OBJECTIVE-DRIVEN DISCRIMINATIVE TRAINING AND ADAPTATION BASED ON AN MCE CRITERION FOR SPEECH RECOGNITION AND DETECTION

A Dissertation
Presented to
The Academic Faculty

By

Sunghwan Shin

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2013

OBJECTIVE-DRIVEN DISCRIMINATIVE TRAINING AND ADAPTATION BASED ON AN MCE CRITERION FOR SPEECH RECOGNITION AND DETECTION

Approved by:

Dr. Biing-Hwang (Fred) Juang, Advisor
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Nikil S. Jayant
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Xiaoli Ma
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Francesco G. Fedele
Associate Professor, School of Civil and Environmental Engineering; and School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Mark A. Clements
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 22 August 2013

Dedicated to my family:

Ji-Yeon, Hye-Lynn, and Jung-Min.

ACKNOWLEDGEMENTS

First and foremost, I would like to greatly appreciate my advisor, Prof. Biing-Hwang (Fred) Juang, for his great supports and guides during my Ph.D. study. I am very fortunate to have him as my advisor. He has always encouraged me to look closely at the research problems and has patiently waited for me to produce the results. This work would not have been possible without his encouragement and support during the last several years. It has been truly honor to be his student.

I would also like to specically thank Dr. Mark Clements and Dr. Xiaoli Ma for serving on the reading committee of this dissertation. Their valuable feedback on my research has provided me an opportunity to improve the quality of this work. I would also like to thank Dr. Nikil Jayant and Dr. Francesco Fedele for serving on my thesis committee and for providing me with insightful comments.

I owe great acknowledge to Dr. Ho-Young Jung, Hyung-Bae Jeon, and Dr. Yun-Keun Lee in Electronics and Telecommunications Research Institute (ETRI), and Dr. Shinji Watanabe in Mitsubishi Electric Research Labs (MERL). The collaboration and communication with them broadened my horizon in the research area of acoustic modeling.

I owe my special thanks to all my former and current collegues at the CSIP: Soohyun Bae, Ted S. Wada, Antonio Moreno-Danniel, Dwi Sianto Mansjur, Yong Zhao, Chao Weng, Wung Jason, Umair Altaf, Mingyu Chen, Mehrez Souden, Seok-Chul Sean Kwon, Yong-Jun Chang, Sungeun Lee, Byungki Byun, Ilseo Kim, Jonathan Kim, and Haejoon Jung, for their great support over my Ph.D. Study.

And finally, my greatest gratitude goes to my wife, Ji-Yeon Kim, who has given me endless love, enormous support and sacrifice during my Ph.D. studies. Without her, I would not have achieved this work. I would like to thank my parents, Geum-Za Choi, Chul-Kyu Kim, and Jin-Sang Jung. Their encouragement will always accompany me in my future career.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Acoustic modeling in state-of-the-art speech recognition systems is commonly based on discriminative criteria. Different from the paradigm of the conventional distribution estimation such as maximum *a posteriori* (MAP) and maximum likelihood (ML), the most popular discriminative criteria such as MCE and MPE aim at direct minimization of the empirical error rate. As recent ASR applications become diverse, it has been increasingly recognized that realistic applications often require a model that can be optimized for a task-specific goal or a particular scenario beyond the general purposes of the current discriminative criteria. These specific requirements cannot be directly handled by the current discriminative criteria since the objective of the criteria is to minimize the overall empirical error rate.

In this thesis, we propose novel objective-driven discriminative training and adaptation frameworks, which are generalized from the minimum classification error (MCE) criterion, for various tasks and scenarios of speech recognition and detection. The proposed frameworks are constructed to formulate new discriminative criteria which satisfy various requirements of the recent ASR applications. In this thesis, each objective required by an application or a developer is directly embedded into the learning criterion. Then, the objective-driven discriminative criterion is used to optimize an acoustic model in order to achieve the required objective.

Three task-specific requirements that the recent ASR applications often require in practice are mainly taken into account in developing the objective-driven discriminative criteria. First, an issue of individual error minimization of speech recognition is addressed and we propose a direct minimization algorithm for each error type of speech recognition. Second, a rapid adaptation scenario is embedded into formulating discriminative linear transforms under the MCE criterion. A regularized MCE criterion is proposed to efficiently improve the generalization capability of the MCE estimate in a rapid adaptation scenario. Finally,

the particular operating scenario that requires a system model optimized at a given specific operating point is discussed over the conventional receiver operating characteristic (ROC) optimization. A constrained discriminative training algorithm which can directly optimize a system model for any particular operating need is proposed. For each of the developed algorithms, we provide an analytical solution and an appropriate optimization procedure.

CHAPTER 1

INTRODUCTION

The technology of automatic speech recognition (ASR) by a machine has advanced substantially in the last two decades [1, 2, 3], thanks to the mathematical formalization of the statistical modeling approach that forms the foundation of the ASR system design methodology. Most of the research in the ASR system design has concentrated on hidden Markov models (HMMs) [1, 2]. The statistical estimation approaches in solving the estimation problem of the HMM parameters, such as the maximum *a posteriori* (MAP) estimation and the maximum likelihood (ML) estimation [4, 5], have made HMMs become mainstream in ASR.

In practice, the statistical modeling approaches based on the paradigm of *distribution estimation* such as ML and MAP often cannot lead to optimal performance of the ASR system because of several limitations. One fundamental issue is that the lack of knowledge associated with the choice of the functional form of the real-data distribution would impede the optimal distribution estimation. Furthermore, maximizing the likelihood or the posterior of the observations does not guarantee the minimum error rate in ASR since there is no direct relationship between the training criterion and the system evaluation criterion, which is normally defined by the phone or word error rate (PER/WER) in ASR.

An effective alternative to the conventional distribution estimation approaches is discriminative training (DT) [6, 7, 8], such as maximum mutual information (MMI) [9, 10, 11], minimum classification error (MCE) [12, 13, 14], and minimum phone/word error (MPE/MWE) [15, 16], of which MCE and MPE/MWE aim at direct minimization of the *empirical error rate* rather than fitting the distributions while MMI tries to maximize the mutual information that is utilized as a measure of association between data and their corresponding labels. The DT methods construct a *discriminative objective function* corresponding to the task evaluation measure and obtain the required recognition models by

minimizing or maximizing the given objective function. These methods have shown successful results in various speech recognition tasks. Thus, the training of acoustic models in state-of-the-art speech recognition systems is commonly based on the discriminative criteria.

Motivated by the success of DT, the use of the discriminative criteria has been widely investigated for model adaptation [17, 18, 19]. This is referred to as discriminative adaptation [20, 21, 22, 23]. It is well known that the performance of the ASR system severely degrades when the test speech has a different acoustic condition, which is not matched with that of the training data. To deal with the mismatch between the training and testing acoustic conditions, in discriminative adaptation, one of the discriminative criteria is chosen in adapting an acoustic model to a specific test domain. Several studies [24, 25, 26] show that the discriminative adaptation methods generally outperform the conventional adaptation methods based on the distribution estimation.

Most of recent studies on discriminative training and adaptation have focused on training hypothesis structures and optimization algorithms given an objective function, in order to further improve the overall ASR performance. In [8], the use of word lattices, instead of N-best lists, for estimating the parameters of the acoustic model under the MMI, MWE, and MCE criteria was implemented without changing the structure of the lattice. Later, weighted finite state transducers (WFSTs)-based DT methods [27, 28] were proposed to produce much more hypothesis sequences than the word lattices. In addition, several optimization techniques over the DT methods have been compared in a unified framework [8, 6].

However, as recent ASR applications become diverse, it has been increasingly recognized that the realistic applications often require a model that can be optimized for a task-specific goal or a particular scenario beyond the general purposes of the current discriminative criteria. For one example, a level of significance for each type of speech recognition errors is often scaled according to the task-specific direction. In an automatic dialog-enabled

language-learning system [29, 30], a deletion error may be regarded as more serious than a substitution error because currently there exist no evaluation guidelines for the deletion error. For this task, a direct minimization mechanism of the deletion error has to be taken into account in the learning criterion.

Another example is that one may require a model optimized at a particular operating scenario. A speaker identification system [31, 32] often claims to have a very low false alarm rate (FAR) while taking a relatively high false rejection rate (FRR) to ensure that legitimate users are not unduly denied access. For this particular scenario, it is necessary to provide a training algorithm that minimizes a FRR at a very low FAR point (e.g., minimize a FRR at a fixed 1% FAR).

Similarly, the area of model adaptation has been attracted to a particular scenario, in which the amount of adaptation data is severely limited (typically less than 10 seconds of adaptation speech). Such a practical adaptation scenario is referred to as rapid adaptation [33, 34, 35]. However, there has been little effort to develop a discriminative adaptation method for rapid adaptation. It is well known that the discriminative criteria easily cause an over-fitting problem in the parameter estimation given the severely limited adaptation data. To utilize discriminative adaptation for rapid adaptation, a new type of an objective function which can efficiently prevent the over-fitting is required.

The current discriminative criteria cannot directly handle the practical and specific requirements discussed above since the objective of the current DT methods is to minimize the *overall* empirical error rate (e.g., string, word, or phone error rate) or maximize the mutual information between data and their corresponding labels. The critical limitation of these methods lies on the rigid structure of the objective function formulation. It is necessary that the chosen objective function to be optimized can be redesigned depending upon an application specification or requirement. Then, any particular objective would be achieved through the objective-driven learning.

As a consequence, to utilize a discriminative criterion for various task-specific goals

and scenarios, a flexible and versatile design in the objective function formulation is crucial. In this dissertation, several objective-driven discriminative training and adaptation frameworks are constructed to overcome the current limitations in utilizing the discriminative criteria for various tasks and scenarios of speech recognition and detection. In particular, the proposed learning frameworks are generalized from the MCE criterion since the MCE criterion provides the flexible framework in formulating the error objective functions appropriate for various tasks and directly links the error objectives to the empirical error rate.

Generalized from the MCE criterion, the main focus of this thesis is to formulate various objective-driven discriminative criteria, which directly minimize the error objective defined by a task-specific goal or a particular scenario. In this formulation, any objective that an application or a developer requires is directly embedded into the learning criterion. Then, the required objective can be achieved by minimizing the specialized objective function based on the objective-driven discriminative criterion.

In this thesis, three task-specific requirements, briefly discussed above, are mainly taken into account in developing the objective-driven discriminative criteria. First, the issue of individual error minimization of the ASR errors is addressed and we propose a direct minimization algorithm of each error type. The three types of errors are explicitly the deletion error, the insertion error, and the substitution error. Second, the rapid adaptation scenario is embedded into formulating discriminative linear transforms under the MCE criterion. A regularized MCE criterion is proposed to efficiently improve the generalization capability of the MCE estimate in a rapid adaptation scenario. Finally, the particular operating scenario that requires a model optimized at a given specific operating point is discussed over the conventional receiver operating characteristic (ROC) optimization [36, 37, 38]. A constrained discriminative training algorithm which can directly optimize a model for any particular operating need is proposed. For each of the developed algorithms, we provide an analytical solution and an appropriate learning procedure.

This thesis is organized as follows: Chapter 2 introduces the origin of the problems and the related work. In this chapter, the conventional and discriminative approaches to training and adaption of the acoustic model are extensively discussed since the focus of this thesis is on discriminative training and adaptation. Chapter 3 presents individual error minimization learning frameworks for speech recognition and detection. Discriminative training for direct minimization of deletion, insertion, and substitution errors is first presented as a direct solution to minimizing each type of the ASR errors. Then, as a natural extension to the detection and verification problem, an adaptive utterance verification framework is described. Chapter 4 provides discriminative linear transform-based adaptation using MVE and MCE criteria. MVE linear regression (MVELR) is first presented as an effective discriminative adaptation method to the detection and verification problem. Then, the regularized MCE linear regression (RMCELR) is proposed to directly deal with a rapid adaptation scenario. Additionally, a structural framework to the prior density estimation is incorporated into the RMCE criterion and thus the structural RMCELR (SRMCELR) is formulated as more efficient discriminative adaptation method for rapid adaptation. Chapter 5 presents a new constrained discriminative training algorithm for particular operating point optimization. The MVE criterion is reformulated by the augmented Lagrange multiplier (ALM) to deal with a particular operating scenario. Finally, the summary and the contributions of the entire thesis are presented in Chapter 6.

CHAPTER 2

BACKGROUND AND RELATED WORK

In this chapter, we first give an overview of an automatic speech recognition (ASR) system. Among the building blocks in the ASR system, acoustic modeling based on hidden Marcov models (HMMs) and maximum likelihood (ML) training is mainly discussed. Then, acoustic model adaptation methods to a different acoustic environment or a new speaker are described. Finally, the conventional discriminative training (DT) criteria and discriminative linear transform-based adaptation methods are extensively discussed since the focus of this thesis is on discriminative training and adaptation.

2.1 An Automatic Speech Recognition System

Rapid progress in the technology of automatic speech recognition (ASR) has been witnessed for the last few decades [1, 2, 3] by the advance of the mathematical formalization of the statistical modeling approaches to learning acoustic and linguistic characteristics from the speech data.

The aim of the ASR systems is to transcribe speech into words. It can be seen as recognizing a word sequence W in a spoken speech waveform X. This can be formulated as a well-known maximum a *posterior* (MAP) decision rule as follows:

$$\widehat{W} = \arg\max_{W} P(W|X) = \arg\max_{W} \Big\{ \overbrace{P(X|W)}^{\text{acoustic score}} \cdot \overbrace{P(W)}^{\text{LM score}} \Big\} \tag{1}$$

where $P(W|X)$ is the posterior probability, $P(X|W)$ is the likelihood of the observation sequence X for the hypothesis W to have produced the observation sequence X, and $P(W)$ is the prior probability of the hypothesis. The likelihood $P(X|W)$ is computed by using the acoustic model (AM), which models the distribution of observations X, and the prior probability $P(W)$ is approximated by the language model (LM), which indicates the probability of the occurrence of the underlying hypothesis W. In this statistical ASR system, a decoder

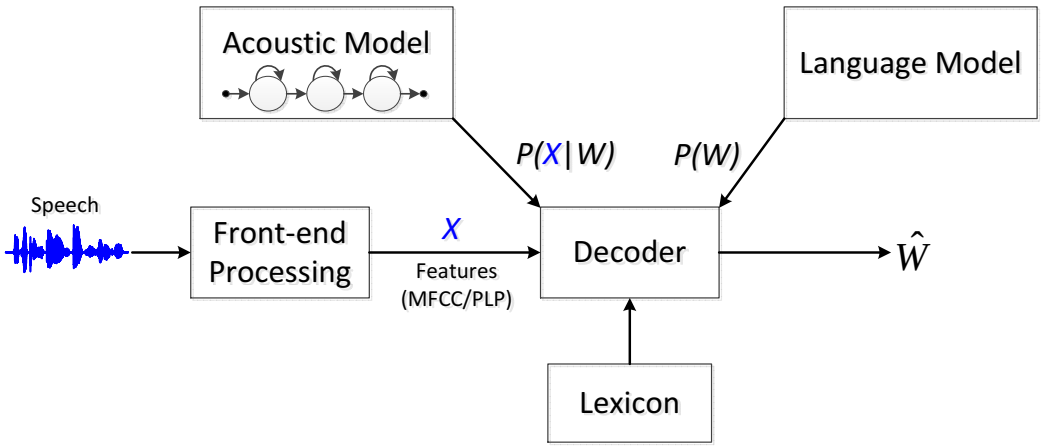


Figure 1: A Block Diagram of an Automatic Speech Recognition (ASR) System.

produces the most probable word sequence $\widehat{W}$ as the output which has the largest probability from AM and LM. Figure 1 presents a block diagram of the statistical ASR system described above.

2.1.1 Feature Extraction

In an ASR system as shown in Figure 1, the input continuous speech waveform is first converted into an appropriate form that is normally defined as a sequence of discrete parameter vectors. These parameter vectors capture the essential discriminating characteristics of the raw speech signal and are referred to as *acoustic features*. This feature extraction module is often referred to as the front-end processing of the ASR system. The feature extraction which carries out compact and effective acoustic features is a common and important process so as to model the statistical properties well and construct a good recognition system.

The most widely used acoustic features in state-of-the-art speech recognition systems are Mel-frequency cepstral coefficients (MFCC) [39] and perceptual linear prediction (PLP) [40]. Both types of feature extraction generate cepstrum-based features, which use the perceptually motivated parameterization, and have been used successfully in most ASR systems. In this thesis, our work is based on the MFCC feature vectors whose default stream is the basic parameter vector, the first (delta) and second (acceleration) difference coefficients and the log energy.

