **PA** threshold only degraded the ASR performance, due to the increase of errors. The best ASR performance is achieved at **PA**=100%. It shows the advantage of the proposed method that it can effectively select the most usable utterances and makes the data selection easy without tuning the threshold in the lightly-supervised acoustic model training.

On the other hand, as shown in Figure 4.4, the selection based on CMS is not straightforward, since the optimal threshold is not easy to determine.

4.3.3 ASR Performance with Enhanced Model Training

Next, lightly-supervised training of the acoustic model is conducted after classification on CCLR-LSV and utterance selection. The same setting with the baseline system described in Chapter 3 is used for acoustic model training as well as the lexicon and language model. Two DNN systems with different features are trained for comparative study and verification of the experimental results. ASR performance of the DNN model enhanced by the selected data is evaluated on CCLR-TST. The proposed data selection method is compared with other three methods as follows:

- **Baseline**: the model trained by only using CCLR-SV as described in Chapter 3. It is an expected lower bound of the proposed method.
- **No selection**: simply pool the CCLR-SV lectures and entire CCLR-LSV lectures together, and directly use the ASR hypothesis of CCLR-LSV without any selection.
- **CMS filtering**: For the ASR hypothesis of CCLR-LSV, the word-level CMS computed by the baseline ASR system is distributed to all phones in each word, and is averaged over the utterance unit for data selection. A series of GMM-HMM models (MLE) have been trained by adding the selected utterances, with different threshold values on CMS, from CCLR-LSV to CCLR-SV. An optimum point at CMS≥0.6 where the best ASR performance on CCLR-DEV is found, as shown in Figure 4.4.
- **Conventional matching**: the conventional lightly-supervised training which selects the data based on simple matching of the ASR hypothesis and the caption text (upper part of Figure 4.2).

ASR performance in CER% is listed for DNN models in Table 4.5 and Table 4.6. The DNN models trained from Fbank feature outperform the DNN models trained from PLP feature. The results of both feature types show that the proposed lightly-supervised training method outperforms all other methods. The percentage of data selected from CCLR-LSV by the proposed method is 78.9%, which is almost double of the data by the conventional method (41.9%). However, without any selection, ASR performance is degraded due to inclusion of erroneous segments. This result demonstrates that the classifiers work effectively for CCLR-LSV. Compared with the CMS filtering, the

proposed method selects usable data more effectively, as confirmed both in Table 4.5 and Table 4.6.

Another advantage of this proposed method is it can select usable data effectively without tuning threshold parameters. Comparing Figure 4.3 and Figure 4.4, it is apparently difficult to find the optimal point in the CMS threshold, which depends on the ASR system and the training data.

Table 4.5 ASR Performance (CER%) by Lightly-supervised Trained DNN Acoustic Model (Fbank).

	Amount of data (hours)		ASR Performance
	CCLR-SV	CCLR-LSV	CER% CCLR-TST
Baseline	35.2	0.0	30.2
No selection	35.2	62.0	27.5
CMS filtering	35.2	46.3	27.7
Conventional matching	35.2	26.5	29.0
Proposed	**35.2**	**48.9**	**27.2**

Table 4.6 ASR Performance (CER%) by Lightly-supervised Trained DNN Acoustic Model (PLP).

	Amount of data (hours)		ASR Performance
	CCLR-SV	CCLR-LSV	CER% CCLR-TST
Baseline	35.2	0.0	31.6
No selection	35.2	62.0	28.8
CMS filtering	35.2	46.3	28.7
Conventional matching	35.2	26.5	29.2
Proposed	**35.2**	**48.9**	**28.4**

4.4 Conclusion

A novel data selection method for lightly-supervised training of an acoustic model has been proposed. The method uses dedicated classifiers for data selection, which are trained with the training database of the baseline acoustic model. A cascaded classification scheme based on a set of binary classifiers is designed by incorporating a variety of features. Experimental evaluations show that the proposed lightly-supervised training method effectively increases the usable training data and improves the accuracy from the baseline model and in comparison with the conventional method. This means

the proposed method can effectively identify the most credible data in huge archives of unfaithful data.

5 SEMI-SUPERVISED ACOUSTIC MODEL TRAINING

5.1 Introduction

Semi-supervised training combines a small set of labelled data with a large set of unlabelled data. The conventional paradigm of semi-supervised acoustic model training dealing with the unlabelled data includes pre-processing (e.g. speech segmentation, non-speech removal, speaker diarization, etc.), automatic transcription generation, data selection and model training. A number of studies have been conducted to address these processes [8][9][10][11][12][13]. However, they still do not solve the crucial part of automatic transcription generation and data selection. This chapter focuses on these issues of the conventional paradigm of the semi-supervised training method.

For data selection, the most commonly used method is based on the CMS computed by the ASR system [22][23][24][25][26][27][28]. The word-level CMS is averaged over the utterance unit for data selection. When tuning the threshold of CMS, there is a trade-off between the data increase and the growth of noise in the label. It is not straightforward to find the optimal threshold and it is not practical to conduct exhaustive search. Moreover, the optimum threshold depends on the available data size. This means that it is needed to tune the threshold every time the data size is increased and the

ASR system is updated. Instead of using CMS, this chapter addresses a novel method that uses dedicated classifiers to select usable data for model training.

The previous chapter addresses the approach in the lightly-supervised training [32] setting, where closed caption text is available and combined with an ASR hypothesis [33]. However, the assumption of closed caption text limits the applicability of the method. This work extends to the more general semi-supervised setting. The text quality can be leveraged by combining hypotheses from a set of complimentary ASR systems with similar accuracy and enough diversity on recognition patterns [34]. Deng et al. [35] mentioned enough diversity exists between GMM and DNN systems. Conveniently, we can reuse the GMM-HMM system that is produced in the process of the DNN-HMM acoustic model training as a complementary system. Conventionally, ROVER-based system combination [36] has been used, but it is not robust to the small number of complementary systems with different distributions of CMS. In this study, the problem is solved by using a cascade of CRF classifications. In the proposed method, the CRF-based classifiers are prepared for two sub-tasks: selector CRF and verifier CRF. The selector CRF is trained to select a correct (or better) hypothesis either from GMM-HMM or DNN-HMM on the character/word level. The verifier CRF is then used to determine whether the selected result is reliable or not. Data selection for acoustic model training is conducted according to the verification result.

5.2 Comparative Analysis of Different Baseline Systems

We use the system enhanced by the previous chapter. Here language model is further enhanced by adding Phoenix corpus (text recordings of 1,300 broadcasted lectures from the Phoenix-HK official website[12]) as shown in Table 5.1.

Table 5.1 Component Language Models and Their Interpolated Model.

	Corpora	#Words	PPlex.	Weights
Component Language Models	CCLR	1.07M	374	0.31
	HUB4	0.34M	710	0.01
	TDT4	4.75M	923	0.04
	GALE	1.03M	426	0.16
	Phoenix	4.12M	352	0.48
Interpolated Language Model		11.31M	248	/

This baseline system achieved an average Character Error Rate (CER) of 24.2% and 27.5% with MLLR speaker adapted GMM-HMM model, and 22.7% and 25.7% with the DNN-HMM model for CCLR-DEV and CCLR-TST.

Hypothesis combination requires a set of complimentary ASR systems with similar accuracy and enough diversity on recognition patterns [34]. Two other DNN systems with the different feature types are also trained. One uses 13-dimensional MFCC features (with the first and second derivatives) and the other uses 13-dimensional PLP features (with the first and second derivatives). Performance (CER) on CCLR-DEV is listed in Table 5.2. Difference in the CER% is less than 2%. The pair-wise edit distances of these systems are listed in Table 5.3. The largest diversity exists between GMM and DNN systems with similar accuracy as mentioned by [35].

Table 5.2 ASR Performance on CCLR-DEV.

System	CER%	Feature
DNN	22.7	40 dim fbank+Δ+ΔΔ, CMVN
	23.5	13 dim MFCC+Δ+ΔΔ, CMVN
	24.0	13 dim PLP+Δ+ΔΔ, CMVN
GMM (MLLR)	24.2	13 dim PLP+Δ+ΔΔ, CMVN

Table 5.3 Pair-wise Edit Distance of ASR Results on CCLR-DEV (Character Level)

	GMM	DNN(PLP)	DNN(MFCC)	DNN(fbank)
GMM	/	/	/	/
DNN(PLP)	**24.5%**	/	/	/
DNN(MFCC)	**24.6%**	14.7%	/	/
DNN(fbank)	**24.3%**	17.3%	16.1%	/

Conveniently, it is possible to reuse the GMM-HMM system (MLLR adapted) that is produced in the process of the DNN-HMM (filterbank feature) as a complementary system.

[12] Available at http://v.ifeng.com/gongkaike/sjdjiangtang/

5.3 CRF-based Hypothesis Combination and Data Selection for DNN Model Training

The flowchart of the proposed system combination and data selection with CRF-based classifiers is shown in Figure 5.1.

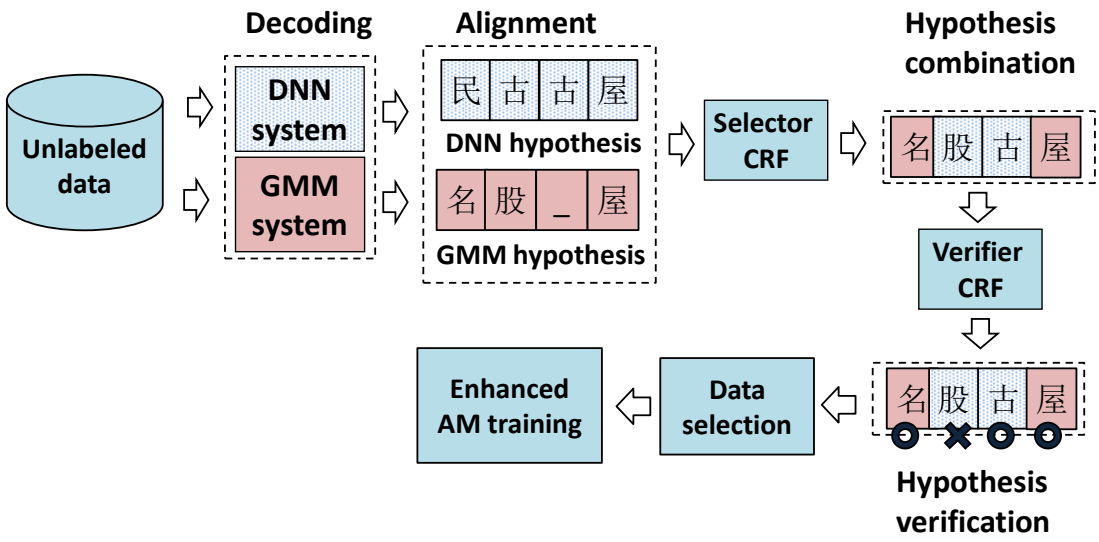


Figure 5.1: Process Flow of the Proposed Semi-supervised Training.

5.3.1 Proposed Framework

5.3.1.1 Pre-processing and Hypothesis Generation

For pre-processing, speech segmentation is first conducted. Then the unlabelled data in CCLR-USV is decoded by the DNN system and the speaker adapted GMM system, respectively.

5.3.1.2 Hypothesis Combination and Verification

Since different recognition patterns are observed between GMM and DNN based recognition hypotheses, CRF models are used to combine these diversities with their contextual information and determine which hypothesis should be selected for acoustic model training. At first, features are extracted from pair-wise aligned texts on the character level. Note that each Chinese character represents a syllable and has a corresponding meaning [49][50][51]. The character unit is adopted in order to avoid the mis-alignment due to different word segmentations and OOV problem. Moreover, as the size of characters is much smaller than the vocabulary size, training CRF models can be more efficient. Then, a correct (or better) hypothesis is selected from complementary hypotheses and verified.

5.3.1.3 Post-processing and Acoustic Model Training

Data selection for acoustic model training is conducted by aggregating the result of the CRF classification in the utterance level. The DNN system is retrained by adding the selected data.

5.3.2 Categories of Alignment Patterns

CCLR-SV is automatically transcribed data, and a three-way character alignment among these two ASR hypotheses by the GMM-based system and the DNN-based system and also the faithful transcripts (reference) is conducted. The alignment patterns can be categorized into five classes, as listed in Table 5.4. The insertion and deletion cases are handled by using a null token.

Table 5.4 Category of Alignment Patterns.

Category	DNN hypothesis		GMM hypothesis		Reference text	Percent %
C_1	学	√	学	√	学	75.2%
C_2	鳕	×	鳕	×	学	6.8%
C_3	鳕	×	穴	×	学	6.6%
C_4	学	√	鳕	×	学	7.7%
C_5	鳕	×	学	√	学	3.7%

(√ means matching with reference, × means mismatching)

The definitions of the categories are as follows:

- C_1: the DNN hypothesis is matched with the GMM hypothesis and also the correct transcript. A majority of the samples falls in this category.
- C_2: although the DNN hypothesis is matched with the GMM hypothesis, neither of them is correct. This case is rare.
- C_3, C_4 and C_5: the DNN hypothesis is different from the GMM hypothesis. In C_3, neither of them is correct. In C_4, the DNN hypothesis is correct. In C_5, the GMM hypothesis is correct.