Additionally, there exist various methods as feature-domain post-processing to further enhance the acoustic features. Cepstral mean normalization (CMN) [41], spectral subtraction (SS) [42], and vocal tract length normalization (VTLN) [43] are used to make speech features robust to environmental or speaker variations.

2.1.2 Acoustic Modeling

In Figure 1, the acoustic model (AM) provides the likelihood of a hypothesis W that has led to the acoustic features extracted from the speech observations X. From the figure and Eq. (1), we can see that the acoustic score $P(X|W)$ computed from the acoustic model plays a very critical role in the ASR system to finally obtain the accurate and reliable word sequence $\widehat{W}$. Hence, acoustic modeling has become one of the most active research topics in ASR, in order to build a high performance speech recognition system.

Most of the research in the acoustic model design has concentrated on *Hidden Markov Models* (HMMs) [1, 2]. HMMs provide a flexible formulation of the acoustic model in which the short-time stationarity of a speech signal can be well characterized as a parametric random process. In particular, the statistical estimation approaches in solving the estimation problem of the HMM parameters, such as the *Maximum a posterior* (MAP) estimation and the *Maximum Likelihood* (ML) estimation [4, 5], have made HMMs become mainstream in ASR. HMMs have been widely used in various ASR tasks and as the most popular and successful acoustic model so far.

In general, given a speech observation sequence $X = \{x_{1,}x_{2,}\ldots,x_T\}$, where x_t is a feature vector within a specific time window, an N-state HMM with a state sequence $\mathbf{q} = \{q_0, q_1, \ldots, q_T\}$ is characterized by a parameter set $\lambda = \{\pi, A, B\}$, where $\pi = \{\pi_i = P(q_0 = i) : 1 \leq i \leq N\}$ is the initial state distribution, $A = \{a_{ij} : 1 \leq i \leq N, 1 \leq j \leq N\}$ is the transition probability matrix, and $B = \{b_j(x_t) : 1 \leq j \leq N\}$ is the observation probability of the states. For instance, the HMM used as a distribution of the speech utterance X is

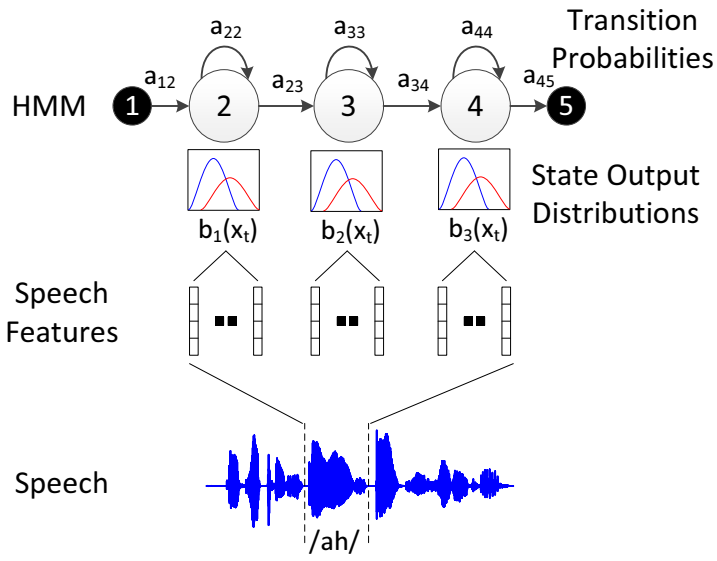


Figure 2: A three-state left-to-right Hidden Markov Model (HMM).

defined as

$$\begin{aligned} P(X|\lambda) &= P(X|\pi, A, B) = \sum_{\mathbf{q}} P(X|\mathbf{q}, \lambda) P(\mathbf{q}|\lambda) \\ &= \sum_{\mathbf{q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(x_t). \end{aligned} \tag{2}$$

Each state in the N-state HMM is associated with an output probability distribution and a state transition probability is attached for transitions from any given state to each of N possible states. Figure 2 depicts a widely used three-state left-to-right HMM. In this figure, an HMM is used to model one acoustic unit, a phoneme /ah/, as described above.

Most state-of-the-art ASR systems are based on a continuous density HMM (CDHMM) and use a multivariate *Gaussian Mixture Model* (GMM) as the output probability distribution defined by $b_j(x_t) = \sum_{m=1}^{M_j} c_{jm} \mathcal{N}(x_t; \mu_{jm}, R_{jm})$, where M is the number of mixture components in state j, c_{jm} is the weight of the mixture component m for state j, and μ_{jm} and R_{jm} are the mean vector and the covariance matrix for the m-th component of the j-th state. The weights c_{jm} are constrained to add up to one for each state. To use HMMs for ASR, there are three fundamental problems [1]: the evaluation, decoding, and estimation problems. In this thesis, the last problem, referred to as the estimation or training issue, will be mainly

discussed according to the objective of the proposed research.

Generally, the estimation problem concerns how to adjust a set of HMM parameters, λ, so as to best fit the given set of observations (called the training set). This parameter estimation can be viewed as training the model that maximizes the probability of the observations. For this goal, in the ASR literature, the ML estimation has been widely used because of its attractive attributes, such as easy implementation and excellent properties of convergence. In the ML criterion, a set of HMM parameters, λ, are estimated by maximizing the likelihood given a set of training data $\mathcal{D} = \{X_{1,}X_{2,}\ldots,X_K\}$ as follows:

$$\hat{\lambda}_{\mathrm{ML}} = \arg\max_{\lambda} P(\mathcal{D}|\lambda) = \arg\max_{\lambda} \left\{ \sum_{k=1}^{K} \log P(X_k|\lambda) \right\}, \tag{3}$$

where K is the amount of training data, and X_k is the training utterance for utterance k. As direct optimization of the ML objective function is difficult, the Baum-Welch algorithm [44], which is a practical implementation of the expectation maximization (EM) algorithm [45], is iteratively used to estimate the HMM parameters. In this approach, an auxiliary function, which provides a lower bound on the log-likelihood, is defined by the current model parameters λ_k at the k-th iteration. The new estimates of the model parameters λ_{k+1} at the $(k+1)$-th iteration are then achieved by maximizing this lower bound, which in turn, increases the log-likelihood. This procedure iterates until the log-likelihood converges to a local optimum. The ML criterion has been adopted to many ASR applications as a standard training method [1, 2].

However, this ML training may provide an optimal solution for the density estimation, but it often does not lead to the optimal performance of the ASR system, meaning the minimum recognition error rate. As a remedy, several discriminative training (DT) methods [12, 46, 15, 47] have been proposed to directly minimize the recognition error rate instead of maximizing the likelihood of the observations. Since the focus of this thesis is on discriminative training and adaptation, some widely used discriminative training and adaptation algorithms will be extensively reviewed in Sections 2.3 and 2.4.

2.1.3 Language Modeling

In Eq. (1), the prior probability $P(W)$ is approximated by the language model (LM) and the LM score as well as the AM score is important factor in recognizing a word sequence W. The language model is a statistical model which represents the syntactic and semantic information in spoken word sequences. In Figure 1, a lexicon, or called a dictionary, defines how each word is pronounced and formed by a set of HMMs in the allowed vocabulary set. On the other hand, the language model determines what sequences of words are grammatically formed and assigns a probability to the word sequence as the LM score. This knowledge information about language is especially important to large vocabulary continuous speech recognition (LVCSR) and spontaneous speech recognition systems.

The most popular language model in the state-of-the-art speech recognition systems is the N-gram language model [1, 2]. Suppose a word sequence $W = \{w_1, w_2, \ldots, w_M\}$ constitutes a sequence of words w_m. In an N-gram model, the probability $P(w_1, \ldots, w_M)$ is approximated as follows:

$$P(w_1, \ldots, w_M) = \prod_{i=1}^{m} P(w_i | w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i | w_{i-(N-1)}, \ldots, w_{i-1}). \tag{4}$$

It is assumed that the probability of the i-th word w_i in the word sequence W can be approximated by constraining the history to the preceding (N-1) words. The simplest case is not to use any history and thus every word has an equal probability as an uniform distribution. In Eq. (4), when $N = 1$, it is referred to as a *unigram* language model while a *bigram* language model is assigned for $N = 2$. In this thesis, a simple loop network, unigram and bigram language models are used for the ASR experiments.

2.2 Acoustic Model Adaptation

Although the HMMs in an ASR system are well trained by the effective optimization algorithm with a sufficient amount of training data, the performance of the ASR system severely degrades when the test speech is from a different acoustic environment or a new speaker who is not matched with the original speakers or environment during training. This serious

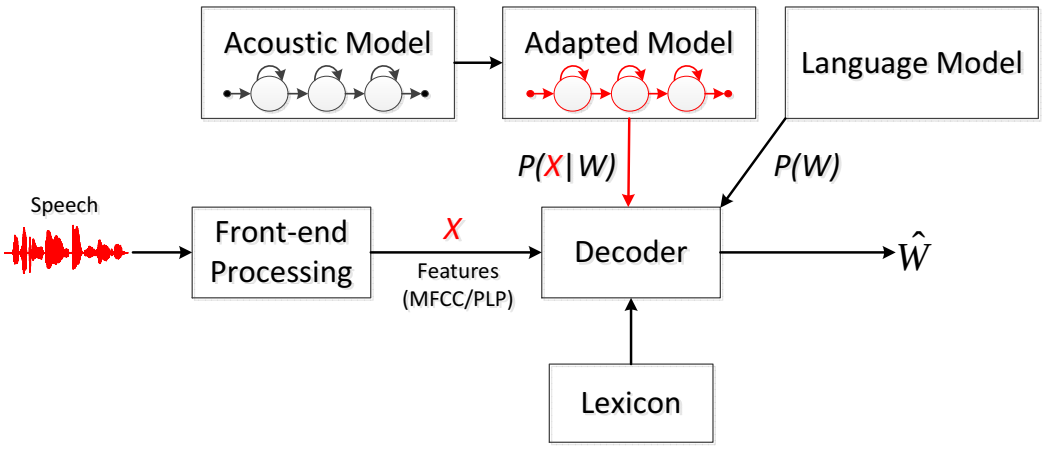


Figure 3: A Block Diagram of an Automatic Speech Recognition (ASR) System When the Adapted Model is Applied to a Noisy Speech Input.

degradation is due to the mismatch between acoustic conditions of the training and testing. In real-world applications, such mismatched scenarios are unavoidable [48, 49, 2]. One direct and effective solution to deal with this mismatch is to adapt the acoustic models to the environmental distortion or the speaker variation. This approach is usually referred as *model adaptation* [17, 18, 19, 33], which is an active research area in ASR.

Figure 3 presents a block diagram of a speech recognition system when the adapted model is applied to a noisy speech input. From the figure, we can see that the original acoustic model is adapted so as to match the testing acoustic condition prior to recognizing the test speech. This model adaptation can be performed with the allowed adaptation data which can represent the acoustic characteristics of the testing domain. In practice, the amount of adaptation data is very limited. The critical issue is thus how the initial acoustic model can be accurately and rapidly adapted to the different acoustic condition with the limited adaptation data.

In the ASR literature, there are two popular model adaptation methods, *Maximum a posteriori* (MAP) adaptation [18] and *Maximum Likelihood Linear Regression* (MLLR) adaptation [19]. A fundamental procedure of these adaptation methods is to adapt the initial HMMs to a specific test domain – either a new environment or a new speaker – using a small amount of test domain data. The MAP approach directly adapts the HMM parameters by

maximizing the *a posteriori* distribution of the HMM parameters given the adaptation data while the MLLR approach estimates an affine transformation by maximizing the likelihood of the adaptation data, so as to shift the parameters closer to those for the test condition.

2.2.1 Maximum *a posteriori* (MAP) Adaptation

In the MAP criterion, the adapted HMM parameters can be obtained by

$$\hat{\lambda}_{\text{MAP}} = \arg\max_{\lambda} P(\lambda|\mathcal{D}) = \arg\max_{\lambda} P(\mathcal{D}|\lambda)P(\lambda), \tag{5}$$

where $\mathcal{D}$ is the adaptation data, and $P(\lambda)$ is the *a priori* distribution over the HMM parameters. In the MAP estimate, the prior term, $P(\lambda)$, represents prior knowledge about the distribution of model parameters and imposes constraints on the values of the parameters. Thus, the MAP adaptation method can prevent the parameters from being over-trained when the amount of adaption data is limited. Note that if $P(\lambda)$ assumes a uniform distribution, then this MAP criterion becomes identical to the ML criterion defined in Eq. (3).

An important issue for MAP adaptation is the choice of the prior distribution. In practice, the initial HMM parameters are generally used as the informative priors [18]. For example, the final update formulation of the ML estimate, here no prior yet, with respect to the mean vector of the m-th Gaussian mixture component for the j-th state is written as

$$\hat{\mu}_{mj}^{ML} = \frac{\sum_t \gamma_{jm}(t)x(t)}{\sum_t \gamma_{jm}(t)} \tag{6}$$

where $\gamma_{jm}(t)$ is the occupation probability of m-th mixture component for the j-th state and $x(t)$ is t-th observation vector. On the other hand, if we assume the prior mean is μ_0, the MAP estimate can be then written as follows:

$$\hat{\mu}_{mj}^{MAP} = \frac{\tau\mu_0 + \sum_t \gamma_{jm}(t)x(t)}{\tau + \sum_t \gamma_{jm}(t)} \tag{7}$$

where τ is a hyper-parameter that controls the balance between the prior mean and the ML estimate of the mean. From Eqs. (6) and (7), we can see that the update formulation of the MAP adaptation is a weighted sum of the prior mean with the ML estimate of the mean

vector. Therefore, as the amount of adaptation data increases, the MAP solution approaches the ML solution. On the contrary, if the amount of adaptation data is very small, then the MAP estimate will remain close to the initial HMM parameters.

2.2.2 Maximum Likelihood Linear Regression (MLLR) Adaptation

On the other hand, a linear transform-based adaptation method has been a widely used alternative to MAP adaptation when there are limited adaptation data. The aim of the linear transform is to adapt the mean parameters of a Gaussian-mixture HMM system using an affine transformation:

$$\hat{\mu} = \mathrm{W}\xi = A\mu + b, \tag{8}$$

where $\xi = [\mu \ 1]^T$ is the extended mean vector, and $\mathrm{W} = [A \ b]$ is the linear transform matrix, which includes both a linear transformation matrix A and a bias vector b. By using the above transform strategy, the mean parameters μ of the initial ASR system are adapted, depending on the availability of the adaptation data. The MLLR method estimates the linear transform by maximizing the likelihood associated with the adaptation data as follows:

$$\hat{\mathrm{W}}_{MLLR} = \arg\max_{\mathrm{W}} P(\mathcal{D}|\lambda, \mathrm{W}). \tag{9}$$

In order to solve the maximization problem, the expectation-maximization (EM) algorithm [44, 45] is applied into the auxiliary function as defined by

$$Q(\lambda, \hat{\lambda}) = K - \frac{1}{2}\sum_{r=1}^{R}\sum_{i=1}^{d}\left(\mathrm{w}_{ri}G_r^{(i)}\mathrm{w}_{ri}^T - 2\mathrm{w}_{ri}k_r^{(i)T}\right) \tag{10}$$

where w_{ri} is the i-th row of W_r along with:

$$G_r^{(i)} = \sum_{m\in M_r}\sum_t \gamma_m(t)\xi_m\xi_m^T\frac{1}{\sigma_{mi}^2} \tag{11}$$

$$k_r^{(i)} = \sum_{m\in M_r}\sum_t \gamma_m(t)x_i(t)\xi_m^T\frac{1}{\sigma_{mi}^2} \tag{12}$$

where $\gamma_m(t)$ and $x(t)$ is the occupation probability and the t-th observation vector, and σ_{mi}^2 is the i-th element of the covariance matrix of the m-th Gaussian, respectively. Note that

HMM label m corresponds to regression class r with membership set M_r. By differentiating the auxiliary function with respect to W_r, MLLR calculates the i-th row of W_r as follows:

$$\begin{aligned} w_{ri}^{ML} &= k_r^{(i)} \left(G_r^{(i)}\right)^{-1} \\ &= \left(\sum_{m \in M_r} \sum_t \gamma_m(t) x_i(t) \xi_m^T \frac{1}{\sigma_{mi}^2}\right) \left(\sum_{m \in M_r} \sum_t \gamma_m(t) \xi_m \xi_m^T \frac{1}{\sigma_{mi}^2}\right)^{-1}. \end{aligned} \tag{13}$$

This method is particularly effective for a small amount of adaptation data because a single transform W can be shared across a set of Gaussian components. In MLLR, all of the Gaussian components are dynamically clustered into several regression classes as specified by a regression-class tree [19], depending upon the amount of adaptation data available.