5.3.3 Classifier Design

CRF [29] is used as the classifier for this task. It can model the relationship between the features and labels by considering sequential dependencies of contextual information. For this reason, it is used for many applications such as confidence

measuring [30][31], ASR error detection [52][55], and automatic narrative retelling assessment [53].

The objective is to accept effective data (C_1, C_4 and C_5) and remove erroneous data (C_2 and C_3). A flat classifier is initially designed and the data selection and verification problem is cast as a five-class classification problem, but it turned to be difficult because of the complex decisions and the data imbalance (see Table 5.4). Therefore, a cascaded approach is adopted.

In the cascaded approach, two kinds of binary classifiers are designed: selector CRF and verifier CRF. The selector CRF is for selection between the hypotheses, and the verifier CRF is for verification of the selected hypothesis. As described in the previous subsection, C_1 and C_2 are the matching cases between two different ASR hypotheses. In these cases, the data selection problem is reduced to whether to accept or discard the character hypothesis. On the other hand, C_3, C_4 and C_5 are the mismatching cases between these two ASR hypotheses. A binary classifier is trained to make a choice between these ASR hypotheses. Then, the other classifier is applied to verify it. For more general purpose, this classifier is the same as the one used for C_1 and C_2. There is not enough training samples to train an individual classifier.

The classification is organized by the two binary classifiers in a cascaded structure as illustrated in Figure 5.2. The binary classifiers are focused on specific classification problems, so they are easily optimized. This design also mitigates the data imbalance problem. In Figure 5.2, one classifier is used for selection of the character hypothesis with highest credibility either from the DNN hypothesis or the GMM hypothesis, and the other one is used for verification of the selected (or matched) hypothesis.

To make binary classification in the selector CRF (**CRF-1**), C_3 is merged into C_5, because it makes the data distribution more balanced. Erroneous patterns in C_3 (i.e. GMM hypothesis is incorrect) will be rejected by the verifier CRF (**CRF-2**).

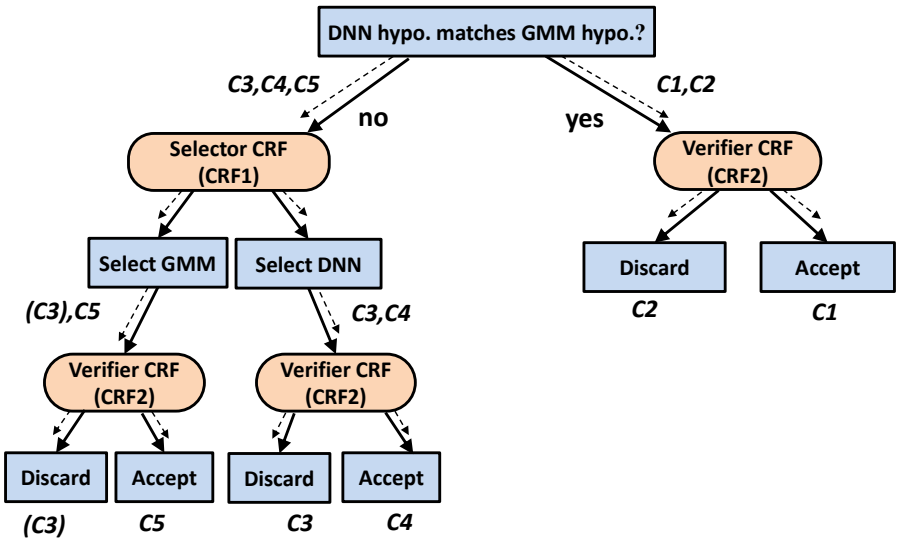


Figure 5.2: Cascaded Classification Scheme for Data Selection.

5.3.4 Feature Design

The input features used in **CRF-1** and **CRF-2** are listed in Table 5.5 and Table 5.6. These features are categorized into two groups: ASR-based and text-based features.

Table 5.5 Feature Design for CRF-1

Feature Type	Features
ASR-based feature	1. Confidence measure score (**CMS**). 2. Duration of the current word (**DUR**). 3. Word trigram LM score (**WLM**). 4. Averaged acoustic model score (**AM**). 5. Number of left competing words (**NLW**). 6. Number of right competing words (**NRW**). 7. Density within word duration (**DEN**).
Text-based feature	1. Lexical feature (**LEX**). 2. Part-Of-Speech (**POS**). 3. 5-gram char LM probability (**CLM**). 4. 5-gram char LM back-off behavior (**BO**).

These features are explained below. The ASR-based features are extracted for the word unit, and distributed to each character in the word. The Julius decoder is modified to output the ASR features when decoding[13]. They are numeric features:

[13] Available at http://github.com/halspeech/Julius_feat_extractor

- The confidence measure score (**CMS**) is output by the Julius decoder [25] of the baseline ASR system. The value is between [0, 1] approximating a posterior probability of the hypothesis word.
- The word duration (**DUR**) feature is the number of frames of the word.
- The word trigram LM (**WLM**) feature is the word trigram language model score of the word while decoding.
- Averaged acoustic model score (**AM**) feature is the acoustic likelihood score averaged for each frame.
- The left competing words (**NLW**) feature is the number of the competing words to the left side of the current word in the word graph.
- The right competing words (**NRW**) feature is the number of the competing words to the right side of the current word in the word graph.
- The density (**DEN**) feature is how many words overlapping between the start time and the end time of the current word in the word graph.

Table 5.6 Feature Design for CRF-2

Feature Type	Features
ASR-based feature	1. Confidence measure score of DNN system and posterior output of CRF-1 (**CMS**).
Text-based feature	1. Lexical feature (**LEX**). 2. Part-Of-Speech (**POS**). 3. 5-gram char LM probability (**CLM**). 4. 5-gram char LM back-off behavior (**BO**).

The text-based features are extracted by rescoring and syntactic analysis in the character level:

- The lexical feature (**LEX**) is a lexical entry (ID) of the current character. It is a symbolic feature.
- The Part-Of-Speech (**POS**) feature is obtained for each character unit by a CRF classifier trained with a character-based Chinese-Tree-Bank (CTB) 4 [54]. This feature is symbolic.

- The language model probability feature (**CLM**) is a negative log probability of the current character rescored by a character 5-gram language model. This feature is numeric. When back-off is used, it is recorded as back-off behavior feature (**BO**). This feature is symbolic.

Because most of the CRF implementations are designed to work with symbolic features, the numeric features (**CMS, DUR, WLM, AM, NLW, NRW, DEN, CLM**) should be converted into discrete features. Moreover, for the symbolic features (**LEX, POS, BO**), the contextual information of the current unit (character) is also incorporated by adding features of the preceding two characters and the following two characters.

For the selector CRF (**CRF-1**), features from the GMM hypothesis and the DNN hypothesis are concatenated together, and the complementary information from both independent ASR systems can help make better classification.

For the verifier CRF (**CRF-2**), it is difficult to use the ASR-based features for the selected hypothesis, because the features from two different types of ASR system have different dynamic ranges [55][56]. The text-based features are also recalculated after classification by the selector CRF (**CRF-1**) because of the context change. Additional feature used is the posterior probability output of **CRF-1** (for the mismatching cases) and the CMS of the DNN system (for the matching cases) as shown in Table 5.6.

5.3.5 Data Selection for Acoustic Model Training

The ASR hypotheses are merged into a single character sequence after the matching and selection processes, and every character in the sequence will have a label, either "*accept*" or "*discard*", based on the verification process according to Figure 5.2. Then, a decision should be made whether or not this sequence of the data should be used for acoustic model training. Two kinds of data selection scheme are investigated as follows:

5.3.5.1 Utterance-level Selection

The most commonly used utterance-level selection is based on utterance-level CMS, which is formulized as follows:

$$C_{sent} = \frac{1}{N}\sum_{i=1}^{N} C_{wi}$$

where C_{wi} is the posterior probability of word w_i obtained by confusion network decoding [26] and N is the number of words in the utterance. Then the utterances can be sorted by utterance-level CMS and select a certain percentage of top utterances for model training.

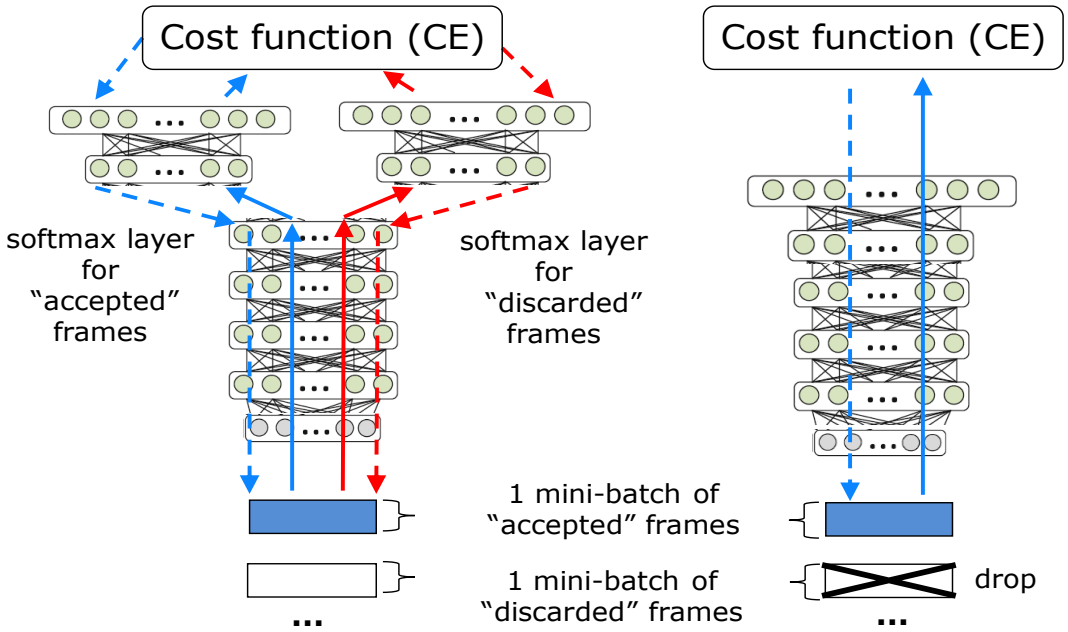
In the proposed method, the character acceptance rate (**CA**) is computed for every utterance to select utterances. Since Chinese is syllabic language and each character is a syllable, the "**CA**" actually means the ratio of "accepted" syllables over the total number of syllables in an utterance.

However, it is not practical to tune the **CA** threshold by using the development set, as it would take so long to train the DNN model for each **CA** threshold value. Considering spoken Chinese is highly homophonic, some character errors existing in the utterances are tolerated and the **CA** threshold is set to 70%.

5.3.5.2 Frame-level Selection

The frame-level data selection based on frame dropping and multi-task training methods is also implemented. The acceptance of each frame is determined, so the parameters of DNN are updated on the selected frame-level mini-batches. Using forced-alignment, the state-level labels and their boundaries are obtained. In this way, the character-level labels can be distributed to all frames. With the frame-level selection, we can train DNN model by either multi-task training method shown in Figure 5.3 (a) or frame dropping method shown in Figure 5.3 (b).

Each mini-batch (256 frames) consisting of either "accepted" frames or "discarded" frames is prepared, and all of the mini-batches are shuffled. In the multi-task training method, the "accepted" mini-batches and the "discarded" mini-batches update the shared hidden layers but update different softmax layers. Then, only the softmax layer for "accepted" frames is preserved after training. In the frame dropping method, only the "accepted" mini-batches are used to update the whole network.



(a) multi-task training (b) frame dropping

Figure 5.3: Frame-level data selection methods.

5.4 Experimental Evaluations

The proposed method is applied to CCLR-USV (184 lectures, 115 hours) to make an enhanced acoustic model, which are tested on CCLR-DEV and CCLR-TST.

5.4.1 Classifier Implementations

5.4.1.1 Training and Testing Data for Classifiers

In the implementation, the CRF classifiers are trained using CCLR-SV: **CRF-1**, which is trained to discriminate ***C3***+***C5*** vs. ***C4***, and **CRF-2**, which is trained to verify the output of **CRF-1** (***C4***+***C5*** vs. ***C3***) and to discriminate ***C1*** vs. ***C2***.

Since the feature of **CRF-2** depends on the result of **CRF-1**, a five-fold cross-validation method is used to get the features of **CRF-2**. Specifically, the training data is partitioned into five subsets, and an individual **CRF-1** trained using 4/5 of the data applied to the rest 1/5 data.

5.4.1.2 Training Data Resampling

In the training data set (CCLR-SV), there is serious imbalance in training samples between classes. The distribution of these patterns in CCLR-SV is shown in Table 5.4. It

is observed that 75.2% of them are categorized into ***C1***. Other four classes are 6.8% (***C2***), 6.6% (***C3***), 7.7% (***C4***) and 3.7% (***C5***), respectively. This distribution will bias the training of the classifiers. Thus, a re-sampling technique is introduced. Specifically, part of samples which appear too frequently is discarded in ***C1***. As a result, the calibrated distributions are as follows: ***C1***: 60.3%, ***C2***: 10.9%, ***C3***+***C5***: 16.6% and ***C4***: 12.2%.

5.4.1.3 Incorporating Data from Captioned Data by using Partial Annotation

For improving **CRF-2**, data from CCLR-LSV is also incorporated to enlarge the training data. This process is not direct, because there are only closed caption texts instead of faithful transcripts.

This issue can be addressed by adopting partial annotation, which has been widely explored for active learning or semi-supervised training CRF (see Appendix 2) in NLP field, e.g. CRF-based part-of-speech (POS) tagging, word segmentation [128][129], named entity recognition tasks [130]. In these works, conditional probabilities over partially annotated data are formulated. Training is achieved by modification to the learning objective function, incorporating partial annotation likelihood, so that a single model can be trained consistently with a mixture of full and partial annotation [131].

Here, a voting mechanism is proposed to generate both of the full and partial annotations from CCLR-LSV (see Figure 5.4). A three-way character alignment is made among the two ASR hypotheses by the GMM-based system (ASR1) and the DNN-based system (ASR2) and the closed caption texts. The insertion and deletion cases are regarded as a null token. The all-matching cases are regarded as positive samples, the all-mismatching cases are regarded as negative samples and others are partial label, and then they are added to the re-sampled training data (see Subsection 5.4.1.2) of **CRF-2**.

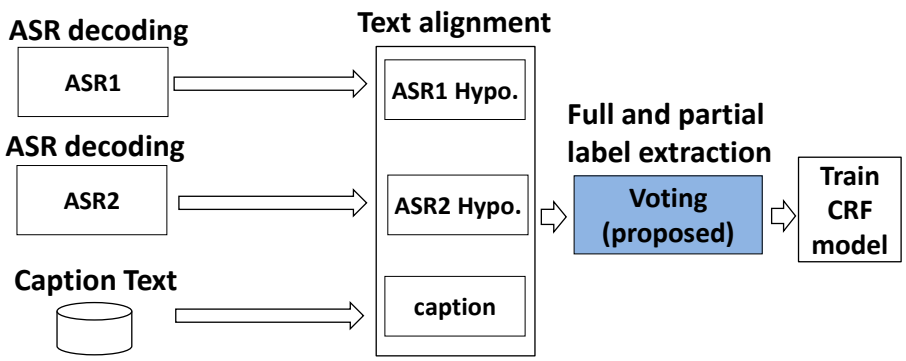


Figure 5.4 Extraction of Partially Annotated Data by Voting

Table 5.7 and Table 5.8 is the training data of **CRF-1** and **CRF-2** after data resampling and incorporating data from CCLR-LSV.

Table 5.7 Training Data of CRF-1

Data Resource	#(***C3***+***C5***) Select GMM	# ***C4*** Select DNN
CCLR-SV	51.3K	38.0K

Table 5.8 Training Data of CRF-2

Data Resource	#positive	#negative	#partial
CCLR-SV	288.0K	60.2K	/
CCLR-LSV	250.9K	0.1K	324.3K

5.4.1.4 Training Settings for CRF Classifiers

In the experiments, the implement of partial CRF [128], which is based on an open source liner-chain CRF toolkit CRFSuite[14], is used. The standard Limited-memory BFGS (L-BFGS) [57] algorithm and L2 regularization are used to train the CRF models with the sparse features of a high dimension. The cut-off threshold for the occurrence frequency of feature is 1. The maximum number of iterations for L-BFGS optimization is 100. To minimize the information loss in the quantization, these numeric values are discretized with the method[15] described in [58]. The same kind of numeric features from the DNN and GMM based systems can have different quantization levels.

[14] Available at http://www.chokkan.org/software/crfsuite/
[15] Available at http://www.irisa.fr/texmex/people/raymond/Tools/tools.html

5.4.2 Classification Accuracy of CRF Classifiers

Classification performance with various feature sets is evaluated on CCLR-DEV, as shown in Table 5.9 and Table 5.10. Performance is measured by precision, recall and F-score:

$$Precision = TP / FP$$
$$Recall = TP / (FP + FN)$$
$$F - score = 2 \times Precision \times Recall / (Precision + Recall)$$

where *TP* is true positive (correct output), *FP* is false positive (false alarm), and *FN* is false negative (miss). It is observed that the overall performance of **CRF-2** (Table 5.10) is higher than that of **CRF-1** (Table 5.9). It suggests selection of the hypothesis is more difficult than verification of the hypothesis.

Table 5.9 Feature Set Evaluation of CRF-1 on CCLR-DEV.

	CRF-1					
	Select GMM (*C3* + *C5*)			Select DNN (*C4*)		
Feature	Recall	Precision	F-score	Recall	Precision	F-score
LEX	0.504	0.498	0.501	0.711	0.716	0.713
POS	0.458	0.449	0.453	0.681	0.689	0.685
CLM	0.471	0.530	0.499	0.763	0.717	0.739
BO	0.300	0.481	0.370	0.816	0.673	0.738
All Text	0.546	0.560	0.553	0.756	0.746	0.751
CMS	0.518	0.541	0.529	0.750	0.733	0.741
DUR	0.491	0.511	0.501	0.733	0.717	0.725
WLM	0.410	0.485	0.444	0.753	0.692	0.721
AM	0.468	0.498	0.483	0.732	0.708	0.720
NLW	0.491	0.455	0.472	0.667	0.697	0.682
NRW	0.491	0.465	0.478	0.679	0.701	0.690
DEN	0.483	0.458	0.470	0.677	0.697	0.687
All ASR	0.572	0.569	0.570	0.754	0.756	0.755
All Features	**0.610**	**0.617**	**0.613**	**0.785**	**0.780**	**0.782**

Table 5.10 Feature Set Evaluation of CRF-2 on CCLR-DEV.

	CRF-2					
	Discard (*C2+ C3*)			Accept (*C1+C4+C5*)		
Feature	Recall	Precision	F-score	Recall	Precision	F-score
LEX	0.044	0.697	0.082	0.996	0.832	0.907
POS	0.002	0.730	0.003	0.999	0.826	0.905
CLM	0.088	0.684	0.155	0.992	0.838	0.908
BO	0.013	0.679	0.025	0.999	0.828	0.905
All Text	0.237	0.662	0.350	0.975	0.859	0.913
CMS(ASR)	0.631	0.588	0.609	0.907	0.921	0.914
All Features	0.621	0.627	**0.624**	0.922	0.920	**0.921**

Among the feature sets, the text-based features and their combinations are generally less effective than the ASR-based feature in **CRF-1** and **CRF-2**. However, for both classifiers, combination of both feature sets shows further improvement. As an individual feature, the **CMS** feature is the most effective for **CRF-1** and **CRF-2**.

From these results, the complete feature set is adopted. Although errors by **CRF-1** in the first stage of the classification is inevitable, part of them are detected and discarded in the second stage of classification by **CRF-2**, as shown in Figure 5.2.

5.4.3 Performance of Hypothesis Selection and Verification

Next, the performance of selection and verification of ASR hypotheses is evaluated using CCLR-DEV and CCLR-TST.

The GMM-HMM and DNN-HMM baseline systems are described in Chapter 3. Other methods compared with the proposed method are as follows:

- **Combine-ROVER:** the hypothesis and CMS derived from the ROVER-based system combination (the conventional method).
- **Combine-single-CRF:** a five-class CRF model to combine the ASR hypothesis is trained.
- **Combine-CRFs:** two classifiers for system combination (the proposed method) are trained. Different stages on the proposed cascade classification are also tested: **Combine-CRFs$_{(CRF-1)}$** for evaluating the effectiveness of the selection process only and **Combine-CRFs$_{(CRF-1+CRF-2)}$** to evaluate the effectiveness of the verification process.

Following metrics are used for the evaluation.

- **Character Error Rate (CER):** ASR evaluation measure after the hypothesis combination.
- **Normalized Cross Entropy (NCE):** It assigns the information gain to each of the hypothesis word to evaluate the quality of the confidence score distribution [59]. Higher values of NCE indicate better ASR confidence estimation. Perfect ASR confidence estimates give an NCE of 1. The definition of NCE is as follows:

$$NCE = \left\{ H_{max} + \sum_{correct} log_2(\hat{p}(w)) + \sum_{incorrect} log_2(1 - \hat{p}(w)) \right\} / H_{max}$$

$$H_{\max} = -n \log_2(p_c) - (N - n) \log_2(1 - p_c)$$

where n is the number of correct hypothesis words, N is the total number of hypothesis words, p_c is the average probability that an output word is correct ($=n / N$), $\hat{p}(w)$ is the confidence measure output of output word w.

- **Equal Error Rate (EER):** the false alarm rate or the miss rate at the confidence score threshold where the false alarm and the miss rate get equal. Lower values of EER indicate better ASR confidence estimation. Perfect ASR confidence estimates give an EER of zero.

The results are listed in Table 5.11. The proposed method **Combine-CRFs** outperforms the other methods. It is observed the combination of hypotheses by ROVER method (**Combine-ROVER**) can effectively reduce the recognition error rate (around absolute 1%) from the best single system (**DNN-HMM**), but it does not improve the confidence estimation. Using a single CRF classifier (**single-CRF**) can largely improve the confidence estimation, but it does not lead to the reduction of the recognition error rate. The proposed method (**Combine-CRFs**) shows robustness to the small number of complementary systems and different distributions of CMS between the DNN-based system and GMM-based system. The **CRF-1** improves the recognition

result of the ROVER method (around absolute 1%) and **CRF-2** further improves the confidence estimation quality based on the **CRF-1** classification result.

Table 5.11 Evaluation of the Data Selection and Verification.

	CCLR-DEV			CCLR-TST		
	CER (%)	NCE	EER (%)	CER (%)	NCE	EER (%)
GMM-HMM	24.2	-0.45	35.1	27.5	-0.29	29.5
DNN-HMM	22.7	-0.08	28.7	25.7	-0.13	30.2
Combine-ROVER	21.7	-0.21	34.9	24.9	-0.18	35.4
single-CRF	21.9	0.28	21.7	25.7	0.25	22.8
Combine-CRFs $_{(CRF-1)}$	**20.5**	0.28	18.2	**24.0**	0.25	19.3
Combine-CRFs $_{(CRF-1+CRF-2)}$	**20.5**	**0.37**	**17.1**	**24.0**	**0.34**	**18.5**

5.4.4 Performance of DNN Acoustic Model Enhanced by Selected Data

Then, the DNN-based acoustic model training is conducted by adding the data selected from CCLR-USV to the CCLR-SV and CCLR-LSV. ASR performance of the model enhanced by the selected data is evaluated on both of CCLR-DEV and CCLR-TST. The proposed data selection method is compared with other methods as follows:

- **Baseline GMM and baseline DNN:** the models are trained by only using CCLR-SV and CCLR-LSV as described in Chapter 3. The only difference is the language model is updated with newly collected lecture caption texts.
- **DNN (CMS):** the utterances are selected from CCLR-USV using the baseline DNN system based on a threshold of averaged CMS score (CMS≥0.6). The optimal threshold was determined by using GMM (MLE) models and CCLR-DEV [33]. It is the most commonly used method. All of the ASR hypotheses of CCLR-USV from DNN based system are also used without any selection (CMS≥0.0).
- **Combine-ROVER:** combine the ASR hypotheses of CCLR-USV from the baseline GMM and the baseline DNN systems using ROVER [36]. The utterances are selected according to the optimal threshold of the averaged CMS score (CMS≥0.6). It is the conventional method for leveraging hypotheses and data selection. All of the combined ASR hypotheses of CCLR-USV are also used without any selection (CMS≥0.0). The hypothesis and CMS are derived from the ROVER-based system combination.

- **Combine-CRFs ($_{CA=1.0,\ CA\geq0.0\ and\ CA\geq0.7}$):** combine the ASR hypotheses of CCLR-USV from two different baseline systems by using a set of CRF models. This is the proposed method for leveraging hypotheses and data selection. Effect of data selection is investigated on three thresholds: CA≥0.0 (no selection), CA=1.0 (use utterances with all characters accepted), and CA≥0.7.

In this experiment, the same setting with the baseline system described in Chapter 3 is used for DNN acoustic model training and testing. The only difference is the language model is updated with newly collected lecture caption texts.

Table 5.12 ASR Performance (CER%) of Cross-Entropy DNN Model by Utterance-level Selection

	Amount of data (hours)		CER%	
	labeled	unlabeled	DEV	TST
Baseline GMM	97.2	0	24.2	27.5
Baseline DNN	97.2	0	22.7	25.7
DNN (CMS≥0.0)	97.2	114.7	22.3	25.4
DNN (CMS≥0.6)	97.2	97.1	22.2	25.4
Combine-ROVER (CMS≥0.0)	97.2	114.7	21.9	24.9
Combine-ROVER (CMS≥0.6)	97.2	82.3	21.9	25.0
Combine-CRFs (CA≥0.0)	97.2	114.7	21.5	24.4
Combine-CRFs (CA=1.0)	97.2	32.5	21.6	24.7
Combine-CRFs (CA≥0.7)	97.2	71.5	21.3	24.2

ASR performance in CER is listed in Table 5.12. The results show that the proposed semi-supervised training method significantly improved the baseline DNN system. It also outperforms all other methods on both evaluation data sets.