For a small amount of adaptation data, MLLR usually outperforms MAP since MLLR makes use of the pooled Gaussian transformation approach. A major drawback of MAP adaptation is that MAP can only adapt the models that are observed in the allowed adaptation data. State-of-the-art ASR systems normally have many thousands of Gaussians, and thus MAP adaptation will require a substantial amount of adaptation data to update all parameters.

In this thesis, adaptation tasks, with the small amount of adaptation data or the extremely limited adaptation data, have been mainly taken into account. Therefore, a linear transform-based adaptation approach is the main focus of this research.

2.3 Conventional Discriminative Training

As discussed in Section 2.1.2, ML training estimates HMM parameters by maximizing the likelihood $P(\mathcal{D}|\lambda)$ of observations given the labeled training data. This training criterion usually cannot lead to optimal recognition performance because it has several limitations. First, the chosen distribution form does not really match the real-data distribution. Since no one can really ascertain the distribution of speech parameters, the chosen distribution is almost surely of a wrong form. Second, the training data are normally limited, and hence the optimality properties of the ML criterion may not mean much of substance. Last,

maximizing the likelihood does not guarantee the minimum error rate in an ASR system since there is no direct relationship between the training criterion and the system evaluation criterion, which is normally defined by the phone/word error rate (PER/WER) [12].

To overcome the fundamental limitations of the traditional distribution estimation approach, *Discriminative Training* (DT) criteria have been proposed as an alternative to the ML criterion. Instead of fitting the distributions to the data, DT attempts to construct an *objective function* corresponding to the system performance measure and obtain the required models by maximizing or minimizing the given objective function. This discriminative objective function-based approach makes it easier to directly embed the discriminative criterion related to the task evaluation measure into the model optimization. Furthermore, while the ML training only considers the labeled training data on a correct transcription as reference hypotheses, DT utilizes the recognition results provided from a recognizer, as competing hypotheses, over the reference hypotheses.

DT of HMMs has been found to outperform ML training [6, 7, 8, 47] and has been widely used in state-of-the-art ASR systems. In the ASR literature, there are three popular discriminative training methods: the MMI [10, 11], MCE [12], and MPE/MWE [16] methods. Among them, MCE and MPE/MWE focus on direct minimization of the empirical error rate while MMI aims at maximizing the mutual information between data and their corresponding labels/symbols. In this section, we review each of the three popular DT methods.

2.3.1 Maximum Mutual Information (MMI)

The MMI method [10] was derived from information theory rather than decision theory. MMI training optimizes the *a posteriori* probability of training utterances by maximizing mutual information between the observations and the corresponding class labels. The MMI criterion can be defined by the sum over the logarithms of the posterior probabilities of each

observation as follows:

$$
\begin{aligned}
F_{MMI}(\lambda) &= \frac{1}{K}\sum_{k=1}^{K}\log P(W^{(k)}|X_k,\lambda) \\
&= \frac{1}{K}\sum_{k=1}^{K}\log \frac{P(X_k|W^{(k)})P(W^{(k)})}{\sum_{\overline{W}} P(X_k|\overline{W})P(\overline{W})}
\end{aligned} \tag{14}
$$

where $W^{(k)}$ is the correct transcription and $\overline{W}$ is a set of all possible word sequences for utterance X_k. MMI training maximizes the above objective function. From Eq. (14), we can see that the MMI criterion is equivalent to maximizing the ratio of the likelihood of the correct hypotheses (numerator) to that of the possible hypotheses (denominator).

In [9], the MMI method was initially applied for an isolated-word-recognition task. It was then successfully applied to connected digit recognition [10], continuous phone recognition [11], and large-vocabulary continuous speech recognition (LVCSR) [46], with the aid of an efficient optimization algorithm, the Extended Baum-Welch (EBW) algorithm [50].

2.3.2 Minimum Classification Error (MCE)

Although the MMI method demonstrated significant performance advantages over conventional ML training, MMI is not based on direct minimization of the empirical training error rate. Since the underlying objective function in MMI is the mutual information which is utilized as a measure of association between data and their corresponding labels, there is no direct relationship between the optimization criterion and the system performance measure defined by the recognition error rate in ASR.

In [13], the MCE method was first proposed by formulating an objective function that allows direct minimization of the empirical training error rate. The MCE objective function is constructed by a smooth loss function, which is a differentiable function of *class misclassification measure* defined as a close approximation to the actual classification error between the labeled model and other competing models. The MCE objective function can

be written as follows:

$$\begin{aligned} F_{MCE}(\lambda) &= \frac{1}{K}\sum_{k=1}^{K}\ell\left(d(X_k|W^{(k)})\right) \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{1+\exp\left(-\alpha\left(-g(X_k,W^{(k)}|\lambda)+G(X_k,\overline{W}|\lambda)\right)+\beta\right)} \end{aligned} \tag{15}$$

where $\ell(\cdot)$ is a smoothed loss function normally defined by a sigmoid function, $g(X_k, W^{(k)}|\lambda)$ is a discriminant function for the correct transcription $W^{(k)}$, and $G(X_k, \overline{W}|\lambda)$ is an anti-discriminant function, which is a weighted sum over all competing hypotheses defined as follows:

$$G(X_k,\overline{W}|\lambda) = \frac{1}{\eta}\log\left[\frac{1}{N}\sum_{n=1}^{N}\exp\left[g(X_k,\overline{W}_{(n)}|\lambda)\eta\right]\right] \tag{16}$$

where $\overline{W}_{(n)}$ is the n-th best string in the given N-best list. MCE training minimizes the smoothed loss function as shown in Eq. (15), which approximates the number of misclassification utterances.

Hence, MCE training can achieve the minimum training error rate by minimizing the misclassification measure given the training data. To optimize the MCE criterion, the generalized probabilistic descent (GPD) algorithm [13, 12] is generally used. The MCE criterion was originally proposed for isolated word recognition [13] and was extended to continuous speech recognition by making use of N-best lists [51, 12] or lattices [52]. It was also applied to LVCSR tasks [8].

2.3.3 Minimum Phone/Word Error (MPE/MWE)

The MPE/MWE method, which is directly related to the empirical training error rate similar to the MCE method, was proposed in [15, 16]. The objective function of this method is a weighted string posterior probability as follows:

$$\begin{aligned} F_{MWE}(\lambda) &= \frac{1}{K}\sum_{k=1}^{K}P(W^{(k)}|X_k,\lambda)A(\overline{W},W^{(k)}) \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{P(X_k|W^{(k)})P(W^{(k)})A(\overline{W},W^{(k)})}{\sum_{\overline{W}}P(X_k|\overline{W})P(\overline{W})} \end{aligned} \tag{17}$$

where $A(\overline{W}, W^{(k)})$ is a phone or word accuracy function. The weighting function, $A(\overline{W}, W^{(k)})$, can be defined at either the phone (MPE) or word (MWE) level and referred to as a "raw accuracy" function that is measured by the accuracy of the competing hypothesis $\overline{W}$ given the reference transcription $W^{(k)}$. Therefore, the MPE/MWE criterion can be viewed as the weighted sum over the posterior probability of each sentence.

As a result, MPE/MWE is also intended to minimize classification error similar to the MCE method, but is weighted by the accuracy function. Similar to the MMI method, the EBW algorithm is used to optimize the entire MPE/MWE estimation process. The MPE criterion has been shown to yield better performance than the MMI criterion [15, 16]. However, it is still not clear whether or not the MPE/MWE method is better than the MCE method in LVCSR tasks since these methods have competed with each other in several different experiments [8, 53, 54].

In this thesis, the MCE criterion is used as the specific discriminative criterion and is generalized to formulate the objectives of this research. Among all those DT methods investigated above, the MCE method provides the most flexible framework in formulating the error objective functions appropriate for various tasks and scenarios. The MCE criterion also directly links the error objectives to the empirical error rate while following the minimum error principle for acoustic modeling.

2.4 Discriminative Linear Transform-based Adaptation

Inasmuch as the DT methods had shown several promising results in state-of-the-art ASR systems, there has been increased interest in *discriminative adaptation* [20, 21, 22, 23], in which discriminative criteria are employed to adapt HMM parameters, instead of the ML criterion. The limitations of the ML criterion discussed in Section 2.1.2 still remain in the adaptation of HMMs. Furthermore, in most adaptation scenarios, the amount of adaptation data is limited; thus, it is very difficult to achieve reliable and robust estimates in such scenarios. As discussed in Section 2.2, when a small amount of adaptation data is available,

MLLR outperforms MAP since MLLR makes use of the pooled Gaussian transformation approach. Therefore, the use of discriminative criteria, such as MCE and MPE/MWE, has been widely investigated in estimating adaptation transforms. This adaptation strategy is referred to as discriminative linear transform (DLT) based adaptation.

Discriminative linear transform-based adaptation mainly uses one of the discriminative criteria, such as MMI, MCE or MPE/MWE, to estimate linear transforms given the adaptation data. Discriminative linear transforms adapt either Gaussian means, variances, or both, in a regression-class tree structure, similar to MLLR adaptation. As a result, discriminative linear transform-based adaptation methods take advantage of MLLR by sharing the same tree structure and overcome the limitations of MLLR by adopting the discriminative criteria. The three popular methods are referred to as MMI linear regression (MMILR) [25], MCE linear regression (MCELR) [55, 24], and MPE linear regression (MPELR) [26]. In addition, some new discriminative linear transform-based adaptation methods, minimum Bayes risk linear regression (MBRLR) [56] and soft margin estimation linear regression (SMELR) [57], have been recently proposed. All of these methods have been primarily applied for speaker adaptation and have been found to outperform the MLLR method.

However, there has been little effort in the ASR literature to apply discriminative linear transform-based adaptation for various ASR tasks and practical scenarios. As mentioned, the use of discriminative linear transforms has been mainly investigated for speaker adaptation. Furthermore, the use of discriminative linear transforms in a practical situation, where the amount of adaptation data is extremely limited (less than 10 seconds of adaptation speech), has not yet been addressed in detail. It is well known that linear transforms suffer from the data-sparseness problem, and it is very hard to increase generalization capability in such a practical scenario [17, 58], called rapid adaptation. In this thesis, these two issues will be discussed in detail, and new discriminative linear transform-based adaptation methods will be proposed to overcome the current limitations.

2.5 Chapter Summary

This chapter describes the origin of the problems and the related works. We first reviewed the conventional automatic speech recognition system. We introduced each building block in the ASR system: feature extraction, acoustic modeling, and language modeling. Among them, acoustic modeling based on the hidden Marcov models and maximum likelihood criterion was mainly described. Acoustic model adaptation based on the ML and MAP criteria was also revisited. Finally, the conventional discriminative training criteria and discriminative linear transform-based adaptation methods were extensively discussed.

CHAPTER 3

INDIVIDUAL ERROR MINIMIZATION LEARNING FOR SPEECH RECOGNITION AND DETECTION

In this chapter, a new discriminative training paradigm for direct minimization of three types of ASR errors, namely, the insertion error, the deletion error and the substitution error, is first proposed. We follow the minimum error principle for acoustic modeling and formulate error objectives in insertion, deletion, and substitution separately for minimization during training. This new training paradigm is generalized from the minimum verification error (MVE) criterion and can explain the direct relationship between recognition errors and detection errors. In the end, by minimizing each objective function, we can obtain three individual error minimization learning algorithms: MD(eletion)E, MI(nsertion)E, and MS(ubstitution)E, respectively. In addition, as a natural extension to the detection and verification problem, an utterance verification (UV) task is chosen to evaluate the proposed individual error minimization algorithm, especially MSE for the UV task. An integrated solution to enhance the overall UV performance, which is defined by a keyword recognition rate and an out-of-vocabulary (OOV) word rejection rate, is proposed by utilizing the MSE-trained models in both recognition and verification stages.

3.1 Direct Minimization of Deletion, Insertion, and Substitution Errors

In continuous speech recognition, recognition errors can be classified into three types after alignment between the transcription and the recognized string by a dynamic programming (DP) procedure. The three error types are deletion, insertion, and substitution. In various ASR applications, a level of significance for each of the errors is often scaled according to the task-specific direction and performance target. For example, a deletion error by the ASR system may be regarded as more serious than a substitution error in an automatic

dialog-enabled language-learning system because currently there are no evaluation guidelines for deletion errors, and the system does not know how to respond to such errors. Thus, it is desirable to formulate a training algorithm that can directly minimize each of these three types of errors.

As discussed in Section 2.3, several discriminative training (DT) methods, such as MMI, MCE, and MPE/MWE, have achieved success in various speech-recognition tasks over the years. Among them, MCE and MPE/MWE focus on direct minimization of mainly the substitution error on the chosen unit class, say a word, either on the same level as the unit, or at a level above (e.g., a string of words) or below (e.g., a string of phonemes) word. It is considered very difficult to present a natural solution to directly minimize deletion and insertion errors.

However, if we re-interpret the three types of the recognition errors in the context of a detection problem, deletion, insertion, and substitution errors can be respectively explained as miss, false alarm, and miss/false-alarm errors happening together. Then, each of the errors can be minimized under the framework of detection theory. The difference between the two problem descriptions is detailed in Table 1 (adopted from [59]). First, in terms of error type, the recognition problem is associated with only one misclassification error while the detection problem is associated with both the Type I error (miss) and Type II error (false alarm). Second, in the presence of alignment errors, the recognition output will inevitably contain deletion, insertion and substitution errors, and each of the errors in the recognition problem can be viewed as a miss, false alarm, and both in the detection problem. Last, in the traditional training criterion, normally during recognition, only the substitution error is minimized, whereas the training criterion for the detection problem can be formulated to minimize a combination or the total of the detection errors associated with miss and false-alarm, respectively. As a result, we may rethink the recognition problem as a detection problem.

In this section, based on the above analysis, a multi-objective DT method using the

Table 1: Comparison between different problem descriptions.

	Error Type	Alignment Errors	Training Criterion
Recognition Problem	Misclassification Error	Deletion Insertion Substitution	Minimize sub-errors only
Detection Problem	Type I/II (Miss/FA) Errors	Type I Type II Type I&II	Minimize Type I&II both

minimum verification error criterion (MVE) [60, 61, 59] is proposed not only to directly deal with each type of the recognition errors from a detection viewpoint, but also to minimize each of the errors and the composite recognition error rate. Under the MVE criterion, a multi-objective training framework is developed by applying two mis-verification measures for miss and false alarm errors selectively, along with the types of the recognition error definitions. In contrast to a string-level MCE [12], the proposed training framework is performed only on error segments between the transcription and the recognized string after DP matching. This training framework provides a direct measure of each type of the three errors and significantly reduces the computational complexity, compared to the string-level MCE. Hence, each objective criterion is named for the minimum deletion error (MDE), minimum insertion error (MIE), and minimum substitution error (MSE), respectively.

3.1.1 Recognition Errors from a Detection Viewpoint

The conventional, well-established MCE objective function was designed to mainly reduce the empirical substitution errors on the training data. For every training utterance X_k, a string-level misclassification measure, $d(X_k|\lambda)$, [12, 14] compares two discriminant functions, $g(X_k, S_r|\lambda)$ for the known reference string S_r and $G(X_k, S_n|\lambda)$ for the competing N-best strings S_n, as follows:

$$d(X_k|\lambda) = -g(X_k, S_r|\lambda) + G(X_k, S_n|\lambda), \tag{18}$$

where λ is the HMM parameter set, and $G(X_k, S_n|\lambda)$ is a weighted sum over the competing N-best strings. Given the misclassification measure, only the local accumulation of the string-level errors can be minimized. However, as argued, it is not appropriate to ignore a direct measure of deletion and insertion errors in discriminative training.

As an alternative, the so-called enhanced minimum classification error (E-MCE) training algorithm was proposed in [62]. MDE, MIE, and MSE were constructed by training three sets of competing strings from the constrained N-best search within the conventional MCE framework. However, E-MCE is not a direct individual error minimization method, but a balanced method for the three types of the recognition errors. Furthermore, since E-MCE explicitly follows the conventional string-based MCE framework based on the misclassification measure in Eq. (18), the objective function of the E-MCE still focuses on minimizing the empirical average loss of the three errors in the given competing string.