It is observed that both of Combine-CRFs and Combine-ROVER outperform the simple CMS-based selection DNN ($_{CMS\geq0.0\ and\ CMS\geq0.6}$). This suggests the system combination effectively leverages the quality of automatically generated transcription. The fact that the proposed method Combine-CRFs ($_{CA\geq0.0}$) further outperforms the Combine-ROVER ($_{CM\geq0.0}$) demonstrates the effectiveness of the CRF models using many features. The Combine-ROVER ($_{CMS\geq0.6}$) and Combine-ROVER ($_{CMS\geq0.0}$) has no significant difference, while the improvement by Combine-CRFs ($_{CA\geq0.7}$) is statistically

significant compared with the other two models ($CMS \geq 0.0$ and $CA = 1.0$) among the proposed method. This confirms the data selection with the verifier CRF has some effect for further improvement.

Finally, the frame-level verification result is also conducted as described in Subsection 5.3.5.2, where "accepted" frames are used for supervised learning. The frame dropping and the multi-task training methods are implemented. These two different methods are referred to Combine-CRFs (multi-task) and Combine-CRFs (drop-frames) respectively. Their ASR performance shows no significant difference compared with Combine-CRFs ($CA \geq 0.7$) in Table 5.13. However, the frame-level selection methods do not require any threshold tuning.

Table 5.13 ASR Performance (CER%) of Cross-Entropy DNN Model by Frame-level Selection.

	Amount of data (hours)		CER%	
	labeled	unlabeled	DEV	TST
Combine-CRFs (CA≥0.7)	97.2	71.5	21.3	24.2
Combine-CRFs (multi-task)	97.2	114.7	21.3	24.3
Combine-CRFs (drop-frames)	97.2	90.4	21.4	24.3

On the other hand, utterance-level selection is advantageous for conducting sequence discriminative training. The sMBR DNN models are trained by using three Cross-Entropy (CE) DNN models listed in Table 5.12: Combine-CRFs ($CA \geq 0.7$, $CA \geq 0.0$ and $CA = 1.0$), and their ASR performance is shown in Table 5.14. A significant improvement by Combine-CRFs ($CA \geq 0.7$) over the other two models ($CMS \geq 0.0$ and $CA = 1.0$) is achieved. The effectiveness of the proposed method is still maintained after sMBR training. That means the proposed data selection method also works for sequence discriminative DNN training.

Table 5.14 ASR Performance (CER%) of sMBR DNN Model.

	Amount of data (hours)		CER%	
	labeled	unlabeled	DEV	TST
Combine-CRFs (CA≥0.0)	97.2	114.7	20.9	23.3
Combine-CRFs (CA=1.0)	97.2	32.5	21.0	23.6
Combine-CRFs (CA≥0.7)	97.2	71.5	20.7	23.1

5.5 Conclusion

A novel method is proposed for hypothesis leveraging and data selection for semi-supervised training of DNN acoustic model. The method uses dedicated classifiers, which are trained with the training database of the baseline acoustic model, to combine complementary ASR hypotheses and select usable data for model training.

A cascaded classification scheme is designed based on a set of binary classifiers, which incorporates a variety of features. Experimental evaluations show that the proposed semi-supervised training method effectively filters usable data, and improves the ASR accuracy from the baseline model and in comparison with the conventional ROVER-based method and the CMS-based selection method.

6 CONCLUSIONS

6.1 Contribution of Thesis Work

This thesis addresses effective acoustic model training targeted on Chinese spoken lectures. While there are many projects on spoken lecture transcription, works on Chinese lectures or spontaneous speech in general were limited. There was no large corpus public available in this category. For a comprehensive study on ASR of spontaneous Chinese, a corpus of Chinese spoken lectures (CCLR) is compiled in this work.

Then, a progressive framework for acoustic model training by using all possible resources step by step is presented. This work uses limited annotated data to train a seed acoustic model by the conventional supervised training. Considering the automatic generated label will have a low accuracy when directly using the seed acoustic model, lightly-supervised training method is introduced to derive an enhanced seed acoustic model from the media archives with closed caption texts. Finally, semi-supervised training is adopted for further improvement from a large amount of media archives without any related text resource.

Since the conventional lightly-supervised acoustic model training discards too much data, a novel data selection method is proposed by exploiting a large amount of data with closed caption texts but not faithful transcripts. In the proposed method, a sequence of the closed caption text and that of the ASR hypothesis by the baseline system are aligned. Then, a set of dedicated classifiers is designed and trained to select the correct one among them or reject both. It is demonstrated that the classifiers can effectively

filter the usable data for acoustic model training without tuning any threshold parameters. A significant improvement in the ASR accuracy is achieved from the baseline system and also in comparison with the conventional method of lightly-supervised training based on simple matching and CMS.

Moreover, a novel semi-supervised training method is also investigated. In the proposed method, ASR hypotheses by complementary GMM and DNN based ASR systems have been obtained. Then, a set of CRF-based classifiers are trained to select the better hypothesis and verify the selected data. The combined hypothesis for acoustic model training shows higher quality compared to the conventional system combination method (ROVER). Moreover, compared with the conventional data selection based on CMS, the proposed method is demonstrated more effective for filtering usable data. A significant improvement in the ASR accuracy is achieved over the baseline system and in comparison with the models trained from the conventional system combination and data selection methods.

6.2 Summary of Performance Improvement

To summarise the effect of the proposed progressive framework for acoustic model training, the performance improvement on the CCLR-TST is shown in Figure 6.1.

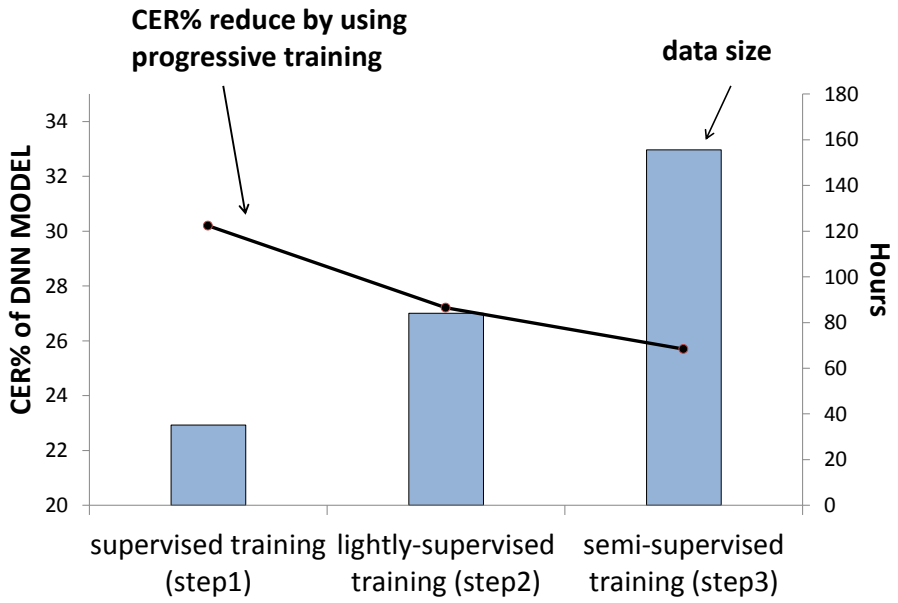


Figure 6.1: Performance Improvement by the Proposed Training Method.

As shown in Figure 6.1, this baseline system (step1) using 35.2 hours faithful data achieved an average Character Error Rate (CER) of 39.3% with the GMM (MLE) model, 36.7% with the GMM (MPE) model, and 30.2% with the DNN model.

The CER of the proposed lightly-supervised training (step2) applied to the DNN model is 27.2% with an absolute CER reduction of 3.0% from the baseline DNN model (30.2%) and the data increase is 48.9 hours.

The CER of the proposed semi-supervised training (step3) applied to the DNN model is 25.7% when the influence from new language model and decoder is removed (around 1.5%). The CER is reduced by absolute 1.5% from the DNN model with the lightly-supervised training (27.2%) and the data increase is 71.5 hours. The sMBR training can further improve the DNN model by 1.1%.

Without the progressive training scheme, the improvement cannot be so large only with the semi-supervised training. It is shown that the lightly-supervised training improved the baseline model by 3.0% with additional data of 48.9 hours, while the semi-supervised training improved only 1.5% with additional data of 71.5 hours. It is not difficult to figure out the upper bound of the improvement will be no more than 3.0%, if all data (120.4 hours) is used for the semi-supervised training without the progressive training scheme.

6.3 Future Work

For building a high-performance automatic transcription system of spoken lectures, there is still much room for further improvement.

First, while semi-supervised learning tries to exploit unlabelled data, the ensemble learning tries to achieve strong generalization by using the combination of multiple learners. An ensemble is typically constructed in two steps. First, a number of component learners are generated; then, the component learners are combined for prediction. Generally, to get a good ensemble, the component learners should be as more accurate as possible, and as more diverse as possible [153]. However, how to measure and control the diversity and co-training of complementary systems remain open problems, but worth investigating.

Second, the discriminative data selection method proposed in this work has potential to be extended to many ASR post-processing tasks, e.g. system combination. The most

recent work on log-linear system combination [150] and QE (quality estimation)-guided system combination method [154] may suggest the promising directions. For example, iCNC (improved confusion network combination) proposed in [160] can improve the combination accuracy by replacing the decision rules by classifiers, which can make favourable use of all the available information. Alternatively, the combination can be performed at the frame level as proposed in [146], which is based on the definition of a time frame-wise word error cost function in a minimum Bayes risk framework.

Third, it is observed that the proposed discriminative data selection method is effective for integrating diverse knowledge resources, either multiple ASR-based (experimental settings for semi-supervised training) or non-ASR based with ASR-based (experimental settings for lightly-supervised training). Thus, it is believed that this discriminative data selection method can motivate more generalized approaches for integrating diverse knowledge resources, such as fusion of transcriptions from multiple human annotators in a crowd-sourcing cooperation. Many researches, e.g. the algorithm proposed in [147] using the linear finite state transducers framework and the algorithm proposed in [148] using the WFST framework, can be adopted for these purposes.

Moreover, the proposed discriminative data selection method is actually independent of any existing classifiers, and thus it is possible to deploy the state-of-the-art classifiers such as LSTM-RNN [149] in this framework.

7 REFERENCES

[1] 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.

[2] H.Nanjo and T.Kawahara. Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition. IEEE-TSAP, Vol.12, No.4, pp.391-400, 2004.

[3] 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.

[4] J.Glass, T.J.Hazen, S.Cyphers, I.Malioutov, D.Huynh, and R.Barzilay. Recent Progress in the MIT Spoken Lecture Processing Project. In Proc. INTERSPEECH, pp. 2553-2556, 2007.

[5] H.Yamazaki, K.Iwano, K.Shinoda, S.Furui, and H.Yokota, Dynamic Language Model Adaptation Using Presentation Slides for Lecture Speech Recognition. In Proc. INTERSPEECH, pp. 2349-2352, 2007.

[6] T.Kawahara, Y.Nemoto, and Y.Akita, Automatic Lecture Transcription by Exploiting Slide Information for Language Model Adaptation. In Proc. ICASSP, pp.4929-4932, 2008.

[7] M.Paul, M.Federico, and S.Stucker, Overview of the IWSLT 2010 Evaluation Campaign. In Proc. IWSLT, pp. 3-27, 2010.

[8] K. Yu, M. Gales, L. Wang and P. Woodland, Unsupervised training and directed manual transcription for LVCSR. Speech Communication, Vol52 (7), pp.652-663, 2010.

[9] H. Liao, E. McDermott and A. Senior, Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In Proc. ASRU, pp. 368-373, 2013.

[10] Y. Huang, D. Yu, Y. Gong and C. Liu, Semi-supervised GMM and DNN acoustic model training with multi-system combination and confidence re-calibration. In Proc. INTERSPEECH, pp. 2360-2364, 2013.

[11] K. Vesely, M. Hannemann and L. Burget, Semi-supervised training of deep neural networks, In Proc. ASRU, pp267-272, 2013.

[12] D. Imseng, B. Potard, P. Motlicek, A. Nanchen, H. Bourlard, Exploiting un-transcribed foreign data for speech recognition in well-resourced languages. In Proc. ICASSP, 2014.

[13] P. Zhang, Y. Liu, and T. Hain, "Semi-supervised dnn training in meeting recognition," in Proc. SLT, 2014.

[14] H. Su, H. Xu, "Multi-softmax Deep Neural Network for Semi-supervised Training", In Proc. INTERSPEECH, 2015.

[15] V. Manohar, D. Povey and S. Khudanpur, "Semi-supervised Maximum Mutual Information Training of Deep Neural Network Acoustic Models", In Proc. INTERSPEECH, 2015.

[16] M. Harper, "IARPA Babel Program," 2014.

[17] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," In Proc. ICASSP, pp. 8619–8623, 2013.

[18] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in Proc. ICASSP, pp. 7304–7308, 2013.

[19] S. Thomas, M. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in Proc. ICASSP, pp. 6704–6708, 2013.

[20] Y. Liu and K. Kirchhoff. Graph-based Semi-supervised Acoustic Modeling in DNN based Speech Recognition. In Proc. SLT, USA, 2014.

[21] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels and J. Bilmes. Submodular Subset Selection for Large-Scale Speech Training Data. In Proc. ICASSP, Italy, 2014.

[22] H. Jiang, Confidence measures for speech recognition: a survey, Speech Communication, vol. 45, no. 4, pp. 455-470, Apr. 2005.

[23] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," In Proc. EUROSPEECH, vol. 2, pp. 827–830, September 1997.

[24] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," IEEE Trans. Speech & Audio Process., vol. 9, no. 3, pp. 288–298, March 2001.

[25] A. Lee, K. Shikano, and T. Kawahara. Real-time word confidence scoring using local posterior probabilities on tree trellis search, In Proc. ICASSP, Vol.1, pp.793-796, 2004.

[26] L. Mangu, E. Brill, and A. Stolcke. Finding Consensus Among Words: Lattice-Based Word Error Minimization. In Proc. EUROSPEECH, pp. 495-498, Budapest, 1999.

[27] H. Xu, D. Povey, L. Mangu, and J. Zhu. "An improved consensus-like method for Minimum Bayes Risk decoding and lattice combination." In Proc. ICASSP, 2010.

[28] V. Goel and W. J. Byrne. "Minimum Bayes-risk automatic speech recognition." Computer Speech and Language 14 (2), pp. 115–135, 2000.

[29] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. ICML, 2001.

[30] M. Seigel and P. Woodland, Combining Information Sources for Confidence Estimation with CRF Models, In Proc. INTERSPEECH, 2011.

[31] J. Fayolle, F. Moreau, C. Raymond, and G. Gravier, "CRF-based combination of contextual features to improve a posteriori wordlevel confidence measures," In Proc. INTERSPEECH, 2010.

[32] L. Lamel, J.L. Gauvain and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training", Computer Speech and Language, vol.16, pp. 115-129, January 2002.

[33] S. Li, Y. Akita, and T. Kawahara. Discriminative data selection for lightly supervised training of acoustic model using closed caption texts. In Proc. INTERSPEECH, pp.3526-3530, 2015.

[34] K. Audhkhasi, A Zavou, P Georgiou, and S. Narayanan, “Theoretical analysis of diversity in an ensemble of automatic speech recognition systems,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, volu.22, no. 3, March 2014.

[35] L. Deng and J. Platt, "Ensemble deep learning for speech recognition." In Proc. INTERSPEECH, 2014.

[36] J. Fiscus, A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER), In Proc. ASRU, 1997.

[37] G.E. Dahl, D. Yu, L. Deng and A. Acero, Context dependent pre-trained deep neural networks for large vocabulary speech recognition, IEEE Transactions on Audio, Speech, and Language Processing, 20(1):30–42, 2012.

[38] G. Hinton, S. Osindero, and Y. Teh, “A Fast Learning Algorithm for Deep Belief Nets”, Neural Computation, vol. 18, pp. 1527–1554, 2006.

[39] G. Hinton, “A Practical Guide to Training Restricted Boltzmann Machines”, Technical Report UTML TR 2010–003, Univ. of Toronto, 2010.

[40] D. Rumelhart, G. Hinton, and R. Williams, “Learning Representations by Back-Propagating Errors,” Nature, vol. 323, Oct. 1986.

[41] H. Su, G. Li, D. Yu, F. Seide, Error Back Propagation for Sequence Training of Context-Dependent Deep Neural Networks for Conversational Speech Transcription, In Proc. ICASSP, 2013.

[42] K. Vesely, A. Ghoshal, L. Burget and D. Povey, Sequence-discriminative training of deep neural networks, In Proc. INTERSPEECH, 2013.

[43] S. Li, Y. Akita, and T. Kawahara, Corpus and transcription system of Chinese lecture room, In Proc. ISCSLP, 2014.

[44] J. Zhang, H. Chan, P. Fung and L. Cao. A Comparative Study on Speech Summarization of Broadcast News and Lecture Speech. In Proc. INTERSPEECH, pp. 2781-2784, 2007.

[45] S. Kong, M. Wu, C. Lin, Y. Fu, and L. Lee. Learning on Demand - Course Lecture Distillation by Information Extraction and Semantic Structuring for Spoken Documents. In Proc. INTERSPEECH, pp. 4709-4712, 2009.

[46] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, The Kaldi speech recognition toolkit, In Proc. ASRU, 2011.

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

[48] M. Mimura and T. Kawahara, Fast Speaker Normalization and Adaptation Based on BIC for Meeting Speech Recognition, IEICE TRANSACTIONS on Information and Systems, vol.J95-D No.7.2012.

[49] J. Luo, L. Lamel and J-L. Gauvain,"Modeling Characters versus Words for Mandarin Speech Recognition." In Proc. ICASSP, Taipei, Taiwan, 2009.

[50] 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, Volume 133, Issue 1, 519-528, January 2013.

[51] M. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," In Proc. INTERSPEECH, 2006.

[52] W. Chen, S.Ananathakrishnan, R.Kumar, R.Prasad, and P.Natarajan, ASR error dection in a conversational spoken language translation system, In Proc. ICASSP, 2013.

[53] M. Lehr, I. Shafran, E. Prud'hommeaux, and B. Roark, Discriminative Joint Modeling of Lexical Variation and Acoustic Confusion for Automated Narrative Retelling Assessment, In Proc. NAACL, 2013.

[54] M. Shen, H. Liu, D. Kawahara, and S. Kurohashi. 2014. Chinese Morphological Analysis with Character-level POS Tagging. In proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL), Short Paper, pages 253–258, Baltimore, USA, 2014.

[55] Y.C. Tam, Y. Lei, J. Zheng and W. Wang, ASR Error Detection using Recurrent Neural Network Language Model and Complementary ASR, In Proc. ICASSP, 2014.

[56] J. Kim, J. Chong, I. Lane, "Efficient On-The-Fly Hypothesis Rescoring in a Hybrid GPU/CPU-based Large Vocabulary Continuous Speech Recognition Engine," In Proc. INTERSPEECH, 2012.

[57] J. Nocedal. "Updating Quasi-Newton Matrices with Limited Storage". Mathematics of Computation. 35. 151. 773-782. 1980.

[58] U. Fayyad and K. Irani, “Multi-interval discretization of continuous attributes for classification learning,” In Proc. IJCAI, pp1022-1027, 1993.

[59] M. Siu, H. Gish, and F. Richardson, “Improved estimation, evaluation and applications of confidence measures for speech recognition,” in Proc. EUROSPEECH, 1997.

[60] A.Rush, R. Reichert, M. Collins, and A. Globerson. Improved Parsing and POS Tagging Using Inter-Sentence Consistency Constraints, In Proc. EMNLP, 2012.

[61] S. Volkova, P. Choudhury, C. Quirk, B. Dolan, and L. Zettlemoyer. Lightly supervised learning of procedural dialog systems. In Proc. ACL, 2013.

[62] Y. Kim, M. Jeong, K. Stratos, and R. Sarikaya. Weakly Supervised Slot Tagging with Partially Labelled Sequences from Web Search Click Logs, In Proc. NAACL, 2015.

[63] R. Sproat, Lightly supervised learning of text normalization: Russian number names, In Proc. SLT, page 436-441, 2010.

[64] M. Huck, D. Vilar, D. Stein, and H. Ney. Lightly-Supervised Training for Hierarchical Phrase-Based Machine Translation. In EMNLP 2011 Workshop on Unsupervised Learning in NLP, pages 91-96, Edinburgh, UK, July 2011.

[65] H. Schwenk, Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In Proc. IWSLT, pages 182–189, 2008.

[66] Y. Yu, F. Zhu, X. Li, Y. Liu, J. Zou, Y. Yang, G. Yang, Z. Fan and X. Wu, Overview of SHRC-Ginkgo speech synthesis system for Blizzard Challenge 2013, In Proc. Blizzard Challenge, 2013.

[67] P. Bell, P. Swietojanski, J. Driesen, M. Sinclair, F. McInnes, and S. Renals, “The UEDIN ASR Systems for the IWSLT 2014 Evaluation, ” In Proc. IWSLT, 2014.

[68] P. Bell, F. McInnes, S. Gangireddy, M. Sinclair, A. Birch, S. Renals, “The UEDIN English ASR System for the IWSLT 2013 Evaluation,” In Proc. IWSLT, 2013.

[69] J. Driesen, P. Bell, and S. Renals, “Description of the UEDIN system for german ASR,” In Proc. IWSLT, 2013.

[70] A. Aue, Q. Gao, H. Hassan, X. He, G. Li, N. Ruiz, F. Seide, “MSR-FBK IWSLT 2013 SLT system description”, In Proc. IWSLT, 2013.

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

[72] P. Shen, X. Lu, X. Hu, N. Kanda, M. Saiko and C. Hori, "The NICT asr system for IWSLT 2014," In Proc. IWSLT 2014, pp. 113–118, Dec. 2014.

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

[74] J.Zhang, H.Chan, P.Fung and L.Cao. A Comparative Study on Speech Summarization of Broadcast News and Lecture Speech. In Proc. INTERSPEECH, pp. 2781-2784, 2007.

[75] S.Kong, M.Wu, C.Lin, Y.Fu, and L.Lee. Learning on Demand - Course Lecture Distillation by Information Extraction and Semantic Structuring for Spoken Documents. In Proc. INTERSPEECH, pp. 4709-4712, 2009.

[76] J. Zhang and P. Fung, Automatic Parliamentary Meeting Minute Generation Using Rhetorical Structure Modeling. In IEEE Transactions on Audio, Speech and Language Processing, Volume 20, Issue 9, November 2012.

[77] S.Li, X.Lu, Y.Akita, and T.Kawahara. Ensemble speaker modeling using speaker adaptive training deep neural network for speaker adaptation. In Proc. INTERSPEECH, pp.2892--2896, 2015.

[78] M. Seltzer and J. Droppo, “Multi-task learning in deep neural networks for improved phoneme recognition,” In Proc. ICASSP, pp. 6965–6968, 2013.

[79] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” In Proc. ICASSP, pp. 7304 – 7308, 2013.

[80] T. Ochiai, S. Matsuda, X. Lu, C. Hori and S. Katagiri, “Speaker adaptive training using deep neural networks,” In Proc. ICASSP, 2014.

[81] J. Xue, J. Li, D. Yu, M. Seltzer, and Y. Gong, “Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network,” In Proc. ICASSP, 2014.

[82] J. Xue, J. Li, and Y. Gong, “Restructuring of deep neural network acoustic models with singular value decomposition,” In Proc. INTERSPEECH, pp. 2365-2369, 2013.

[83] 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.

[84] D.Yu, K.Yao, 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.

[85] S. Davis and P. Mermelstein, “Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. ASSP, vol. 28, no. 4, pp. 357–366, 1980.

[86] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” Journal Acoust. Soc. Amer., vol. 87, no. 4, pp. 1738–1752, 1990.

[87] P. Woodland, M. Gales, and D. Pye, “Improving environmental robustness in large vocabulary speech recognition,” In Proc. ICASSP, vol. 1, 1996, pp. 65–68.

[88] F. Honig, S. George, H. Christian, B. Fabio. Revising perceptual linear prediction (PLP). In Proc. INTERSPEECH, pp. 2997–3000, 2005.

[89] C. Chelba and F. Jelinek. 2000. Structured language modeling. Computer Speech & Language, 14(4):283– 332.

[90] M. Collins. 2003. Head-driven statistical models for natural language parsing. Computational Linguistics, 29(4):589–637.

[91] Y. Bengio and R. Ducharme. A neural probabilistic language model. In Advances in Neural Information Processing Systems (NIPS 13), 2001.

[92] Viterbi AJ (April 1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory 13 (2): 260–269. doi:10.1109/TIT.1967.1054010. (note: the Viterbi decoding algorithm is described in section IV.)

[93] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, “Sphinx-4: A flexible open source framework for speech recognition,” Sun Microsystems Inc., Technical Report SML1 TR2004-0811, 2004.

[94] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Loof, R. Schluter, and H. Ney, “The RWTH Aachen University Open Source Speech Recognition System,” In Proc. INTERSPEECH, 2009, pp. 2111–2114.

[95] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for version 3.4). Cambridge University Engineering Department, 2009.

[96] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," Computer Speech and Language, vol. 20, no. 1, pp. 69–88, 2002.

[97] M. Mohri, "Finite-state transducers in language and speech processing," Computational Linguistics, vol. 23, no. 2, pp. 269–311, 1997.

[98] D. Paul and J. Baker, "The design for the Wall Street Journal based CSR corpus," In Proc. ICSLP, 1992.

[99] D. Graff, et al. 1996 English Broadcast News Speech (HUB4) LDC97S44. Web Download. Philadelphia: Linguistic Data Consortium, 1997.

[100] D. Moore, J. Dines, M. Magimai Doss, J. Vepa, O. Cheng, and T. Hain, "Juicer: A weighted ſnite-state transducer speech decoder," in MLMI'06, 2006.

[101] P. Dixon, D. Caseiro, T. Oonishi and S. Furui. 2007. The Titech Large Vocabulary WFST Speech Recognition System, In Proc. ASRU, pp. 1301-1304.

[102] P.R. Dixon, C. Hori, and H. Kashioka, "Development of the SprinTra WFST Speech Decoder," NICT Research Journal, pp 15-20, 2012.