To construct individual direct objective functions for deletion and insertion errors, a new training-event-selection scheme is proposed as illustrated in Figure 4. Suppose that the reference string is W^r, and the one-best decoded string from ASR is W^d. After a DP-based string alignment procedure, one deletion error W_2^r and one insertion error W_2^d are counted as shown in Figure 4. If we interpret the two recognition errors from a detection viewpoint, the deletion error W_2^r can be regarded as a miss error in the detection problem since W_2^r has to exist on the decoded string, but it is missed with respect to the decoded output sequence. On the other hand, W_2^d has to be rejected, but it is inserted on the decoded output sequence. Thus, the insertion error W_2^d can be viewed as a false alarm error in the detection problem. Then, from the MVE criterion, the segments of the deletion error W_2^r and the insertion error W_2^d are trained by the first mis-verification measure $d_I(X_k, W_2^r|\lambda)$ and the second mis-verification measure $d_{II}(X_k, W_2^d|\lambda)$, respectively, as follows:

$$d_I(X_k, W_2^r|\lambda) = -g_t(X_k, W_2^r|\lambda_t) + g_a(X_k, W_2^r|\lambda_a), \tag{19}$$

$$d_{II}(X_k, W_2^d|\lambda) = +g_t(X_k, W_2^d|\lambda_t) - g_a(X_k, W_2^d|\lambda_a), \tag{20}$$

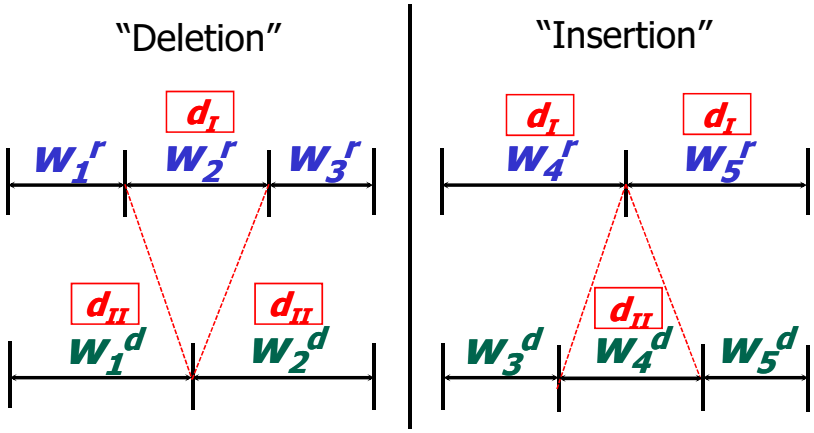


Figure 4: Error count and corresponding mis-verification measures under MVE criterion.

where d_I and d_{II} are the type I and type II mis-verification measures [59, 63, 64], respectively. In Eqs. (2) and (3), g_t and g_a are the normalized log likelihoods, and λ_t and λ_a are the parameter sets of the target model and the anti-model [59, 63, 64, 65] for the given segment, respectively.

This new training paradigm generalized from the MVE criterion can explain the direct relationship between the recognition and detection errors. Nevertheless, it is intuitively obvious that counting only error segments, W_2^r and W_2^d, may not reflect effective model separation and error minimization in the DT phase since the deletion and insertion errors are directly related to the preceding and succeeding segments. In addition, there is a prominent need in identifying the part of speech data containing the potential deletion and insertion errors for the purpose of discriminative parameter optimization. Therefore, a new training framework covering the segments right before and after the error segment is proposed as shown in Figure 4. One can further extend this framework by associating the preceding and succeeding segments with non-uniform error costs or by containing more connected segments with d_I and d_{II} than proposed.

3.1.2 Derivation of Multi-objective Discriminative Training

Segment-based MVE has shown its effectiveness in constructing detectors [63, 64] and rescoring hypotheses [66, 67] from an ASR system for improved continuous speech recognition. In this section, the multi-objective discriminative training generalized from the segment-based MVE criterion is derived in detail.

Suppose there are M classes and K training samples in a given training data set. After DP matching, the given K training samples are assigned into $\{X_1^r, X_2^r, \ldots, X_k^r\}$ for the reference transcript and $\{X_1^d, X_2^d, \ldots, X_k^d\}$ for the decoded output. From the samples and error assignments of the decoded output, the empirical average loss is defined by

$$L(\tilde{\lambda}) = \frac{1}{K} \sum_{k=1}^{K} \ell_{total}(X_k^d|\lambda), \tag{21}$$

where $\ell_{total}(X_k^d|\lambda)$ is the composite loss function which combines four different types of the recognition outputs from the general DP-based string error assignment. For the multi-objective discriminative learning, the composite loss function can be described as

$$\begin{aligned} \ell_{total}(X_k^d|\lambda) &= \ell_{Del}(X_k^d|\lambda)1(X_k^d \in \text{“}Del\text{”}) + \ell_{Ins}(X_k^d|\lambda)1(X_k^d \in \text{“}Ins\text{”}) \\ &+ \ell_{Sub}(X_k^d|\lambda)1(X_k^d \in \text{“}Sub\text{”}) + \ell_{Hit}(X_k^d|\lambda)1(X_k^d \in \text{“}Hit\text{”}), \end{aligned} \tag{22}$$

where $\ell_{Del}(\cdot)$, $\ell_{Ins}(\cdot)$, and $\ell_{Sub}(\cdot)$ denote respectively individual objective functions: MDE, MIE, and MSE.

First, the objective function for MDE can be written as

$$\begin{aligned} \ell_{Del}(X_k^d|\lambda) &= PW_I \sum_{i=1}^{M} \ell(d_I(X_k^r|\lambda^i))1(X_k^r \in C_i) \\ &+ PW_{II} \sum_{j=-1,1} \sum_{i=1}^{M} \ell(d_{II}(X_{k+j}^d|\lambda^i))1(X_{k+j}^d \in C_i), \end{aligned} \tag{23}$$

where PW_I and PW_{II} are the penalty weights for type I and type II errors, respectively, and $\ell(\cdot)$ is a smoothed loss function normally defined by a sigmoid function [12], i.e.

$$\ell(d_I) = \frac{1}{1 + \exp(-\alpha d_I + \beta)}, \tag{24}$$

where α is a constant which controls the slope of the smoothing function, and β sets an offset of the function. Note that the two kinds of mis-verification measures are separately assigned to the reference segment X_k^r and decoded segment X_{k+j}^d as defined by

$$d_I(X_k^r|\lambda^i) = -g_t(X_k^r|\lambda_t^i) + g_a(X_k^r|\lambda_a^i), \tag{25}$$

$$d_{II}(X_{k+j}^d|\lambda^i) = +g_t(X_{k+j}^d|\lambda_t^i) - g_a(X_{k+j}^d|\lambda_a^i). \tag{26}$$

Unlike Eq. (18), in Eqs. (25) and (26), g_t and g_a are the segment-based normalized log likelihood, and λ_t^i and λ_a^i are the parameter set of the target and the anti-model for the i-th class, respectively. In HMMs described in Section 2.1.2, $g(X|\lambda^i)$ can be described as the maximum log likelihood of the state sequence obtained by Viterbi alignment [1]. For example, a set of the *class discriminant functions* $g(X|\lambda^i)$, $i = 1, 2, \ldots, M$ can be expressed by

$$g(X|\lambda^i) = P(X, \mathbf{q}|\lambda^i) = \pi_{q_0}^i \prod_{t=1}^{T} a_{q_{t-1}q_t}^i b_{q_t}^i(x_t), \tag{27}$$

where $\mathbf{q}$ is any state sequence being generated by the Markov chain, and λ^i is the HMM parameter set for the i-th class. In this research, the maximum joint observation-state probability is chosen for the discriminant function $g(X|\lambda^i)$ such that

$$\begin{aligned} g(X|\lambda^i) &= \log\left\{\max_{\mathbf{q}} g(X, \mathbf{q}|\lambda^i)\right\} = \log\left\{g(X, \bar{\mathbf{q}}|\lambda^i)\right\} \\ &= \sum_{t=1}^{T}\left[\log a_{\bar{q}_{t-1}\bar{q}_t}^i + \log b_{\bar{q}_t}^i(x_t)\right] + \log \pi_{\bar{q}_0}^i, \end{aligned} \tag{28}$$

where $\bar{\mathbf{q}} = \{\bar{q}_0, \bar{q}_1, \ldots, \bar{q}_T\}$ is the optimal state sequence that achieves $\max_{\mathbf{q}} g(X, \mathbf{q}|\lambda^i)$. In addition, the output likelihood $b_j^i(x_t)$ of the K-mixture Gaussian can be defined by

$$b_j^i(x_t) = \sum_{k=1}^{K} c_{jk}^i \mathcal{N}(x_t; \mu_{jk}^i, R_{jk}^i) = \sum_{k=1}^{K} \frac{c_{jk}^i}{(2\pi)^{D/2}\left|R_{jk}^i\right|^{1/2}} \exp\left[-\frac{1}{2}\sum_{\ell=1}^{D}\frac{\left(x_{t\ell} - \mu_{jk\ell}^i\right)^2}{\left(\sigma_{jk\ell}^i\right)^2}\right], \tag{29}$$

where $\mathcal{N}(\cdot)$ denotes a normal distribution, D is the dimension of $x_t = [x_{t1}, x_{t2}, \ldots, x_{tD}]'$, c_{jk}^i are the mixture weights, $\mu_{jk}^i = \left[\mu_{jk\ell}\right]_{\ell=1}^D$ the mean vector, and R_{jk}^i the covariance matrix which, for simplicity, is assumed to be diagonal, i.e. $R_{jk}^i = \left[\sigma_{jk\ell}^2\right]_{\ell=1}^D$, of the k-th mixture component in the j-th state for the i-th HMM model.

Similar to MDE, the objective function of MIE can be written as

$$\begin{aligned}\ell_{Ins}(X_k^d|\lambda) &= PW_I \sum_{j=-1,1} \sum_{i=1}^{M} \ell\left(d_I(X_{k+j}^r|\lambda^i)\right) 1(X_{k+j}^r \in C_i) \\ &+ PW_{II} \sum_{i=1}^{M} \ell\left(d_{II}(X_k^d|\lambda^i)\right) 1(X_k^d \in C_i). \end{aligned} \tag{30}$$

For MSE, as discussed, the substitution error can be regarded as miss and false alarm errors happening together at the given segments. As is done above, the objective function of MSE can be formulated as

$$\begin{aligned}\ell_{Sub}(X_k^d|\lambda) &= PW_I \sum_{i=1}^{M} \ell\left(d_I(X_k^r|\lambda^i)\right) 1(X_k^r \in C_i) \\ &+ PW_{II} \sum_{i=1}^{M} \ell\left(d_{II}(X_k^d|\lambda^i)\right) 1(X_k^d \in C_i). \end{aligned} \tag{31}$$

Last, as in the conventional segment-based MVE, the hit tokens can be optionally trained either on the reference transcript or on the decoded output.

Finally, the minimization of each objective function can be accomplished through the generalized probabilistic descent (GPD) algorithm [12, 60, 51, 14]. According to an iterative procedure with the given training data, all the parameters in λ_t and λ_a follow the update rule of the GPD algorithm as defined by

$$\lambda_{k+1} = \lambda_k - \epsilon_k \nabla \ell(X_k|\lambda)\big|_{\lambda=\lambda_k}, \tag{32}$$

where ϵ_k is a learning rate, and k is the cumulative number of the processed training samples at time t. In this research, the optimization algorithm above is operated on a sample-by-sample update. For brevity, here the updating process is derived only for the mean vector in the parameter set. The discriminative adjustment of the mean vector in the target model parameter set λ_t^i follows

$$\tilde{\mu}_{jk\ell}^i(n+1) = \tilde{\mu}_{jk\ell}^i(n) - \epsilon_n \frac{\partial \ell(X_n|\lambda)}{\partial \tilde{\mu}_{jk\ell}^i}\bigg|_{\lambda=\lambda_k}, \tag{33}$$

where $\tilde{\mu}_{jk\ell}^i = \mu_{jk\ell}^i / \sigma_{jk\ell}^i$ satisfying the internal constraints [12, 1] in the HMMs. If $X_n \in class\ i$, and $\ell(\cdot)$ is associated with $d_I(\cdot)$ as defined in Eq. (24), then the partial derivative part in

Eq. (33) is expressed in detail as follows:

$$\frac{\partial \ell(X_n|\lambda^i)}{\partial \tilde{\mu}^i_{jk\ell}} = \alpha \ell(X_n|\lambda^i)\left(1-\ell(X_n|\lambda^i)\right)\left(-\frac{\partial g(X_n|\lambda^i_t)}{\partial \tilde{\mu}^i_{jk\ell}} + \frac{\partial g(X_n|\lambda^i_a)}{\partial \tilde{\mu}^i_{jk\ell}}\right). \tag{34}$$

In Eq. (34), the mean vector $\tilde{\mu}^i_{jk\ell}$ is associated only with the output likelihood functions, and the gradient of $g(X_n|\lambda^i)$ is therefore written as

$$\frac{\partial g(X_n|\lambda^i)}{\partial \tilde{\mu}^i_{jk\ell}} = \sum_{t=1}^{T} \delta\left(\bar{q}_t - j\right) \frac{\partial \log b^i_j(x_t)}{\partial \tilde{\mu}^i_{jk\ell}} \tag{35}$$

and

$$\frac{\partial \log b^i_j(x_t)}{\partial \tilde{\mu}^i_{jk\ell}} = \frac{c^i_{jk}}{(2\pi)^{D/2}\left|R^i_{jk}\right|^{1/2} b^i_j(x_t)}\left(\frac{x_{t\ell}}{\sigma^i_{jk\ell}} - \tilde{\mu}^i_{jk\ell}\right)\exp\left[-\frac{1}{2}\sum_{\ell=1}^{D}\left(\frac{x_{t\ell}}{\sigma^i_{jk\ell}} - \tilde{\mu}^i_{jk\ell}\right)^2\right], \tag{36}$$

where $\delta(\cdot)$ is the Kronecker delta function. The last step is to convert $\tilde{\mu}^i_{jk\ell}$ back according to the following equation:

$$\tilde{\mu}^i_{jk\ell}(n+1) = \tilde{\mu}^i_{jk\ell}(n+1)\sigma^i_{jk\ell}(n). \tag{37}$$

Similarly, the derivations for the variance vectors, mixture weights, and transition probabilities can be easily accomplished [12, 60, 14].

In the following experiments, uniform penalty weights for both PW_I and PW_{II} are used. The experiments are conducted on each objective criterion and then a simple combination of the multi-objective criteria. Furthermore, the scheme of recognizer output voting error reduction (ROVER) [68] is tested as a post-processing scheme for the multiple ASR system combination of the proposed MIE/MDE/MSE. One can investigate the non-uniform penalty weights and rule-based combinations of the multi-objective criteria with particular constraints such as [69] over the proposed training framework.

3.1.3 Recognition Experiments on the TIMIT Database

The experiments reported in this section are carried out on the TIMIT database, and the standard experimental setup as specified in [70] is used. Phone and word recognition tasks are conducted, respectively.

As a baseline in phone recognition, context-independent (CI) HMM phone models are trained by the latest version of the hidden Markov model toolkit (HTK) [71]. The CI system consists of 48 monophones defined in [70], and all phones except for the short pause "sp" are modeled by three-state left-to-right HMMs with 70 Gaussians per state. The short pause model "sp" has only one state. The anti-models needed in the likelihood ration test share the same structure as the recognition models, which are regarded as the target models in the proposed training framework. In the phonetic recognizer's evaluation, a bigram language model over phones estimated from the training set is used. In addition, forty-eight monophones are merged into 39 monophones according to the standard mapping described in [70], and the confusion among the merged phones is not considered as errors. The number of training iterations for all MCE and the proposed method in Table 2 is fixed to be five.

On the other hand, in word recognition, context-dependent (CD) target models and CI anti-models are trained. The CI anti-models consist of 41 monophones that are folded from the 48 monophones defined in [70]. Separately, the set of cross-word triphone target models contains a total of 4,328 physical triphone models with 1,024 tied-states. In both the CI and CD models, all phones are modeled by three-state HMMs with each state having eight-mixture Gaussian components. In the word-recognition evaluation, a bigram language model over words estimated from the training set is used. For the proposed discriminative training, a word-loop network is used to generate competing strings in the training data. In addition, the number of training iterations for all MIE/MDE/MSE in Table 2 is fixed to be three.

In all experiments, the speech is represented by 39 dimensional feature vectors with 12MFCC, 12Δ, 12$\Delta\Delta$, and three log-energy values. The standard 3,696 training utterances excluding the "sa" utterances and 192 core-test utterances were used for training and testing, respectively.