[103] Y. Miao, M. Gowayyed, F. Metze. EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-Based Decoding. In Proc. ASRU, 2015.

[104] D. Amodei, et al. "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin." arXiv preprint arXiv:1512.02595 (2015).

[105] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," In Proc. ASRU, 2009.

[106] G. Zweig and Nguyen, "SCARF: A segmental conditional random field toolkit for speech recognition," In Proc. INTERSPEECH, 2010.

[107] 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.

[108] Y.Long, M.J.F.Gales, P.Lanchantin, X.Liu, M.S.Seigel and P.C.Woodland. Improving Lightly Supervised Training for Broadcast Transcription. In Proc. INTERSPEECH, 2013.

[109] J.Driesen and S.Renals. Lightly supervised automatic subtitling of weather forecasts, In Proc. ASRU, 2013.

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

[111] H. Chan, J. Zhang, P. Fung and L. Cao, “A Mandarin Lecture Speech Transcription System for Speech Summarization” In Proc. ASRU, 2007.

[112] 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.

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

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

[115] S. Tseng, “Repairs and repetitions in spontaneous Mandarin”, in Proc DiSS’03, Disfluency in Spontaneous Speech Workshop, Göteborg University, Sweden, 5–8 September 2003.

[116] 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.

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

[118] 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.

[119] L.Nguyen and B.Xiang. Light Supervision in Acoustic Model Training. In Proc. ICASSP, Vol. 1, pp. I-185, 2004.

[120] H.Chan and P.Woodland. Improving Broadcast News Transcription by Lightly Supervised Discriminative Training. In Proc. ICASSP, Vol. 1, pp. 737-740, 2004.

[121] H.Lin, and J.Bilmes, How to select a good training-data subset for transcription: submodular active selection for sequences, In Proc. INTERSPEECH, pp.2859-2862, 2009.

[122] T. Lavergne, O. Cappé, and F. Yvon. Practical Very Large Scale CRFs. In Proc. ACL, pages 504-513, July 2010.

[123] N. Sokolovska, T. Lavergne, O. Cappé, and F. Yvon. Efficient Learning of Sparse Conditional Random Fields for Supervised Sequence Labelling. IEEE J. Sel. Topics Signal Process., 4(6):953-964, December 2010.

[124] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition," PhD thesis, Cambridge University Engineering Dept, 2003.

[125] P. Huang, K. Kumar, C. Liu, Y. Gong, and L. Deng, "Predicting Speech Recognition Confidence Using Deep Learning With Word Identity and Score Features," In Proc. ICASSP, pp. 7413–7417, 2013.

[126] K. Walker, et al. GALE Phase 2 Chinese Broadcast Conversation Speech LDC2013S04. DVD. Philadelphia: Linguistic Data Consortium, 2013.

[127] J. Bergstra, F. Bastien, O. Breuleux, P. Lamblin, R. Pascanu, O. Delalleau, et al., “Theano: Deep learning on gpus with python,” In Proc. Big Learning Workshop, NIPS, vol.11, pp.1–6, 2011.

[128] Y. Liu, Y. Zhang, W. Che, T. Liu, and F. Wu, “Domain adaptation for crf-based chinese word segmentation using free annotations,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 864–874.

[129] F. Yang and P. Vozila, “Semi-supervised chinese word segmentation using partial-label learning with conditional random fields,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 90–98.

[130] T. Sasada, S. Mori, T. Kawahara, and Y. Yamakata. Named entity recognizer trainable from partially annotated data. In Proc. PACLING, pp.10--17, 2015.

[131] Y. Tsuboi, H. Kashima, S. Mori, H. Oda, and Y. Matsumoto, “Training conditional random fields using incomplete annotations,” in Proceedings of the 22nd International Conference on Computational Linguistics, 2008.

[132] L. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. The annals of mathematical statistics, pages 1554–1563. 1966.

[133] J. Kong, and G. David, TDT4 Multilingual Broadcast News Speech Corpus LDC2005S11. DVD. Philadelphia: Linguistic Data Consortium, 2005.

[134] B. Juang, Maximum-likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains, AT&T Technical Journal, 64(6), 1235-1249, 1985.

[135] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," In Proc. A Speaker Odyssey - The Speaker Recognition Workshop, Jun. 2001, pp. 213–218.

[136] C.L. Leggetter and P.C. Woodland: Speaker Adaptation of HMMs using Linear Regression, Computer Speech and Language, pp.171-185, 1995.

[137] N. Xue, , F. Fei and M. Palmer, The Penn Chinese Treebank: Phrase structure annotation of a large corpus. Natural Language Engineering 11(2), 207–238. 2005.

[138] N. Jaitly, P. Nguyen, A.W. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," In Proc. INTERSPEECH, 2012.

[139] C. Huang, P. Dixon, S. Matsuda, Y. Wu, X. Lu, M. Saiko and C. Hori, "The NICT ASR System for IWSLT 2013," In Proc. IWSLT, 2013.

[140] C. Alberti and M. Bacchiani, "Automatic captioning in YouTube," Dec. 2009.

[141] C. Zhang and P. C. Woodland, "Standalone training of context dependent deep neural network acoustic models," In Proc. ICASSP, Florence, Italy, 2014.

[142] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," In Proc. ICASSP, 2014.

[143] A. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling", In Proc. ICASSP, pp. 4273-4276, 2012.

[144] Y. Kubo, T. Hori, A. Nakamura, "Large vocabulary continuous speech recognition based on WFST structured classifiers and deep bottleneck features", In Proc. ICASSP, 2013.

[145] M. Mimura and T. Kawahara, “Investigation on speaker adaptation of DNN-HMM for Japanese spoken lecture transcription by using CSJ”, IPSJ SIG-SLP-97-9, 2013.

[146] B. Hoffmeister et al. “Frame based system combination and a comparison with weighted ROVER and CNC”, In Proc. ICSLP, 2006.

[147] Benoit Favre, Mickael Rouvier, and Frederic Bechet, “Reranked aligners for interactive transcript correction,” In Proc. ICASSP, Florence, Italy, May 2014.

[148] M. Rouvier, B. Favre, F. Bechet, “Joint Decoding of Complementary Utterances,” In Proc. SLT, 2014.

[149] S.Hochreiter and J.Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[150] J. Yang, C. Zhang, A. Ragni, M. Gales and P. Woodland, “System Combination with Log-linear Models”, In Proc. ICASSP, 2016.

[151] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson and O. Vinyals, "Learning the Speech Front-end with Raw Waveform CLDNNs," in Proc. INTERSPEECH, 2015.

[152] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan and M. Bacchiani, "Factored Spatial and Spectral Multichannel Raw Waveform CLDNNs," In Proc. ICASSP, March 2016.

[153] A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, Advances in Neural Information Processing Systems 7, pages 231–238. MIT Press, Cambridge, MA, 1995.

[154] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA), pp 223–231, 2006.

[155] V. Nair and G. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in In Proc. ICML, pp.131–136, 2010.

[156] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” Journal of Machine Learning Research, vol. 15, pp. 315323, 2011.

[157] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, “Maxout networks,” arXiv:1302.4389, 2013.

[158] X. Zhang, J. Trmal, D. Povey and S. Khudanpur, "Improving Deep Neural Network Acoustic Models using Generalized Maxout Networks", In Proc. ICASSP, 2014.

[159] S. Jalalvand, M. Negri, D. Falavigna and M. Turchi “Driving Rover with segment-based ASR quality estimation.", In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), Beijing, China 26 -31 July 2015.

[160] B. Hoffmeister, R. Schluter, and H. Ney, “icnc and irover: The limits of improving system combination with classification?” In Proc. INTERSPEECH, pp. 232–235, 2008.

8 APPENDICES

APPENDIX 1: SPEAKER ADAPTIVE TRAINING DNN 98

APPENDIX 2: TRAINING CRF WITH PARTIALLY ANNOTATION..... 107

LIST OF PUBLICATIONS BY THE AUTHOR..... 111

APPENDIX 1: SPEAKER ADAPTIVE TRAINING DNN

This thesis work introduces an ensemble speaker modelling using a speaker adaptive training (SAT) deep neural network (SAT-DNN). A speaker-independent DNN (SI-DNN) acoustic model can first be trained as a universal speaker model (USM). Based on the USM, a SAT-DNN is used to obtain a set of speaker-dependent models by assuming that all other layers except one speaker-dependent (SD) layer are shared among speakers. The speaker ensemble matrix is created by concatenating all of the SD neural weight matrices. With matrix factorization technique, an ensemble speaker subspace is extracted. When testing, an initial model for each target speaker is selected in this ensemble speaker subspace. Then, adaptation is carried out to obtain the final acoustic model for testing. In order to reduce the number of adaptation parameters, low-rank speaker subspace is further explored. The algorithm is tested on lecture transcription task. Experimental results showed that the proposed method is effective for unsupervised speaker adaptation.

A1.1. Ensemble speaker modelling using speaker adaptive training DNN

Rather than using only one SI-DNN model as an initial model in adaptation, many SD-DNN models should be prepared, and choose the best one among them as an initial model for adaptation. The basic procedure is as follows:

a) Train a USM, i.e., SI-DNN.
b) Taking the USM as an initial model, train speaker-dependent models, i.e., SD-DNNs. For training, a multi-task learning architecture for SAT-DNN is adopted.
c) Factorize the speaker-dependent weight matrices using SVD and obtain speaker-specific coefficient matrices. Then, perform low-rank matrix approximation to reduce the number of adaptation model parameters.
d) Perform adaptation for a testing speaker by picking up an initial model in the speaker subspace.

In the following subsections, each stage of the procedure is described in details.

A1.1.1 Multi-task Learning Architecture for SAT-DNN

In multi-task learning, it is supposed that some model parameters are shared by all tasks and each task has its own task-dependent parameters. It is shown that this multi-task learning strategy achieves better generalization than single-task learning strategy in various task domains such as phone recognition [78] and multilingual speech recognition [17][18]. The SAT-DNN proposed in [80] can be regarded as a multi-task learning. In NICT-SAT-DNN [80], the DNN architecture is configured as shown in Figure 8.1. All of the DNN layers are shared among speakers except one SD layer. The parameters in the SD layer are updated only for a specific speaker while the parameters for all of the shared layers are updated for all speakers. Explicitly specifying one layer as an SD layer in training makes training focus much more on speaker adaptation in DNN. In speaker adaptive training, the initial model parameters are set as the model parameters of an SI-DNN model.

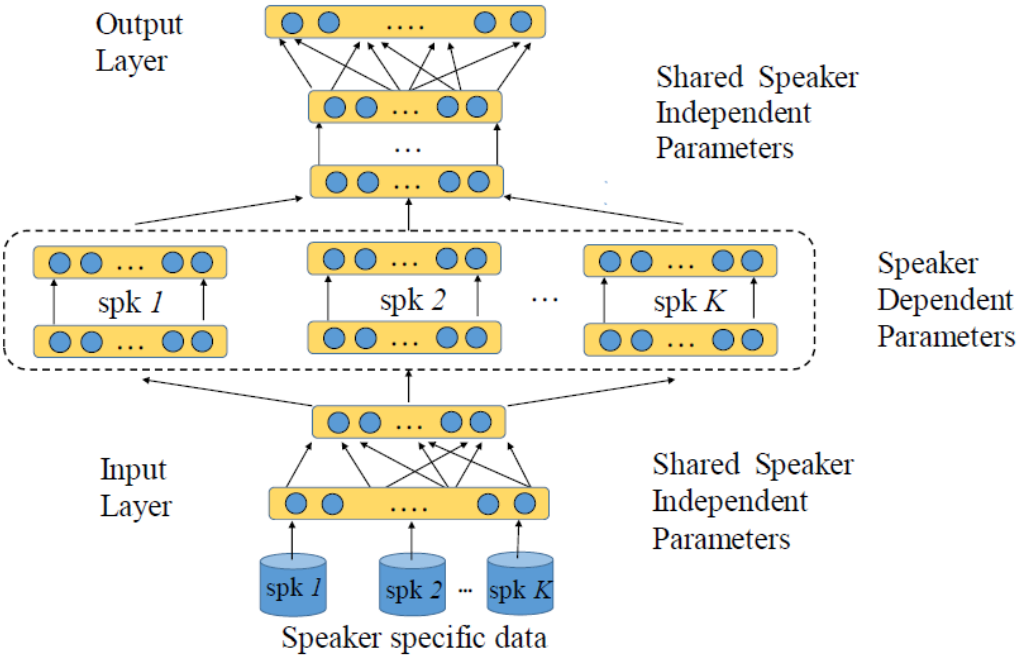


Figure 8.1: Multi-task Learning Architecture for SAT-DNN.