In the phone-recognition task, the phone accuracy rate of the baseline system is 70.57%

Table 2: Phone accuracy rate (%) comparison for ML, MCE, and multi-objective training techniques.

	Del	Ins	Sub
ML	678	170	1289
MCE	674	179	**1265**
MDE	**655**	175	1279
MIE	687	**156**	1278
MSE	691	159	**1273**
D+I+S	687	159	1272
H+D+I+S	521	274	1278

after four iterations with ML estimation using the bigram language model. A performance comparison between the conventional string-based MCE and the proposed multi-objective DT method is detailed in Table 2. In particular, the detailed performance of each objective criterion and two kinds of simple combinations of individual objective criteria such as "D+S+I" and "H+D+S+I" is presented. Note that in the combined multi-objective training methods, the three error segments and "hit" segments are simply incorporated into the DT phase.

As shown in Table 2, it is evident that MCE mainly reduces the substitution error as intended. However, each objective criterion of MDE, MIE, and MSE results in primarily reducing its target error type, respectively. Furthermore, although the simple combinations of the individual objective criteria are constructed, the two combined multi-objective training methods still confirm the effectiveness of the proposed training framework. A rule-based optimization method such as [69], unlike the simple combinations reported here, may bring about a higher overall error reduction.

In the word-recognition task, the word error rate (WER) of the baseline system is 44.59% with the recognition models trained by ML estimation. A performance comparison between the baseline ML and the proposed training methods on the WER is detailed

Table 3: Word error rate (%) comparison between ML and multi-objective training techniques.

	Del	Ins	Sub
ML	89	84	527
MDE	**76**	87	533
MIE	95	**71**	527
MSE	88	83	**514**
ROVER	81	79	522

in Table 3. In Table 3, it appears that the proposed training framework leads to direct minimization of word-level individual errors. However, compared to the ML baseline, MIE and MDE yield more deletion and insertion errors, respectively. One possible cause of such instability is a lack of modeling anti-models with a corresponding discriminability. As mentioned, the anti-models for the limited 41 CI monophones were employed during evaluation with the CD target models in the DT phase. It is likely that use of the CD anti-subword models discriminatively trained with the corresponding CD target models would lead to improved performance as shown in [65].

Furthermore, the ROVER as a post-processing scheme is used to combine the multiple ASR outputs of the proposed MIE/MDE/MSE. The ROVER algorithm was originally proposed to improve the performance of speech recognition by combining multiple speech recognizers. The outputs of multiple ASR systems are aligned into a word transition network (WTN) by dynamic programming, and then majority voting is performed for each correspondence set. The consensus output yields a word error rate (WER) of 43.44%, which is a slight reduction over the best single system MSE of a WER of 43.63%. Note that this combination scheme of individual recognition outputs using ROVER does not extensively explore the issue of sensitivity of individual error minimization since the essence of ROVER is to extract a consensus/unanimity hypothesis from multiple alternatives.

Table 4: The number of word-tokens in the training hypotheses generated by a word-loop network.

Del	Ins	Sub	Hit
1,957	264	7,087	21,088

We have seen that the proposed learning framework leads to direct minimization of the individual errors. In particular, the deletion and insertion errors, which are typically considered very difficult to handle, were directly reduced by the proposed MDE and MIE, respectively. Although the proposed learning framework achieved very encouraging results in this task, there are still many challenging issues.

In this research, a word-loop network is used to generate the competing hypotheses for the proposed discriminative training while the bi-gram language model is used for decoding. In this setup, the number of tokens for MDE and MIE, i.e., tokens that are likely to cause insertion and deletion errors, is limited. Table 4 shows the number of word-tokens in the training hypotheses generated by a simple word-loop network. As can be seen, most of tokens are correctly recognized and a very limited number of the deletion and insertion errors are obtained (e.g., roughly on the order of one percent of the tokens led to insertion errors). These limited tokens have a great impact on the performance of the proposed MDE and MIE since MDE and MIE are performed only on the corresponding error tokens. Recently, weighted finite state transducers (WFSTs)-based discriminative training [27, 28] has been proposed to produce much more hypotheses for discriminative training. We believe that WFST-based approach would significantly improve the performance of the individual error minimization learning.

In addition, while the testing set as well as the training set in TIMIT contains only a small number of the deletion and insertion errors, yet many of them are articles such as “a” and “the” or one short syllable-based word such as “in” and “on”. This is commonly observed in general read speech databases such as TIMIT and wall street journal (WSJ) [72].

Instead of the read speech databases, spontaneous and conversational speech recognition systems [73, 74] that normally contain a wide variety of deletion and insertion errors would be a promising application in use of the proposed individual error minimization learning framework.

3.2 Adaptive Utterance Verification Framework

In the previous section, although the proposed learning method yields discriminatively trained anti-models and target models at the same time, only the recognition performance by the target model has been investigated. As the discriminatively trained target model can be directly used for the recognition task, the simultaneously trained anti-model with the target model, as a set of detectors, can also be used for detection and verification tasks. Since the proposed individual error minimization learning criteria are essentially generalized from the MVE criterion, it is expected that the proposed method has an intrinsic nature of the MVE criterion; thus, a viable application using the proposed method may be extended to detection and verification.

In this section, utterance verification (UV) [75, 76, 77] is chosen as a target task to determine whether the proposed individual error minimization method can be directly applied to a UV task. In particular, not only the recognition performance, but also the *rejection performance* of the recognition errors in UV will be investigated by using both the target and anti-models. Note that MVE in this section can be viewed as MSE in the previous section. Since a UV task considered in this section consists of isolated keyword recognition followed by verification, there are no deletion or insertion errors.

In contrast to the conventional two-stage UV, an integrated solution is proposed to enhance the overall UV system performance. The integration is accomplished by adapting and merging the target model for UV with the acoustic model for ASR based on the MVE principle at each iteration in the recognition stage. The proposed iterative procedure for

UV model adaptation also involves revision of the data segmentation and the decoded hypotheses.

3.2.1 Utterance Verification

Conventional ASR systems are generally task specific with a fixed system construct, such as vocabulary and grammar, which does not provide a user-friendly interface with flexibility in accepting a wide range of user responses. The performance of these systems is seriously degraded by out-of-vocabulary (OOV) words (improper input utterances) spoken by the user or mismatched operating designs, such as different training and testing conditions. To enhance the ASR performance for a friendlier voice user interface, it is necessary to provide a mechanism for verifying the level of confidence in the recognition results. Such a mechanism should reject OOV utterances, as well as potentially misrecognized utterances, to avoid detriments caused by senseless recognition errors. This mechanism is often referred to as utterance verification (UV).

The conventional UV framework consists of a recognition stage and a verification stage as shown in Figure 5. In the recognition stage, the decoder produces a tentatively recognized output for the verification stage. The decoder produces the output using generally trained acoustic (recognition) models. The verification system considers the recognition output as hypotheses and verifies the confidence level for the provided tentative decisions. The UV system determines the scores of the hypotheses by using the corresponding target models and anti-models – a set of verification models – on the segments of the hypotheses provided by the decoder. Finally, in the evaluation stage, a ratio of the scores is compared to a pre-specified operating threshold. Based on the threshold, a final decision is made to either accept or reject the hypothesis.

In this research, UV refers to the ability to accept or reject a hypothesized word corresponding to a correctly decoded keyword, an incorrectly decoded keyword, or an OOV word. This capability, different from the conventional formulation of speech recognition,

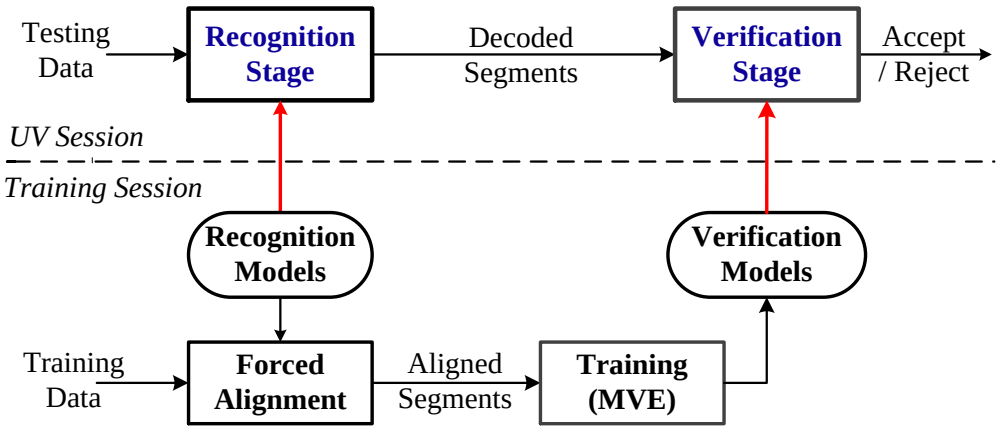


Figure 5: Basic architecture of two-stage system in conventional UV.

is implemented as a likelihood ratio-based hypothesis testing procedure for verifying individual subword units in a decoded word as a result of ASR decoding. In other words, the verification is performed as post-processing after the recognition.

Conventional hypothesis testing in the verification stage is based on the Neyman-Pearson lemma [78], which teaches the use of the likelihood ratio to accept or reject a proposed hypothesis as defined in

$$LR(k) = \frac{P(X_k|H_0)}{P(X_k|H_1)} \gtrless \tau_k \text{ ; Accept or Reject.} \tag{38}$$

A generalized likelihood ratio is computed when testing data X_k is observed and then compared against a decision threshold to decide which one of two hypotheses is to be accepted. The two hypotheses are the null hypothesis H_0 corresponding to the target model and the alternative hypothesis H_1 corresponding to the anti-model. Hypothesis testing is performed by comparing the likelihood ratio $LR(k)$ to a pre-specified operating threshold τ_k. If the two likelihood functions of $P(X_k|H_0)$ and $P(X_k|H_1)$ are known exactly, the above likelihood ratio test is the most powerful test [78]. However, the true likelihood or distribution functions are unknown in a real-world application.

3.2.2 Limitations of Conventional Utterance Verification Framework

As shown above, a reliable estimate of the verification models plays a key role in UV since hypothesis testing is performed by the discrimination (ratio) between the target and the anti-models. In the context of UV, MVE training has shown successful results in several UV tasks. Nevertheless, an additional level of uncertainty needs to be addressed, namely the potential mismatch in the statistical behaviors of the training data and of the field data. Since the pre-labeled data normally consisting of phoneme boundaries – the start and end times of each phoneme on a reference transcription – are at best a limited representation to support the given recognition models, the parameters optimized for a given training set often undergo significant degradation under mismatch operating conditions.

Furthermore, although the two stages in UV may jointly affect the overall verification performance, many researchers have been considering the first stage (recognition stage) and the second stage (verification stage) separately, as shown in Figure 5. Integrating speech recognition and UV in a single decoding scheme is believed to offer substantial performance improvement, particularly for speech signals containing OOV words, ill-formed words, or ill-modeled utterances. Past attempts at such integration include the hybrid decoder proposed in [79] and the one-pass likelihood ratio-based decoder proposed in [77]. Although these proposals take advantage of information from anti-models and likelihood ratio testing, the benefits in general do not materialize simultaneously in terms of recognition and verification performances.

3.2.3 Adaptive Utterance Verification Framework

As shown in Section 3.1, the discriminatively trained target model directly leads to performance improvement in recognition. Thus, it is expected that the label information obtained from the target model can be advantageously utilized to adapt the model parameters to the field data. In contrast to the conventional UV framework, in which the label information obtained from the recognition model is fixed throughout the training stage, in this proposed research, labels and segmentations are sequentially updated along with the target model

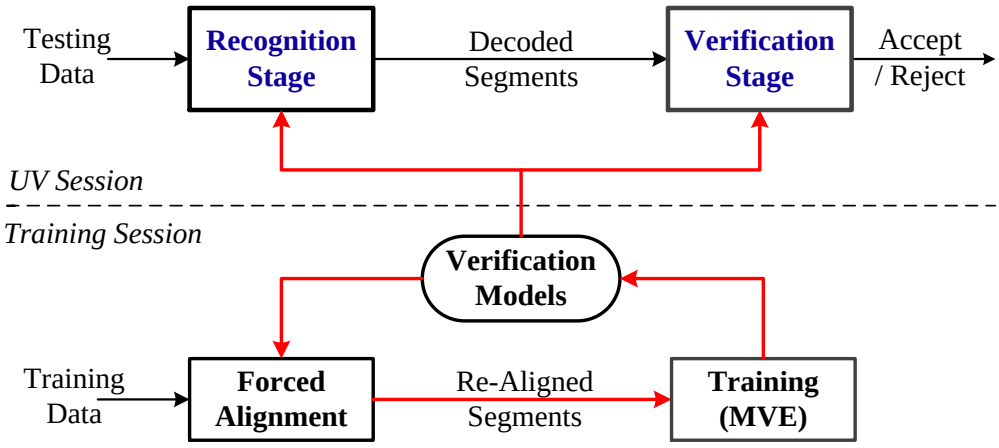


Figure 6: Adaptive UV framework.

refinement in discriminative training with the given adaptation data. That is, the verification models are updated iteratively by discriminative training, using a matched set of data, which are associated with iteratively obtained labels and segmentations, as shown in the training session of Figure 6.

Furthermore, in the context of the conventional UV, the recognized hypotheses do not change regardless of the UV models. This limitation of using only the recognized hypotheses carried out by the recognition models may substantially affect the entire verification framework. It is obvious that improved segmentation and duration in a way consistent with the verification models will directly affect the verification performance. Meanwhile, if the recognition error is improved, resulting in a reduced portion of the incorrectly recognized hypotheses, the entire verification framework will deliver superior performance. Hence, as an integrated solution for the entire verification framework, the use of target models updated in MVE training is proposed for the recognition stage again, as shown in the UV session of Figure 6.

The proposed UV framework can be considered essentially as one integrated stage associated with only the verification models in contrast to the conventional rigid two stages associated with the inconsistent recognition models and verification models as shown in Figure 5. In this new framework, at every iteration during discriminative training, not only

the label information for the next MVE training, but also the recognized output for the hypothesis testing is sequentially updated by the current-stage MVE target model. Hence, throughout the adaptive UV framework with MVE training, improved decoding results and discriminatively trained verification models can be simultaneously obtained. It is obvious that the updated decoder would produce a possibly better set of hypotheses than the fixed decoder for the verification stage.

3.2.3.1 Segment-based Minimum Verification Error (MVE) Training

The MVE training method can be viewed as a special version of the MCE method for detection and verification problems. Similar to the MCE criterion, the objective of MVE training is to directly minimize the empirical average loss. In contrast to the conventional string-based MVE [60, 61], here the segment-based MVE [63, 64] will be derived. Note that the string-based MVE was initially designed to minimize the empirical average loss in the given strings when a pair of detectors is used as a recognizer. Hence, it still focuses on minimizing recognition errors rather than verification errors. Alternatively, segment-based MVE directly minimizes the total verification errors as the weighted sum of type I and type II errors not in the given strings, but in the given segments. An obvious advantage of segment-based MVE is that the intrinsic properties of the speech signal, which is based on segments during recognition and verification, can be directly embedded into the training phase. Accordingly, the total verification errors latent in every given segment are efficiently minimized. In this section, the theoretical framework of the segment-based MVE is briefly reviewed.

Suppose there are M classes and K training tokens (segments) in a training set. For a given training set $\{X_1, X_2, \ldots, X_k\}$, the empirical average loss is defined by

$$L(\tilde{\lambda}) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{M} \ell_{total}(X_k|\lambda^i) 1(X_k \in class\ i), \tag{39}$$

where $1(\cdot)$ is an indicator function that returns one when the condition set in its argument is satisfied and zero otherwise, and $\ell_{total}(X_k|\lambda^i)$ is the composite error estimation function

which combines two different kinds of verification errors: type I error (miss) and type II error (false alarm). The composite error estimation function can be described as

$$\ell_{total}(X_k|\lambda^i) = PW_I \ell_I(X_k|\lambda^i) + PW_{II} \sum_{j=1, j\neq i}^{M} \ell_{II}(X_k|\lambda^j), \tag{40}$$

where PW_I and PW_{II} are the penalty weights for type I and type II errors, respectively, and ℓ_I and ℓ_{II} are smoothed loss functions to approximate the empirical verification error on each training sample X_k defined as follows:

$$\ell_I(X_k|\lambda^i) = \frac{1}{1 + \exp(-\alpha d_I(X_k|\lambda^i) + \beta)}, \tag{41}$$

$$\ell_{II}(X_k|\lambda^j) = \frac{1}{1 + \exp(-\alpha d_{II}(X_k|\lambda^j) + \beta)} \; ; \; j = 1, 2, \ldots, M, \; j \neq i, \tag{42}$$

where $d(X_k)$ is the mis-verification measure for the two types of detection errors. The two misclassification measures for each incoming training token X_k labeled as the i-th class event can be formulated according to Eqs. (25) and (26), respectively. In addition, the discriminant functions, g_t and g_a, for the target and anti-models are defined by Eq. (27). In this research, the maximum joint observation-state probability as defined by Eq. (28) is also chosen for the discriminant functions.