A1.1.2 Ensemble Speaker Matrix Factorization

From SAT-DNN introduced in Section A1.1.1, a set of SD-DNN models (with shared neural weight matrices) are obtained. Suppose the SD-DNN model is represented as the neural weight matrix of the SD layer as

$$\left\{\mathbf{W}_{sd}^{i} \in R^{m*n}, i = 1,2,\ldots,K\right\}$$

where K is the total number of speakers, m and n are the numbers of neurons for input and output, respectively, of the SD layers. The ensemble speaker matrix is composed by concatenating these matrices as

$$\mathbf{W}_{sd} \overset{\Delta}{=} [\mathbf{W}_{sd}^{1}, \mathbf{W}_{sd}^{2}, \ldots, \mathbf{W}_{sd}^{K}] \in R^{m*l}, l = n * K$$

Based on SVD matrix decomposition [17, 18], it is decomposed as

$$\mathbf{W}_{sd} = \mathbf{U} * \mathbf{S} * [(\mathbf{V}_{sd}^{1})^{T}, (\mathbf{V}_{sd}^{2})^{T}, \ldots, (\mathbf{V}_{sd}^{K})^{T}] \tag{A1.1}$$

In this equation, $\mathbf{U} \in R^{m*n}$ is the left singular matrix, $\mathbf{S} \in R^{n*n}$ is a diagonal matrix with elements as singular values. $(\mathbf{V}_{sd}^{i})^{T} \in R^{n*n}$ is the speaker coefficient matrix of the i-th speaker that satisfies:

$$\mathbf{W}_{sd}^{i} = \mathbf{U} * \mathbf{S} * (\mathbf{V}_{sd}^{i})^{T} \tag{A1.2}$$

In DNN, this matrix factorization can be implemented as in Figure 8.2.

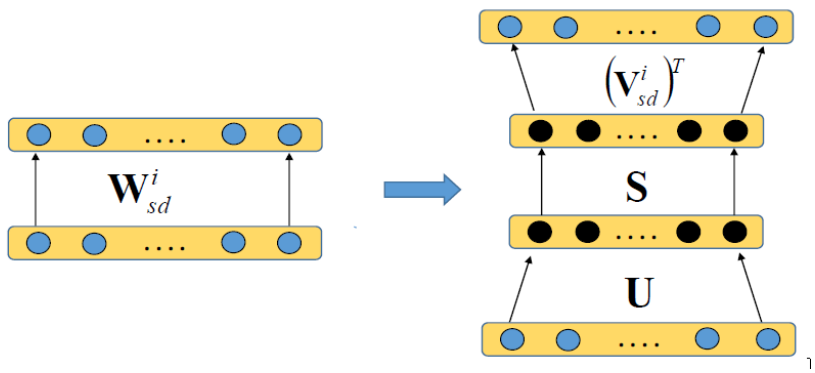


Figure 8.2: Before (left) and After (right) Matrix Factorization in One Layer of DNN.

In this figure, black balls represent linear response neurons. The total transform effect of the factorized matrix is the same as only using one transform matrix $\mathbf{W}_{sd}^{i}$ [82].

A1.1.3 Low-rank Matrix Approximation

In order to reduce the number of model parameters in adaptation, low-rank approximation techniques are used. The ensemble speaker matrix can be approximated in a low-rank form as:

$$\mathbf{W}_{sd} \approx \widetilde{\mathbf{U}} * \widetilde{\mathbf{S}} * [(\widetilde{\mathbf{V}}_{sd}^{1})^{T}, (\widetilde{\mathbf{V}}_{sd}^{2})^{T}, \ldots, (\widetilde{\mathbf{V}}_{sd}^{K})^{T}] \tag{A1.3}$$

where $\widetilde{\mathbf{S}} \in R^{d*d}$ is a diagonal matrix with top d largest singular values of **S**, and $\widetilde{\mathbf{U}} \in R^{m*d}$ is a matrix with column vectors corresponding to singular values in $\tilde{\mathbf{s}}$.

$(\widetilde{\mathbf{V}}_{sd}^{i})^{T} \in R^{d*n}$ is the speaker coefficient matrix and $d << \min\{m,n\}$ is the low-rank value of the matrix. The advantage of using this low-rank approximation is that a small bottleneck layer can be generated in implementation and they may make the model much more robust (or with better generalization ability) than using the full-rank matrix [81].

A1.1.4 Adaptation on SAT-DNN Ensemble Models

1) Updating speaker coefficient matrix of the SD layer

As shown in Section A1.1.2, one direct physical explanation of the ensemble matrix factorization (refer to Equations A1.1 and A1.2) is that: **U*S** is the weighted speaker subspace bases and $(\mathbf{V}_{sd}^{i})^{T}$ is the speaker coefficient matrix. For a test speaker, the adapted model can be regarded as one point in this speaker subspace, and then the weight matrix for the SD layer of the target speaker should be in the form of

$$\mathbf{W}_{sd}^{test} = \mathbf{U} * \mathbf{S} * (\mathbf{V}_{sd}^{test})^{T} \tag{A1.4}$$

This $(\mathbf{V}_{sd}^{test})^{T}$ needs to be estimated in the adaptation model. This matrix is a function of training speakers as:

$$\mathbf{V}_{sd}^{test} \overset{\Delta}{=} \mathbf{F}(\mathbf{V}_{sd}^{1}, \mathbf{V}_{sd}^{2}, \ldots, \mathbf{V}_{sd}^{K}; \Theta) \tag{A1.5}$$

where **F(.)** is a function matrix with parameter Θ. It is difficult to obtain the solution if there is no prior knowledge of this F**(.)**. Suppose this mapping function is a linear regression of all training speakers, it can be formulated as (for simplicity, the bias in linear regression model is omitted):

$$\mathbf{V}_{sd}^{test} \overset{\Delta}{=} \sum_{i=1}^{K} \mathbf{A}_i \mathbf{V}_{sd}^{i} \tag{A1.6}$$

where $\mathbf{A}_i$ is a regression matrix. If $\mathbf{A}_i$ is an identity matrix

($\mathbf{A}_i = \mathbf{I}$) for i=1, 2 … K, the adaptation model is the average of all training speakers as

$$\mathbf{V}_{sd}^{test} \overset{\Delta}{=} \overline{\mathbf{V}}_{sd}^{train} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{V}_{sd}^{i} \tag{A1.7}$$

If $\mathbf{A}_i$=**0** for all i except when $i \neq best$, then

$$\mathbf{V}_{sd}^{test} \overset{\Delta}{=} \mathbf{A}_{best} \mathbf{V}_{sd}^{best} \tag{A1.8}$$

This means only picking up the "best" speaker's model $\mathbf{V}_{sd}^{best}$ for adaptation. In implementation, rather than using the linear regression in Equation A1.8, a direct parameter update algorithm for non-linear regression was applied for more accurate estimation. The matrix in DNN is decomposed into two components as shown in Figure 8.3. Only matrix parameters in $\mathbf{V}_{sd}^{test}$ are updated from an initial model of $\mathbf{V}_{sd}^{best}$ using adaptation data.

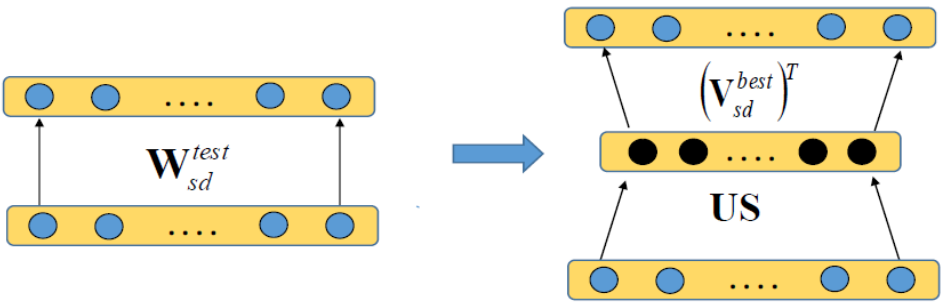


Figure 8.3: Decomposition to $\mathbf{W}_{sd}^{best}$ for Speaker Coef-matrix Adaptation.

In order to reduce the number of adaptation parameters, low-rank form as introduced in Section A1.1.3 can be used. All of the equations and formulations in Equations A1.4, A1.5, A1.6, A1.7 and A1.8 hold by changing corresponding matrix to its low-rank form.

2) Updating Singular Values in the SD Layer

After picking up a "best" speaker's model for adaptation, that the left and right singular vectors can be supposed fixed, only the singular values are adjusted to weight these two singular vectors for a testing speaker. This idea can be formulated as follows:

For an initial model (the "best" one from SAT-DNN ensembles), the factorization of the SD matrix is:

$$\mathbf{W}_{sd}^{best} = \mathbf{U}_{sd}^{best}\Sigma_{\alpha}\left(\mathbf{V}_{sd}^{best}\right)^{T} \tag{A1.9}$$

where $\Sigma_{\alpha} = \mathrm{diag}\left(\alpha_1, \alpha_2, ..., \alpha_i, ..., \alpha_p\right)$, p=min$\{m,n\}$.

For a test speaker, the $\mathbf{U}^{best}$ and $\left(\mathbf{V}_{sd}^{best}\right)^{T}$ can be supposed to be keeping the same and only the singular value matrix is updated as:

$$\mathbf{W}_{sd}^{test} = \mathbf{U}_{sd}^{best}\Sigma_{\beta}\left(\mathbf{V}_{sd}^{best}\right)^{T} \tag{A1.10}$$

where $\Sigma_{\beta} = \mathrm{diag}\left(\beta_1, \beta_2, ..., \beta_i, ..., \beta_p\right)$.Then purpose of adaptation is to find a mapping function as:

$$\beta_i = g_i\left(\alpha_i\right), i = 1,2,..., p \tag{A1.11}$$

In real implementation, it is accomplished by inserting a linear transformation matrix **M** between $\mathbf{U}_{sd}^{best}$ and Σ_{α} according to Equation A1.10. Figure 8.4 shows the decomposition structure in DNN implementation. The transformation matrix **M** can be initialized by using identity matrix.

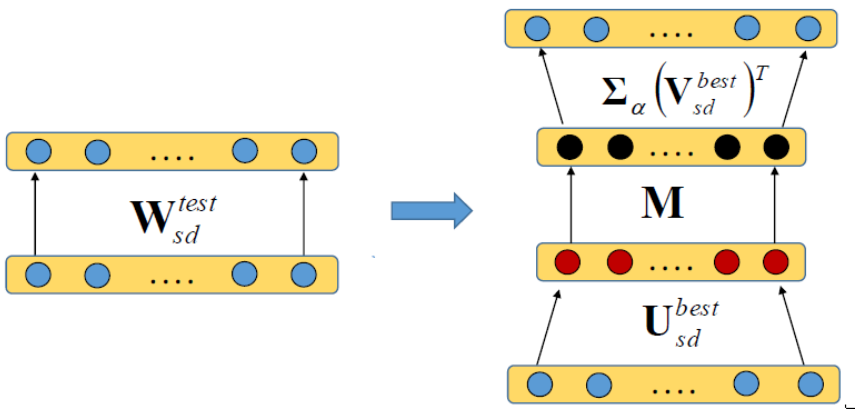


Figure 8.4: Decomposition to $\mathbf{W}_{sd}^{best}$ for Singular Values Adaptation.

This thesis work only updates the diagonal elements of **M**. The advantage of singular value adaptation strategy is that only a small number of p parameters are involved in adaptation, i.e., the number of adaptation parameters is drastically reduced.

A1.2 Implementation and Evaluations

a) SI-DNN Model

The SI-DNN model is the lightly-supervised trained DNN model. The DNN has 1320 neuron nodes in the input layer (5 frames on each side of the current frame), 3000 neuron nodes in the output layer, and 1024 neuron nodes in each hidden layer (7 hidden layers). The testing set is CCLR-TST.

b) Baseline SAT-DNN Model

The corpus and SI-DNN model used here are listed in Chapter 3. The data of total 97.2 hours from CCLR-SV and CCLR-LSV is used to do SAT-DNN training. For DNN model training, 40 dimensional filterbank features, plus their first and second derivatives were used as a feature set. 184 speakers in the training set were used in SAT training based on the SI-DNN model. Finally, 184 SD-DNN models were obtained. In this study, the SAT training is only performed on the second hidden layer. Kaldi DNN toolkit (nnet1) [46] and theano library [126] were used in the implementation. And the training procedures used in [80] were followed in the implementation. In this SAT-DNN method, choosing the second or third hidden layer as the SD layer in adaptation could obtain a better result than choosing other layers as shown in [80], there is no clear theoretical support on which layer should be used as the SD layer in SAT training. In this study, SAT training is only performed on the second hidden layer.

Recognition experiments is conducted on the TST set to see whether the adaptation is effective or not. Julius is modified for fast decoding with the DNN acoustic model. This baseline system achieved an average Character Error Rate (CER) of 28.5% with the DNN-HMM model on the TST set, not so good compared to SI-DNN.

c) Ensemble Speaker Modelling

By concatenating the weight matrices (1024×1024) of these SD layers, a super matrix (1024×188416) was organized. SVD was applied on this super matrix for factorization. Based on the factorization, globally shared speaker subspace **U** ($1024\times$rank), singular value matrix S (rank$\times$rank), and the coefficient matrix related to

each speaker $\left(\mathbf{V}_{sd}^{i}\right)^{T}$ (rank$\times$1024) were obtained. In experiments, four rank values (1024, 500, 300 and 100) were tested and full rank value is 1024.