Finally, according to an iterative procedure with the given training data, all the parameters in the target and anti-models follow the update rule of GPD algorithm as defined by Eqs. (32–37) when minimizing (39).

3.2.4 Experiments

All experiments presented in this section were conducted on distance-talking and noisy-speech databases collected under four different remote talking conditions: 30 centimeters, 60 centimeters, 100 centimeters, and 150 centimeters corresponding to the distance between a talker and the microphone.

In all evaluation sets, the number of keywords and that of OOV words are chosen to be identical. Each of the databases comprises of 1,470 utterances recorded by 49 speakers, with 30 utterances per speaker. Each utterance consists of an isolated word such as a

command or point of interest for a voice control application of an in-car navigation system. For the keyword detection and OOV word rejection experiments, a set of 130 keywords and 50 OOV words are predefined in the 1,470 utterances. Among the 1,470 utterances, a set of 1,113 (75.71%) utterances contains 130 keywords considered as legitimate inputs, and the other 357 (24.29%) utterances contain 50 OOV words considered as invalid inputs to be rejected by the system.

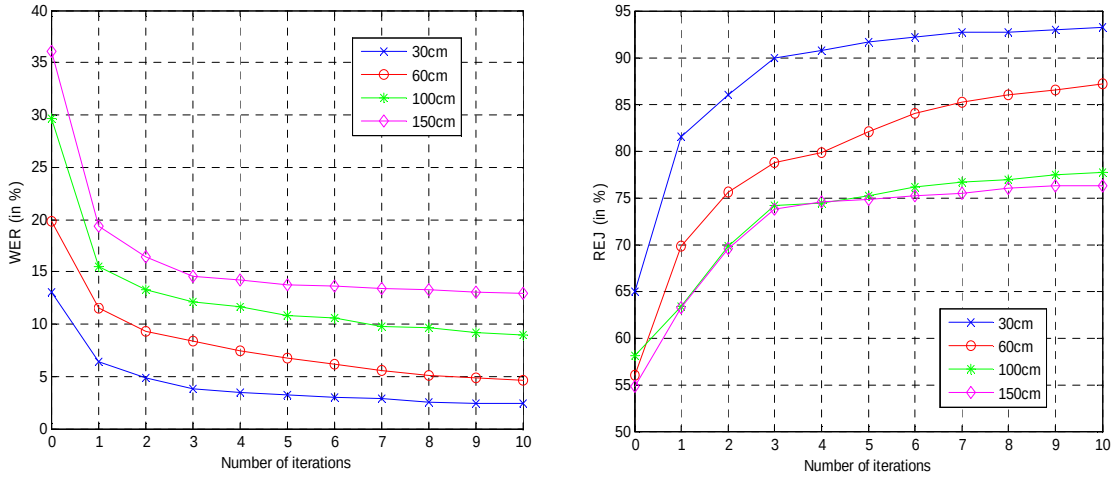
As a baseline, a set of 45 Korean monophone models were used. All models are represented by three-state strict left-to-right HMMs with 16 Gaussian mixture components per state. For the baseline recognition models and verification models, a large-vocabulary speech corpus consisting of 1,700,000 phone-optimized word utterances, 40,000 sentence-based utterances, and 160,000 distant-talking utterances was used for training the initial ML models. Then, the ML-trained models were refined by the conventional MVE method. The refined-MVE models have been used for all adaptation experiments as the baseline models.

On the adaptation side, the baseline models were trained based on the two MVE training scenarios: In the first scenario, conventional MVE training under the two-stage conventional UV framework is performed without updating the transcription during MVE training and the recognition hypotheses in the verification stage. In the second scenario, adaptive MVE (A-MVE) training is performed to yield the updated transcriptions for the next training phase and the improved recognition hypotheses in the verification stage.

Both are trained with 490 utterances (one third of the total 1,470 testing utterances) randomly chosen in the keyword utterances at each iteration. Then, the DT procedure is performed over 10 iterations. As discussed, at each iteration, the label information on the transcription is realigned by the current-stage MVE target model. Also, the updated label information is used for the next DT stage.

The changes of the overall UV performance by A-MVE are illustrated in Figure 7. In detail, the changes of WER and OOV rejection rate (REJ) with increasing number of

iterations about the four different databases are shown in Figure 7 (a) and (b), respectively. In both WER and OOV REJ, there is no performance degradation iteration-by-iteration, and most of the performance gains have been achieved largely in the first three iterations. In addition, after eight iterations, the performance change curves in both Figure 7. (a) and (b) are flat.



(a) Change of WER. **(b) Change of OOV REJ.**

Figure 7: Changes of performance (%) over 10 iterations about four different databases. (a) Change of WER. (b) Change of OOV rejection rate.

3.2.4.1 30 cm Database

An overall performance comparison on three different methods, which are the baseline, the conventional MVE, and the adaptive MVE (A-MVE), respectively, is presented in the Table 5. From the second row in Table 5, with no rejection (that is, REJ=0.0%), the initial WER of 29.05% is observed by the baseline model. On the other hand, with the verification, the WER is reduced to 13.09% at seven percent false REJ and 8.66% at 15% false REJ. Furthermore, after the verification, the REJ of the OOV words is 64.99% at seven percent false REJ and 79.27% at 15% false REJ, respectively.

The overall performance by the MVE-trained model under the conventional UV framework is summarized in the third row (MVE) of Table 5. With the verification, the WER drops to 4.01%, and the OOV REJ is increased to 90.48% at 15% false REJ. Although

Table 5: Overall performance comparison on 30cm database.

	WER at 0% rejection	WER / OOV REJ at 7% false rejection	WER / OOV REJ at 15% false rejection	EER
Baseline	29.05%	13.09% / 64.99%	8.66% / 79.27%	17.08%
MVE	29.05%	7.98% / 78.99%	4.01% / 90.48%	12.49%
A-MVE	**25.37%**	**3.77% / 89.92%**	**1.48% / 96.08%**	**8.26%**

the MVE method under the conventional UV framework produces substantial word error reduction and improved OOV REJ compared to the baseline performance, the proposed method, the A-MVE under the adaptive UV framework, confirms that considerable additional gains of performance can be achieved all over the performance metrics. In particular, the WER has been reduced to 3.77% and 1.48% at seven percent false REJ and 15% false REJ, respectively. In addition, with respect to the OOV REJ, remarkable performance improvement is also obtained. The OOV REJ of 89.92% and 96.08% is achieved by the A-MVE method at seven percent false REJ and 15% false REJ, respectively.

Finally, the EER performance is presented in the rightmost column of the Table 5. It can be shown that the EER of the A-MVE is significantly reduced compared to the baseline as well as the MVE. For details, Figure 8 shows detection error tradeoff (DET) curves of the three different methods on the 30 cm database.

3.2.4.2 60 cm and 100 cm Databases

The second and the third testing sets are the "60 cm database" and the "100 cm database" with a larger recording distance than the 30 cm database. As we have observed in the 30 cm database, the A-MVE significantly reduces the WERs, both with and without the verification, and notably improves the verification performance, the OOV REJ, and the EER on both databases. Details of a performance comparison on these databases are presented in Tables 6 and 7, respectively.

In particular, at 7% false REJ on the 60 cm database and the 100 cm database, the

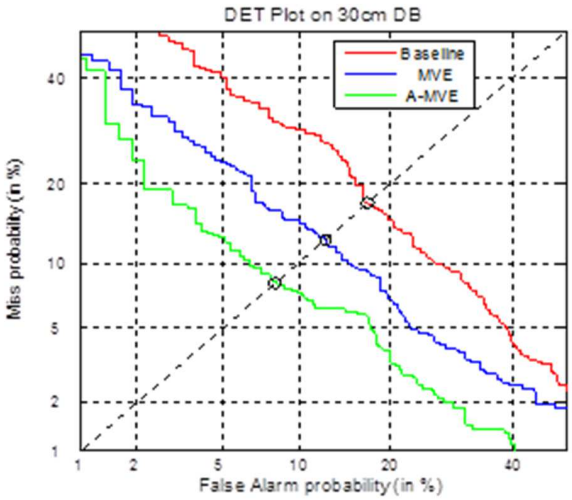


Figure 8: DET curves of three different methods on 30cm database; the circles on the diagonal line are EER points.

Table 6: Overall performance comparison on 60cm database.

	WER at 0% rejection	WER / OOV REJ at 7% false rejection	WER / OOV REJ at 15% false rejection	EER
Baseline	37.69%	19.79% / 56.02%	14.88% / 70.87%	19.03%
MVE	37.69%	15.24% / 65.83%	9.64% / 81.79%	14.97%
A-MVE	**28.11%**	**8.38% / 78.71%**	**3.74% / 91.88%**	**12.22%**

proposed framework using the A-MVE training reduces the WER by further 6.86% and 7.70% and simultaneously increases the OOV REJ by further 12.88% and 3.38%, respectively, over the conventional framework using the MVE training.

From the experimental results on the 30 cm, 60 cm, and 100 cm databases, it is clear that under the proposed adaptive UV framework, the two types of false alarms (misrecognized keywords and OOVs) are minimized while the detection of correctly recognized keywords is maximized.

3.2.4.3 150 cm Database

The last testing set is the 150cm database with the longest distance between a talker and the microphone among all the databases. The performance comparison of ML, MVE, and

Table 7: Overall performance comparison on 100cm database.

	WER at 0% rejection	WER / OOV REJ at 7% false rejection	WER / OOV REJ at 15% false rejection	EER
Baseline	52.15%	29.67% / 58.03%	22.60% / 71.55%	18.68%
MVE	52.15%	19.90% / 70.70%	13.62% / 80.28%	13.56%
A-MVE	**33.06%**	**12.13% / 74.08%**	**5.44% / 89.58%**	**12.60%**

A-MVE is detailed in Table 8. In particular, the baseline performance is seriously degraded from 29.05% to 59.71%, in terms of the WER, compared to the 30cm database. Even with the verification, the performance is limited to the WER of 26.82% and the OOV REJ of 74.01% at a 15% false REJ. By the A-MVE method, the WER rapidly drops from 59.71% to 39.06%, even without the verification. Furthermore, with the verification by the A-MVE, the WER is reduced to 7.66%, and the OOV REJ is increased to 88.14% at a 15% false REJ.

Table 8: Overall performance comparison on 150cm database.

	WER at 0% rejection	WER / OOV REJ at 7% false rejection	WER / OOV REJ at 15% false rejection	EER
Baseline	59.71%	36.09% / 54.80%	26.82% / 74.01%	17.79%
MVE	59.71%	23.00% / 73.16%	15.91% / 83.90%	12.88%
A-MVE	**39.06%**	**14.59% / 73.73%**	**7.66% / 88.14%**	**13.09%**

As a result, the adaptive UV framework reduces the WER without the verification and also provides benefits with the verification by producing improved knowledge such as segmentation for the hypothesis testing. All experimental results confirm that under the adaptive UV framework, the WER is remarkably reduced, and a substantial improvement of the OOV REJ is achieved simultaneously.

3.3 Chapter Summary

In this chapter, we proposed the individual error minimization learning framework and investigated its applications to speech recognition and utterance verification. First, we re-interpreted the commonly known three recognition error types, namely, insertion, deletion and substitution, from an event detection viewpoint. By considering the deletion, insertion, and substitution errors as miss, false alarm, and simultaneous miss/false-alarm, the MVE criterion was generalized to MD(eletion)E, MI(nsertion)E, and MS(ubstitution)E, as the objective functions for direct minimization of each of the three types of errors. This new training paradigm follows the minimum error principle for acoustic modeling and can explain the direct relationship between recognition errors and detection errors. In addition, the adaptive utterance verification (UV) framework was proposed to enhance the overall UV performance. This new UV system fully utilizes the proposed individual error minimization framework by integrating the recognition and verification stages using the MSE-trained models and thus overcomes several limitations of the conventional rigid two-stage UV system.

In evaluation, we first carried out experiments in phone and word recognition on the TIMIT corpus. Experimental results demonstrated that each objective criterion of MDE, MIE, and MSE results in minimization of its target error, respectively. Furthermore, the UV experiments consisting of keyword recognition followed by OOV rejection were conducted on the ETRI distance-talking speech databases. Throughout the proposed adaptive UV framework, we simultaneously obtained an improved overall system decoder with a much reduced recognition error rate and discriminatively trained verification models which significantly enhance the entire verification performance.

CHAPTER 4

DISCRIMINATIVE LINEAR TRANSFORM-BASED ADAPTATION USING MCE AND MVE CRITERIA

In this chapter, several discriminative linear transform (DLT) based adaptation methods using the MCE and MVE criteria are proposed for speech recognition and detection. In the context of discriminative adaptation, most of the research has been limited to speech recognition and speaker adaptation. To further consider discriminative adaptation for detection and verification tasks under a noisy condition, a new DLT-based adaptation method using the MVE criterion is proposed. The proposed MVE linear regression (MVELR) formulates an objective function as a way of keeping consistency between detector training and performance evaluation under a noisy condition. The essence of MVELR is to estimate a set of discriminative linear transformations, which directly minimize the total detection errors, some of which are due to characteristic mismatch in the given adaptation data compared to the original training data.

Despite the effective discriminability and adaptation capability of the DLT-based adaptation methods, it is well known that these methods suffer from the data-sparseness problem. When the adaptation data are severely limited (less than 10 seconds of adaptation speech/called rapid adaptation), it is highly difficult to obtain a solid and consistent performance improvement. The rationale is that the DLTs are easily over-trained given the extremely limited adaptation data. This problem is well known as the generalization issue in the machine learning literature. To overcome the limitation of the DLTs for rapid adaptation, a regularized MCE (RMCELR) criterion is formulated by introducing the *a priori* distribution as a regularization term to the original MCE empirical risk. This RMCE criterion is applied to the DLT-based adaptation and the RMCE linear regression (RMCELR) adaptation method is proposed for rapid adaptation.

Furthermore, a structuring framework to the prior density estimation is proposed to

better estimate the hyper-parameters of the priors. In this framework, the prior densities for the transform matrices are hierarchically structured in a context decision tree according to the amount of the adaptation data available. Then, the transform matrices are derived using the regularized MCE criterion. For this reason, we call the proposed approach structural regularized MCELR (SRMCELR).

4.1 Discriminative Linear Transform-based Adaptation for Detection and Verification Problems

4.1.1 Introduction

As discussed in Section 2.2, linear transform-based adaptation methods have been widely used in automatic speech recognition. The aim of the linear transform is to adapt the mean parameters of a Gaussian-mixture HMM system using affine transformation as defined by Eq. (8). Based on the transform strategy, the mean parameters μ of the initial ASR system are adapted based on the available adaptation data. In particular, maximum likelihood linear regression (MLLR), a common adaptation method, estimates the linear transforms by maximizing the likelihood of the transforms associated with the adaptation data. The basic idea of MLLR in the adaptation of HMM parameters is shown in Figure 9.

The ML-based transform adaptation method has limited performance in accurately estimating the transforms. As discussed in Section 2.4, when the adaptation data are sparse, the estimated transforms may not reliably adapt to the speaker variation or the environmental distortion encapsulated in the given adaptation data. Furthermore, maximizing the likelihood does not guarantee the minimum error rate in an ASR system since there is no direct relationship between the training criterion and the system evaluation criterion. As a result, discriminative criteria, such as MMI, MCE, and MPE/MWE described in Section 2.3, have been investigated for discriminative adaptation.

Among them, the MCE criterion directly minimizes the empirical classification error over a set of training data. Similar to [55, 24], the MCE criterion can be directly applied

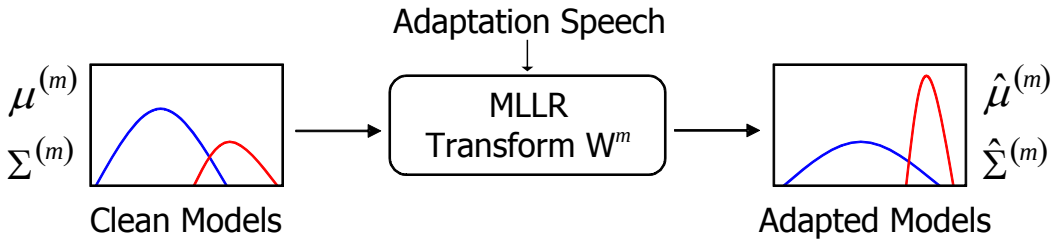


Figure 9: Maximum likelihood linear regression (MLLR) in adaptation of HMM parameters.

to the estimation of linear transforms and model parameter adaptation by using the generalized probabilistic descent (GPD) method. The results show that MCELR outperforms MLLR in recognition accuracy with any given amount of adaptation data. In this research, analogous to MCE-based approaches, the adaptation problem with the minimum verification error (MVE) criterion is investigated thoroughly since there has been little effort in the literature to apply discriminative criteria for detection and verification tasks.

In [63], the effectiveness of the MVE method on various broad phonetic class detection tasks is reported. This paper presents three sets of phonetic category detection as a particular application for detection-based automatic speech recognition (ASR) [80, 81]. These taxonomical sets comprise acoustic-phonetic classes according to their articulatory manners, broad phonetic definition and phonemic identities. The three sets of categorization are often studied in the context of a detection-based approach toward speech recognition. In this section, detectors are designed for adaptation experiments in the same manner as in [63].

Furthermore, in [63], detection errors were significantly reduced in terms of the total error rate since the MVE training method directly minimizes the total verification errors consisting of a combination of type I errors (miss) and type II errors (false alarm). The effectiveness of MVE training in designing detectors has also been confirmed in the previous section. In this research, the MVE criterion is extended for the estimation of linear transforms under the adaptation scenario.

The essence of MVE linear regression (MVELR) is to estimate a set of discriminative linear transformations that achieve the smallest empirical average loss with the given adaptation data. The loss function is minimized by the GPD algorithm according to an iterative procedure. Similar to [63] in the general detection problem, the MVELR directly minimizes the total detection errors, some of which are due to characteristic mismatches in the given adaptation data compared to the original training data. Hence, discriminative linear transforms by MVELR are likely to be more effective in adapting to noisy environments or various speakers than the standard ML-based linear-transform approaches in the minimization of detection errors.

In this section, the MVELR equations are derived, and the MVELR adaptation framework is developed following the MVE criterion. In an adaptation experiment, using MVE-trained target models and anti-models as the initial models for detection of the aforementioned acoustic-phonetic categories, two kinds of adaptation techniques, MLLR and MVELR, respectively, are applied. A comparison study between detectors designed on MLLR and on MVELR is conducted.

4.1.2 MVE linear regression (MVELR) Adaptation

In Section 3.2, MVE training was described in detail, and the effectiveness of the MVE method has been shown in the verification task. In this section, a formulation of MVE linear regression (MVELR) adaptation is derived.

The objective of MVELR is to estimate a set of linear transformations that achieve the smallest empirical average loss with the given adaptation data $\{X_1^r, X_2^r, \ldots, X_k^r\}$. Using the GPD algorithm described in Section 3.1.2, the updated linear transforms W^i can be found by minimizing the empirical average loss defined in Eq. (39). Similar to Eq. (32), the update rule of parameter W^i at epoch k is

$$W^i(k+1) = W^i(k) - \epsilon_k \frac{\partial \ell_{total}(X_k|\lambda)}{\partial W^i}\bigg|_{W=W^i}, \tag{43}$$

where the parameter W^i is defined by $\hat{\mu} = W^i\xi$ and $W^i = \left\{W_t^i, W_a^i\right\}$, $i = 1, 2, \ldots, M$. The

Figure 10: MVE linear regression (MVELR) adaptation.

set of linear transforms, W_t^i and W_a^i, are the linear transforms of the target model and anti-model for the i-th class, respectively. The above partial derivative part is expressed in detail as follows:

$$\frac{\partial \ell_i(X_n|\lambda^i)}{\partial W^i} = \alpha \ell_i(X_n|\lambda^i)\left(1 - \ell_i(X_n|\lambda^i)\right)\left(-\frac{\partial g(X_n|\lambda_t^i)}{\partial W_t^i} + \frac{\partial g(X_n|\lambda_a^i)}{\partial W_a^i}\right), \tag{44}$$

where $X_k \in class\ i$ in the adaptation data set, and W^i is the linear transformations for the i-th class. In Eq. (44), $g(\cdot)$ is the normalized log likelihood function as a class discriminant function defined in Eq. (27). Since the transformation matrices are associated only with output likelihood functions, the gradient of $g(\cdot)$ is written as

$$\frac{\partial g(X_k|\lambda^i)}{\partial W^i} = \sum_{t=1}^{T} \delta\left(\bar{q}_t - j\right) \frac{\partial \log b_j^i(x_t)}{\partial W^i} \tag{45}$$

and the final update equation of W_m^i in each regression class m for class i is written as

$$\frac{\partial \log b_j^i(x_t)}{\partial W^i} = \sum_{r=1}^{R} \frac{c_{m_r}^i}{(2\pi)^{D/2}\left|R_{m_r}^i\right|^{1/2} b_j^i(x_t)} \left(\frac{x_t - \hat{\mu}_{m_r}^i}{R_{m_r}^i}\right) \xi_{m_r} \exp\left[-\frac{1}{2}\frac{\left(x_t - \hat{\mu}_{m_r}^i\right)^2}{R_{m_r}^i}\right], \tag{46}$$

where $\hat{\mu}_{m_r} = W_m \xi_{m_r}$, m_r is the R Gaussian components of a particular regression class m, and W_m^i is a linear transformation of the m-th regression class for the i-th class in the adaptation data set. This MVELR adaptation framework is illustrated in Figure 10. As shown in Figure 10, a set of linear transforms for target and anti-models are separately treated and estimated during discriminative adaptation.

Unlike [24], the scaling of variables in parameter transformation did not show any performance gains. The MVELR parameter transformation is more sensitive than MCELR. One can investigate the scaling in (46) and the optimization problem for MVE framework adaptation.

Table 9: Mapping rule for six-class category.

six-class	Monophones	Percentage (%) in the testing set
fricatives	ch dh f jh s sh th v z zh	16.88
vowels	aa ae ah ao aw ax ay eh er ey ih ix iy ow oy uh uw	39.62
nasals	en m n ng	10.32
stops	b d g k p t	13.08
others	dx el hh l r w y	12.95
silence	sil	7.14

4.1.3 Broad Phonetic Class (BPC) Detection Experiments

Experiments in this section were conducted on the original TIMIT database and one of the distance-talking speech databases [82] recorded from the original TIMIT database. The speech databases, referred to as TIMIT DM, were recorded with a variety of commercial portable devices in a conference/meeting room equipped with sound attenuating wall panels and acoustic ceiling tiles. The original TIMIT database was used for training the baseline models, and the HP1 database was chosen among the five databases [82] for the adaptation and testing.

Similar to [63], three taxonomical phonetic category detectors are defined and trained by the ML method followed by the MVE method on the original TIMIT database. The categories include six classes based on the articulatory manner [1], 14 classes based on the broad phonetic definition from [83], and 48 classes based on monophones defined in [70]. The mapping rules from 48 monophones into the six- and 14-class sets are shown in Tables 9 and 10, respectively. The target models and anti-models in all detectors are constructed by three-state strict left-to-right HMMs and 16-component Gaussian-mixture densities with diagonal covariance matrices.

Based on the initial MVE seed detectors described above, two kinds of adaption techniques, MLLR and MVELR, respectively, are applied using the given adaptation data for comparison. In both MLLR and MVELR, only the mean adaptation was investigated, and the baseline class for each detector is pre-defined to specify the set of components that share the same transform. Three 13×13 block diagonal matrices in all transforms are used for the 39 dimensional feature vectors. In particular, note that in MLLR adaptation, the transforms estimated for the target model are shared in the anti-model. The transforms for the anti-model are not separately considered since the acoustic and environmental distortion is assumed to be properly captured and represented in the transform parameters as part of the signal characteristics. On the other hand, MVELR treats the transforms for the target model and anti-model as separate because the anti-models are constructed strictly for the purpose of facilitating formal hypothesis testing with minimized verification error. MVELR has a different parameter update rule for each transform following Eq. (43).

For training the initial ML and MVE detectors, a total of 3,696 utterances in the training set of the original TIMIT database are used. Regarding the adaptation, some of 3,696 utterances in the training set of TIMIT HP1 are randomly chosen for both MLLR and MVELR adaptations. For all categories, the randomly chosen 200 utterances from TIMIT HP1 were used as the adaptation data. In testing, a total of 1,344 utterances in the testing set of the TIMIT HP1 database are used. All feature vectors have 12MFCCs + energy, and their first- and second-order time derivatives.

Three different kinds of experiments were conducted on the three phonetic categories, six-class, 14-class, and 48-class categories, respectively. The aim of these evaluations is to observe performance gains of MLLR and MVELR under detector-based supervised adaptation scenarios. Performance is obtained based on the minimum total error rate (MTER). As already presented in [63], the MTER is simply the minimum error rate based on an exhausted search of the thresholds when applying the detectors to the test tokens. Hence, the error rate can be seen as a lower bound. In addition, all performance metrics in MVELR

Table 10: Mapping rule for 14-class category.

14-class	Monophones	Percentage (%) in the testing set
front vowels	ae eh ey ih ix iy	20.06
mid vowels	ah ax er	9.33
back vowels	aa ao ow uh uw	7.02
diphthongs	aw ay oy	2.41
voiced fricatives	dh v z	6.50
unvoiced fricatives	f th s sh zh	9.08
affricatives	ch jh	1.30
voiced consonant	b d g	3.72
unvoiced consonant	k p t	7.91
nasals	en m n ng	10.32
liquids	dx el l r	10.91
glides	w y	2.98
whispers	hh	1.31
silence	sil	7.14

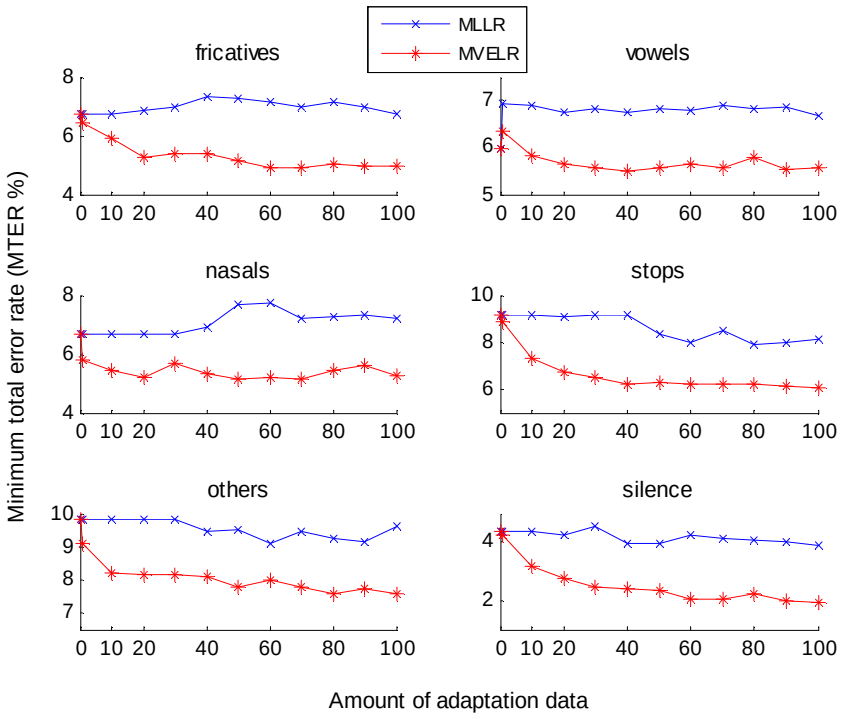


Figure 11: Performance comparison in six-class category with respect to the number of adaptation utterances.

are measured at an iteration of 10.

A performance comparison based on the minimum total error rate with respect to various amount of adaptation data in utterances is shown in Figure 11. It is evident that MVELR performs better than MLLR, even when the amount of adaptation data is seriously limited. The minimum total error rate of all sub-classes in the six-class category is detailed in Table 11. It is clear that the error rate of all sub-classes is significantly reduced when compared to baseline and MLLR performance.

For 14- and 48-class categories, a similar pattern on performance improvement is observed. In the 14 classes, MVELR produces an absolute performance gain of 0.94% compared to using MLLR with respect to the weighted average error rate. On the other hand, in the 48 classes, the weighted average values of the minimum total error rates for MLLR and MVELR are 3.21% and 2.88%, respectively. A detailed performance comparison in the 14 and 48 classes is presented in Tables 12 and 13 [64]. In conclusion, experimental

Table 11: Minimum total error rate (%) for the six-class category.

six-class	MVE Seed	Adaptation	
		MLLR	MVELR
fricatives	6.74	7.00	4.74
vowels	5.99	6.65	5.41
nasals	6.68	7.16	5.15
stops	9.13	8.16	6.10
others	9.81	9.09	7.44
silence	4.36	3.80	1.70
Weighted Average	**6.98**	**7.07**	**5.35**

results confirm that the proposed MVELR method significantly reduces the total error rate and outperforms MLLR over all categories.

Note that no performance gain has been observed in MLLR. One possible reason may be the inconsistency in the estimation of the transforms of anti-models. As previously discussed, in these experiments, no particular transform for the anti-model was considered, and the transform of the target model was simply shared with the anti-model. It is suggested that one investigates the transform for the anti-model with some particular constraints using MLLR.

Table 12: Minimum total error rate (%) for the 14-class category.

14-class	MVE Seed	Adaptation	
		MLLR	MVELR
front vowels	7.62	8.13	7.30
mid vowels	7.59	7.61	7.35
back vowels	5.16	5.30	4.95
diphthongs	2.08	2.00	1.94
voiced fricatives	5.68	6.28	5.57
unvoiced fricatives	5.54	5.14	4.51
affricatives	1.18	1.09	1.06
voiced consonant	3.23	3.25	3.16
unvoiced consonant	6.28	6.03	5.25
nasals	7.57	8.49	5.04
liquids	7.34	7.69	7.26
glides	2.21	2.11	2.25
whispers	1.31	1.31	1.31
silence	4.36	3.79	1.60
Weighted Average	**6.13**	**6.31**	**5.37**

Table 13: Minimum total error rate (%) for the 48-class category.

Name(%)	MVE Seed	Adaptation		Name(%)	MVE Seed	Adaptation	
		MLLR	MVELR			MLLR	MVELR
aa(1.67)	1.55	1.56	1.52	sil(6.01)	4.32	3.40	1.90
ae(1.52)	1.24	1.28	1.22	epi(0.65)	0.64	0.62	0.62
ah(1.70)	1.70	1.68	1.69	k(2.30)	1.99	1.71	1.73
ao(1.50)	1.25	1.25	1.25	l(3.66)	2.51	2.68	2.53
aw(0.43)	0.43	0.43	0.43	m(2.75)	2.45	2.50	2.29
ax(2.81)	2.52	2.54	2.51	n(4.78)	4.35	4.76	4.11
ay(1.35)	1.03	1.04	0.89	ng(0.74)	0.71	0.73	0.71
b(1.07)	0.93	0.89	0.99	ow(1.18)	1.14	1.17	1.13
ch(0.51)	0.51	0.51	0.51	oy(0.25)	0.19	0.24	0.19
d(1.28)	1.25	1.26	1.25	p(1.76)	1.74	1.75	1.69
dh(1.63)	1.59	1.58	1.57	r(3.63)	2.58	2.87	2.46
dx(1.22)	1.00	1.04	0.95	s(4.29)	2.32	2.43	2.16
eh(2.46)	2.28	2.34	2.29	sh(0.91)	0.68	0.69	0.57
el(0.68)	0.64	0.66	0.63	t(2.60)	2.46	2.40	2.24
en(0.43)	0.43	0.43	0.43	th(0.51)	0.51	0.51	0.51
er(3.34)	2.70	2.76	2.68	uh(0.42)	0.42	0.42	0.42
ey(1.58)	1.16	1.29	1.14	uw(1.13)	1.02	1.07	1.00
f(1.80)	1.80	1.59	1.80	v(1.40)	1.39	1.40	1.39
g(0.79)	0.78	0.71	0.76	w(1.77)	1.26	1.21	1.25
hh(1.11)	1.11	1.11	1.09	y(0.74)	0.55	0.66	0.58
ih(2.84)	2.72	2.82	2.78	z(2.44)	2.11	2.30	2.03
ix(4.91)	4.22	4.25	4.20	zh(0.14)	0.14	0.14	0.14
iy(3.57)	1.95	2.04	2.01	vcl(4.98)	4.97	4.93	4.79
jh(0.58)	0.58	0.54	0.53	cl(10.19)	10.10	10.13	8.89
Weighted Average	**3.21**	**3.21**	**2.88**				

4.2 Regularized MCE Linear Regression Adaptation

In the previous section, to overcome the current limitation in utilizing discriminative linear transform-based adaptation for detection and verification tasks under a noisy condition, MVE linear regression was proposed and has generally shown effective discriminability and adaptation capability. However, the MVELR method has shown very limited performance when the amount of adaptation data is less than 10 utterances. In a practical adaptation scenario, in which the amount of adaptation is extremely limited (typically less than 10 seconds of adaptation speech), it is difficult to obtain a solid and consistent performance improvement by DLT-based adaptation. It is well known that DLT-based adaptation methods are subject to the data-sparseness problem [58, 24, 26, 57].