When selecting the initial model for each testing speaker, the SD layer with highest frame accuracy was chosen on the testing data compared to the labels derived in an unsupervised way.

d) Experimental Evaluations

By gradually reducing the adaptation data for each testing speaker from 50 utterances (1 minute on average), to 30 utterances (half a minute on average), and then to 10 utterances (10 seconds on average), experiments was carried out to test the two adaptation algorithms as introduced in Section A1.1.4, i.e., speaker coefficient matrix adaptation, and singular value matrix adaptation. Table 2 shows the results for different experimental conditions. In this table, SAT means baseline SAT-DNN model. SAT-SVD-V denotes adaptation on speaker coefficient matrix **V** (with rank of the matrix specified in bracket), and SAT-SVD-S represents adaptation on the singular values.

The utterances are randomly selected from those sentences with the averaged word confidence score larger than 0.8. The improvements compared to the baseline with statistical significance (by the NIST Scoring Toolkit) are shown in bold fonts.

Table 8.1 Adaptation Performances (CER% on TST).

	Parameter size for adaptation	w/o adaptation	#utterances for adaptation		
			50	30	10
SAT (baseline)	1024*1024	28.5	26.8	27.1	27.7
SAT-SVD-V (r=1024)	1024*1024	28.5	26.8	**26.9**	**27.5**
SAT-SVD-V (r=500)	1024*500	28.5	**26.6**	**26.9**	**27.5**
SAT-SVD-V (r=300)	1024*300	**28.4**	**26.5**	**27.0**	27.6
SAT-SVD-V (r=100)	1024*100	29.6	26.8	27.5	28.4
SAT-SVD-S	1024	28.5	27.1	27.4	27.9

From Table 8.1, it is observed the rank and the adaptation data size exerted large influence to the adaptation results.

For the first method (SAT-SVD-V), most of its performances are higher than or equivalent with the baseline SAT adaptation method, except when the rank is too small

(100). SAT-SVD-V (rank=1024) outperforms SAT baseline on small data cases (30 utterances and 10 utterances), although they have the same number of parameters for adaptation. This result shows selecting the “best” initial model for adaptation is effective.

It is also noticed that SAT-SVD-V is better than the baseline SAT adaptation at rank=300 and rank=500 after adaptation with all data cases. Especially for the rank=300, the bottleneck structure seems to introduce more robustness even without adaptation. The low-rank approximation based adaptation technique shows better accuracy with large reduction on number of adaptation parameters.

The second method (SAT-SVD-S) is more sensitive to the adaptation data size due to its very limited number of parameters for adaptation (1024). But it still mostly outperforms the speaker coefficient matrix adaptation with low-rank case of SAT-SVD-V (r=100) which holds 100*1024 model parameters.

APPENDIX 2: TRAINING CRF WITH PARTIALLY ANNOTATION

A2.1 Training a CRF with Partially Annotated Data

The performance of CRF is significantly affected by the size of annotated data in the conventional supervised training paradigm. It is very costly to prepare accurate label for each word. To address this issue, a scheme of semi-supervised training, which does not require full annotation transcripts but exploits available data from partially annotated data, has been explored for CRF-based part-of-speech (POS) tagging, word segmentation [128][129], named entity recognition tasks [130]. In these works, conditional probabilities over partially annotated data are formulated. Training is achieved by modification to the learning objective function, incorporating partial annotation likelihood, so that a single model can be trained consistently with a mixture of full and partial annotation [131].

In this thesis work, the scheme of adopting partial annotated data is investigated to improve the performance of the CRF-based confidence estimation for ASR results. There is a large amount of audio and video data with closed caption texts. In the field of acoustic model training, data with closed captions can be used effectively for lightly-supervised training the acoustic models [32]. Inspired by this research, a simple but robust method is proposed for constructing partial annotation from a large-scale data set that only has closed caption texts. Then, CRF model is trained by using these data.

A2.1.1 Full and Partial Annotations

Figure 8.5 and Figure 8.6 show examples of full and partial annotations, respectively. In the figures, “T” and “F” stand for the “true” and “false” of the recognized characters. The label sequence is demonstrated as a path consists of nodes and arrows.

By choosing one label for each hypothetic character, full label sequence can be obtained {(true) → (true) → (false) → (true) → (true) → (true) → (false)…} as shown in Figure 8.5.

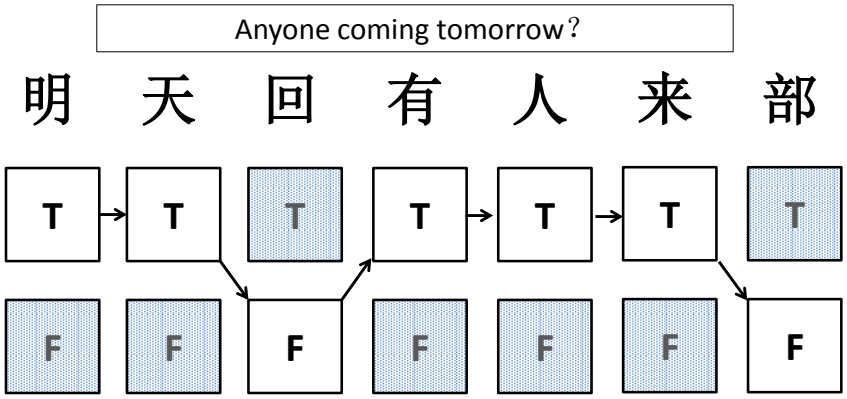


Figure 8.5 Example of Fully Annotated Utterance.

(Label sequence is in white color)

In the case of partial annotation, instead of assigning each hypothetic character a symbolic label, a non-empty subset of the label space {true, false} is assigned to each hypothetic character. In Figure 8.6, the label sequence is as follows: {(true) → (true) → (false) → (true, false) → (true, false) → (true) → (true)…}.

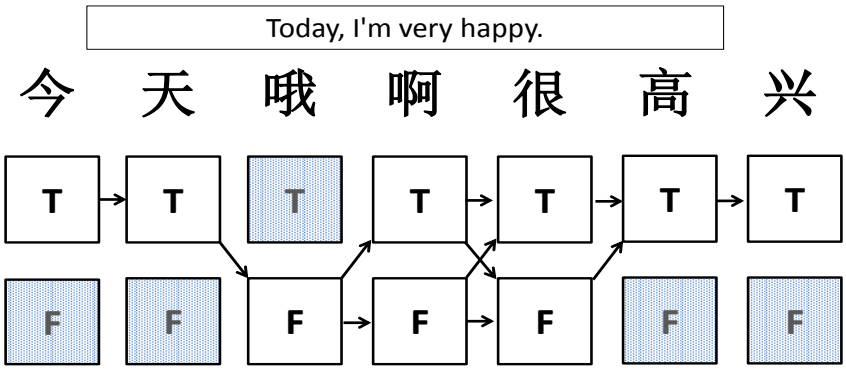


Figure 8.6 Example of Partially Annotated Utterance.

(Label sequence is in white color)

A2.1.2. CRF-based Confidence Estimation with Full Annotation

A CRF is a discriminative model which estimates the conditional probability. Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ be a label sequence given the input feature sequence $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, where N is the sequence length and $y_i \in$ {true, false}. This conditional probability is written as the normalized log-linear function as Equation A2.1.

$$p_\theta(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \exp\left(\sum_k \lambda_k f_k(\mathbf{x}, \mathbf{y}) \right) \tag{A2.1}$$

$$Z_\theta(\mathbf{x}) = \sum_{\mathbf{y}} \exp\left(\sum_k \lambda_k f_k(\mathbf{x}, \mathbf{y}) \right) \tag{A2.2}$$

where $\theta = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ are model parameters, f_k is the k-th feature function, and $Z_\theta(\mathbf{x})$ is the probability normalizer.

For fully-annotated training data, the learning problem of CRF is to maximize the log likelihood over all the training data as:

$$\theta^* = \arg\max_\theta \mathcal{L}(\theta) \tag{A2.3}$$

$$\mathcal{L}(\theta) = \sum_{p=1}^{N} \log p_\theta\left(\mathbf{y}^{(p)} \mid \mathbf{x}^{(p)}\right) \tag{A2.4}$$

Both the likelihood and its gradient can be calculated by performing the forward-backward algorithm [132] on the sequence optimization algorithms can be used to learn the model parameters, e.g. Limited Memory-BFGS [57].

A2.1.3. Train CRF Model with Partial Annotations

The method in [131] is used and it models conditional probabilities over partially annotated data. Training is achieved by a modification to the learning objective function, incorporating partial annotation likelihood, so that a single model can be trained consistently with a mixture of full and partial annotation.

As mention in last section, the possible labels that correspond to the partial annotation as $\boldsymbol{L} = (L_1, L_2, ..., L_N)$, where each L_i is a non-empty subset of the label space {true, false} that corresponds to the set of possible labels for feature $\mathbf{x}_i$. Let $\mathbf{Y}_{\boldsymbol{L}}$ be the set of all possible label sequences where $\forall \mathbf{y} \in \mathbf{Y}_{\boldsymbol{L}}$, $y_i \in L_i$. The conditional probability of $\mathbf{Y}_{\boldsymbol{L}}$ can be modeled as

$$p_\theta(\mathbf{Y}_{\boldsymbol{L}} \mid \mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \sum_{\mathbf{y} \in \mathrm{Y}_L} \exp\left(\sum_k \lambda_k f_k(\mathbf{x}, \mathbf{y}) \right) \tag{A2.5}$$

The normalizer $Z_\theta(\mathbf{x})$ is in the same format as in Equation A2.2. If each element in $\mathbf{Y}_{\boldsymbol{L}}$ is constrained to one single label, the CRF model in Equation A2.5 will roll back to Equation A2.1. So a unified framework to train CRF models can be obtained with both fully and partially annotated data.

The log marginal probability of YL over N partially annotated training examples can be formalized as follows.

$$\theta^* = \arg\max_{\theta} \mathcal{L}(\theta) \tag{A2.6}$$

$$\mathcal{L}(\theta) = \sum_{p=1}^{N} \log p_{\theta}\left(\mathbf{Y}_{\boldsymbol{L}} \mid \mathbf{x}\right) \tag{A2.7}$$

By introducing a modification to the forward-backward algorithm [131] with the same optimization algorithms, the model parameters can be learned.

LIST OF PUBLICATIONS BY THE AUTHOR

Journals:

[1] S. Li, Y. Akita, and T. Kawahara, Semi-supervised Acoustic Model Training by Discriminative Data Selection from Multiple ASR Systems' Hypotheses. (in submission)

[2] S. Li, Y. Akita, and T. Kawahara, Automatic lecture transcription based on discriminative data selection for lightly supervised acoustic model training. IEICE Trans., Vol.E98-D, No.8, pp.1545-1552, 2015.

[3] L. Wang, H. Chen, S. Li and H. Meng, Phoneme-level articulatory animation in pronunciation training, in Journal: Speech Communication, Vol. 54, Issue 7, September 2012, pp. 845–856.

International Conferences:

[1] S. Li, Y. Akita, and T. Kawahara. Data selection from multiple ASR systems' hypotheses for unsupervised acoustic model training. In Proc. ICASSP, 2016.

[2] S. Li, Y. Akita, and T. Kawahara. Discriminative data selection for lightly supervised training of acoustic model using closed caption texts. In Proc. INTERSPEECH, pp.3526-3530, 2015.

[3] S. Li, X. Lu, Y. Akita, and T. Kawahara. Ensemble speaker modeling using speaker adaptive training deep neural network for speaker adaptation. In Proc. INTERSPEECH, pp.2892--2896, 2015.

[4] S. Li, Y. Akita and T. Kawahara, Corpus and Transcription System of Chinese Lecture Room, In Proc. ISCSLP, 2014.

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

[6] S. Li, L. Wang and E. Qi, The Phoneme-level Articulator Dynamics for Pronunciation Animation, In Proc. IALP, pp.283-286, 2011.

[7] J. Chen, L. Wang, C. Li, J. Hu and S. Li, IELS: A Computer-aided Pronunciation Training System for Undergraduate Students, In Proc. ICETC, Vol.1, pp.338-342, 2010.

Technical Reports:

[1] S. Li, M. Mimura, T. Kawahara, Automatic Transcription of Chinese Spoken Lectures, ASJ autumn, 2013.

[2] S. Li, Y. Akita, T. Kawahara, Data Selection Assisted by Caption to Improve Acoustic Modeling for Lecture Transcription, ASJ spring, 2014.

[3] S. Li, Y. Akita, T. Kawahara, Classifier-based data selection for lightly-supervised training of acoustic model for lecture transcription, IPSJ SIG-SLP-102-4, 2014.

[4] S. Li, Y. Akita, T. Kawahara, Unsupervised Training of Deep Neural Network Acoustic Models for Lecture Transcriptions, ASJ autumn, 2014.

[5] S. Li, Y. Akita, T. Kawahara, Incorporating divergences from hypotheses of multiple ASR systems to improve unsupervised acoustic model training, ASJ spring, 2015.

[6] S. Li, Y. Akita, T. Kawahara, Effective combination of multiple ASR hypotheses with CRF-based classifiers, ASJ autumn, 2015.

[7] S. Li, Y. Akita, T. Kawahara, Discriminative data selection from multiple ASR systems' hypotheses for unsupervised acoustic model training, IPSJ SIG-SLP-109-8, 2015.