To overcome the limitation of DLT-based adaptation for rapid adaptation, in this thesis we propose a regularized minimum classification error linear regression (MCELR) algorithm for rapid adaptation in which the amount of adaptation data is severely limited. In regularized MCELR, a regularization term is introduced as a weight penalty to the MCELR risk and the penalized empirical risk is then minimized with respect to the transformation parameters. The regularization term in the penalized empirical risk is regarded as a prior distribution of the transformation parameters. The prior knowledge as a regularization term can serve as constraints on the transformation parameters to prevent over-fitting and as interpolation weights for the MCELR estimation process.

This chapter provides an analytical solution for the regularized MCELR framework by deriving the penalized empirical risk in association with the prior distribution of the transform parameters. Adaptation experiments with a small amount of adaptation data are performed on a supervised adaptation scenario using the noisy and distorted speech database, TIMIT HP [64, 82]. We conduct a comparison of the adaptation capability, in terms of the environmental distortion, of four methods, MLLR, MAPLR, MCELR and regularized MCELR (RMCELR), and confirm the effectiveness of the proposed approach especially in a rapid adaptation scenario.

4.2.1 Regularization of Discriminative Linear Transforms

It is well known that DLT-based adaptation methods are subject to the data sparseness problem. For example, the MCELR adaptation method by Wu and Huo [24] was successfully applied to speaker adaptation. However, it showed very limited performance improvement when the amount of adaptation data was less than one minute. This limitation has been shown in several other studies using DLTs [25, 26, 57, 64]. It is mainly because they may lower the discriminability for unseen data while the parameters for observed data may be overly tuned.

This problem is known as the *generalization* issue in the machine learning literature. To deal with the generalization problem, one of the common approaches is to control the capacity or complexity in model training [84]. In this way, *regularization* involves introducing additional information as a penalty for complexity to avoid over-fitting. Another approach is to maximize the margins of training samples closest to the decision boundary [85, 86]. The use of variability of the error margin [87] and variational bounds [88, 89] for regularization has also been proposed to address the generalization issue.

In addition, from a Bayesian point-of-view, a regularization technique is equivalent to imposing certain prior distributions on model parameters. In the ASR literature, maximum *a posteriori* linear regression (MAPLR) [90] and structural MAPLR (SMAPLR) [91] were proposed under a Bayesian framework. A key idea is to take advantage of additional information on the possible values of the transformation parameters when the amount of adaptation data is limited. In the Bayesian framework, this additional information can take the form of a prior distribution of the transformation parameters. However, these methods are still based on the optimal distribution estimation. It is desirable to take advantage of both the Bayesian perspective and discriminative adaptation.

In [92], the use of different forms of a dynamic prior in linear transforms was investigated for rapid speaker adaptation. Prior information estimated by VTLN [93] was used for fast and robust CMLLR transform estimation. In this thesis, to utilize the prior knowledge

in association with the MCE discriminative objective for rapid adaptation, we propose regularized MCELR by introducing a matrix normal prior distribution as a regularization term to the MCELR empirical risk.

4.2.2 MCE Linear Regression (MCELR) Adaptation

As described in Section 2.2 and 4.1.2, linear transforms, W_m, are assigned to a particular regression class m, which consists of R similar Gaussian components as follows: $\{m_r\}_{r=1}^{R}$. In MCELR adaptation [55, 24], the MCE criterion [12] is employed to estimate a set of discriminative linear transformations, $\hat{W}_m$, which achieve the smallest empirical average loss with the given adaptation data.

For a given adaptation data set $\{X_1, X_2, \ldots, X_k\}$, the empirical average loss of MCELR is defined by

$$L_{emp}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \ell\left(d(X_k|\lambda)\right), \tag{47}$$

where K is the total number of adaptation utterances, $\ell(\cdot)$ is a sigmoid function as shown in Eq. (24), and $d(X_k|\lambda)$ is a mis-classification measure. In the string-based MCE training, a string-level misclassification measure is defined by

$$d(X_k|\lambda) = -g(X_k, S_k|\lambda) + G(X_k, S_{k,n}|\lambda), \tag{48}$$

where S_k is the correct label sequence for the k-th utterance, and $S_{k,n}$ is n-th best competing sequence for the k-th utterance, which is a recognized string not equal to S_k. Here, $g(X_k, S_k|\lambda)$ is a discriminant function as defined in Eq. (28), and $G(X_k, S_{n,k}|\lambda)$ is an *anti-discriminant function*, which is a weighted sum over the competing N-best strings [12, 24], defined as follows:

$$G(X_k, S_{n,k}|\lambda) = \frac{1}{\eta} \log\left[\frac{1}{N} \sum_{n=1}^{N} \exp\left[g(X_k, S_{k,n}|\lambda)\eta\right]\right]. \tag{49}$$

For a k-th utterance, $g(X_k|\lambda) > G(X_k|\lambda)$ implies correct classification, and $g(X_k|\lambda) < G(X_k|\lambda)$ means false classification. When η approaches ∞, the anti-discriminant function becomes $\max_{n,n\neq k} g(X_k, S_{k,n}|\lambda)$, which is the best competitor not equal to $g(X_k, S_k|\lambda)$.

Given the above definitions, MCELR can achieve discriminative linear transforms, $\hat{W}_m$, by minimizing the MCE objective function defined in Eq. (47) with respect to W_m. Finally, the update rule of W_m using the generalized probabilistic descent (GPD) algorithm [12, 14] becomes

$$\begin{aligned} W_m(k+1) &= W_m(k) - \epsilon_k \frac{\partial \ell}{\partial W_m}\bigg|_{W_m = W_m(k)} \\ &= W_m(k) - \epsilon_k \alpha \ell (1-\ell) \left(-\frac{\partial g}{\partial W_m} + \frac{\partial G}{\partial W_m} \right) \bigg|_{W_m = W_m(k)}, \end{aligned} \tag{50}$$

where α is a constant which controls the slope of the sigmoid function, ϵ_k is a learning rate, and k is the cumulative number of the given adaptation samples. The derivatives of the discriminant function with respect to W_m follow Eqs. (45–46).

4.2.3 Regularized MCELR Formulation

As discussed, although the effective adaptation capability of MCELR in estimating the parameters of the transformation matrices has been shown in several studies, MCELR generally suffers from the generalization problem given severely limited adaptation data. In this research, to deal with the generalization issue for rapid adaptation, we formulate regularized MCELR by introducing the regularization term to the MCELR objective function.

In regularization, a penalty term $F_{reg}(\lambda)$, which is called a regularizer, is added to the original MCELR empirical risk and the penalized empirical risk can be written as follows:

$$\min_{\lambda} L_{emp}(\lambda) + F_{reg}(\lambda), \tag{51}$$

where $L_{emp}(\lambda)$ is the MCELR empirical risk defined in Eq. (47). If we use a prior distribution of transformation matrices in the regularization term, we can add $-\log P(W)$ as follows:

$$\min_{\lambda} L_{emp}(\lambda) - \log P(W). \tag{52}$$

Then, we define a penalized empirical loss function for a k-th utterance in a regularized

MCELR criterion as follows:

$$\ell^R(X_k|\lambda) = \ell(X_k|\lambda) - \zeta \log P(W), \tag{53}$$

where $\ell(X_k|\lambda)$ is the original MCELR loss function as shown in Eq. (47), ζ is a regularization factor scaling parameter, and $P(W)$ is the prior distribution of transformation matrices, respectively.

The objective of regularized MCELR is to estimate a set of linear transformations which achieve the smallest penalized empirical average loss with the given adaptation data. Using the GPD algorithm, the update rule of a linear transform, which minimizes the regularized loss defined in Eq. (53), is represented as

$$W_m(k+1) = W_m(k) - \epsilon_k \left(\frac{\partial \ell(X_k|\lambda)}{\partial W_m} - \zeta \frac{\partial \log P(W_m)}{\partial W_m} \right) \Bigg|_{W_m = W_m(k)}. \tag{54}$$

Comparing Eq. (54) with the non-regularized MCELR update shown in Eq. (50), it can be seen that a derivative of the prior distribution is used as constraints and interpolation weights to the original MCELR loss.

Main issues here are how to define the prior distribution and how to obtain the derivative of the prior distribution with respect to W_m. A conjugate distribution as the prior distribution is preferable to obtain an analytical solution. In this research, a matrix variate normal density, which can be viewed as a matrix version of a multivariate normal distribution as shown in [90, 91, 94], is used as the prior distribution. Let p be the feature dimension and W_m be a $p \times (p+1)$ matrix, then the matrix normal distribution is defined as

$$\begin{aligned} P(W_m) &= \mathcal{N}(W_m|M_m, \Phi_m, \Omega_m) \\ &\propto \frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left[\Omega_m^{-1}(W_m - M_m)^T \Phi_m^{-1}(W_m - M_m)\right]\right)}{|\Omega_m|^{(p+1)/2}|\Phi_m|^{p/2}}, \end{aligned} \tag{55}$$

where M_m is a $p \times (p+1)$ matrix, Φ_m is a $p \times p$ matrix, $\Phi_m \geq 0$, and Ω_m is a $(p+1) \times (p+1)$ matrix, $\Omega_m \geq 0$. These three matrices, M_m, Φ_m, and Ω_m, are the hyper-parameters to be carefully chosen. Generally, M_m can be obtained by a mean of W_m, and Φ_m and Ω_m are

estimated by

$$\Phi_m = E\left[(W_m - M_m)(W_m - M_m)^T\right], \quad (56)$$

$$\Omega_m = E\left[(W_m - M_m)^T(W_m - M_m)\right]/c, \quad (57)$$

where c is a scalar coefficient. Then, the partial derivative of $\log P(W_m)$ with respect to W_m in Eq. (54) is obtained as follows:

$$\begin{aligned}\frac{\partial - \log P(W_m)}{\partial W_m} &= \frac{\partial}{\partial W_m} - \frac{1}{2}\mathrm{tr}\left[\Omega_m^{-1}(W_m - M_m)^T\Phi_m^{-1}(W_m - M_m)\right] \\ &= -\frac{1}{2}\left(\Phi_m^{-1}(W_m - M_m)\Omega_m^{-1} + (\Phi_m^{-1})^T(W_m - M_m)(\Omega_m^{-1})^T\right) \\ &= -\Phi_m^{-1}(W_m - M_m)\Omega_m^{-1}. \quad (58)\end{aligned}$$

As can be seen, many hyper-parameters have to be carefully estimated and thus make the implementation difficult. In this research, Φ_m is set to the identity matrix as $\Phi_m = \mathrm{I}$, and Ω_m is set to a scaled identity matrix as $\Omega_m = \mathrm{I}/c$. Then, the Eq. (58) is simplified as

$$\frac{\partial \log P(W_m)}{\partial W_m} = -(W_m - M_m)c. \quad (59)$$

Therefore, the update rule of RMCELR as defined in Eq. (54) can be rewritten as

$$W_m(k+1) = W_m(k) - \epsilon_k\left(\frac{\partial \ell(X_k|\lambda)}{\partial W_m} + \zeta c(W_m(k) - M_m)\right)\Bigg|_{W_m = W_m(k)}, \quad (60)$$

where the derivative of $\ell(X_k|\lambda)$ follows Eqs. (44-46).

From the Eq. (60), it can be seen that the update rule during the regularized MCELR estimation is guided by a linear combination of the MCELR estimate $\frac{\partial \ell(X_k|\lambda)}{\partial W_m}$ and the corresponding constraint $\zeta c(W_m(k) - M_m)$. When the sample size n increases, the influence of the constraint diminishes and thus the MCELR estimate is dominant. On the other hand, if n is small, the prior opinion about W_m is strong and the new estimate of the linear transform is highly influenced by the constraint $\epsilon_k\zeta c(W_m(k) - M_m)$.

4.2.4 Rapid Adaptation Experiments

Experiments are conducted on the original TIMIT database and the TIMIT HP [82] database which is one of the distance-talking speech databases, TIMIT DM [64, 82]. The TIMIT

HP database was introduced in Section 4.1.3 and chosen again for rapid adaptation experiments in this section. The original clean TIMIT database was used for training the baseline models and the TIMIT HP database was chosen for adaptation and testing.

To build the baseline acoustic models with maximum likelihood (ML) training, the HTK is first used with a total of 3,696 utterances from the original clean TIMIT database. The set of clean baseline models contains a total of 3,443 physical triphone models with 865 tied-states, and each state is modeled by a 16-component Gaussian mixture. In decoding, a bi-gram language model over phones estimated from the training set is used. In addition, the standard 48 monophones are merged into 39 monophones according to the standard mapping described in [70], and the confusion among the merged phones is not considered as errors. In all experiments, input speech is represented by 39 dimensional feature vectors with 12MFCC, 12Δ, 12$\Delta\Delta$, and three log-energy values.

We randomly chose 2, 4, 6, 8, 10, 25, and 50 utterances from the training set of TIMIT HP database for adaptation while a total of 192 core-test utterances in the testing set of TIMIT HP were used for testing. MLLR adaptation was performed on a regression tree with 31 base-classes for speech and one base-class for non-speech (silence). A leaf occupation count threshold was differently set according to the number of adaptation utterances. For instance, 2, 4, 6, and 8 utterances had fewer than 500 threshold values while 10, 25, and 50 utterances had more than the threshold.

As an initialization for linear transformations, both MCELR and regularized MCELR (RMCELR) commenced the adaptation process using transformation matrices estimated by MLLR. The total number of training iterations was set to be 20 for both MCELR and RMCELR. The initial learning rate ϵ_k was set to be 5.0×10^{-6} for MCELR while it was set to be 1.5×10^{-6} for RMCELR. Then, similar to [24, 57], the learning rate ϵ_k was gradually decreased as the following schedule:

$$\epsilon_{k+1} = \epsilon_k - \frac{\epsilon_0 T_k}{I \sum_{k=1}^{K} T_k}, \tag{61}$$

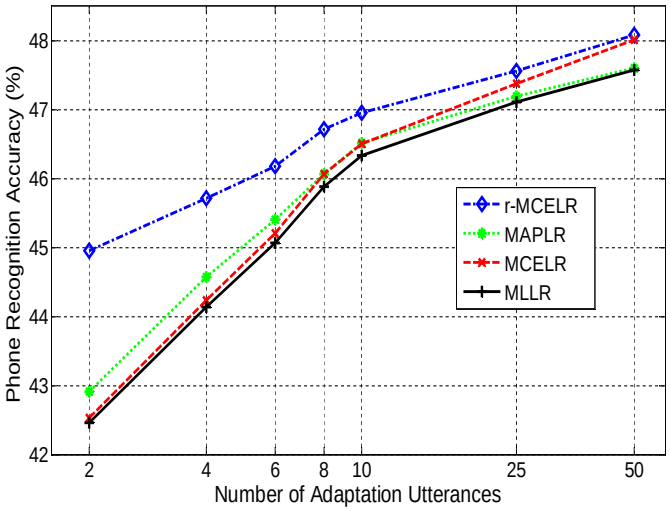


Figure 12: Phone Accuracy Rate (%) of MLLR, MAPLR, MCELR and RMCELR for the number of adaptation utterances.

where I and T_k are the total number of iterations and the number of frames in the k-th cumulative adaptation utterances, respectively. This learning rate is required for the stochastic convergence [95, 96, 97]. The other parameters were set as $\zeta = 1.0$, $\alpha = 0.2$, and $\eta = 20$.

Note that all adaptation experiments were performed under the supervised adaptation scenario. In MCELR and RMCELR, the TIMIT reference transcription was used as a correct sequence, and the *10*-best string lists generated from the baseline recognizer were used as competing sequences. Finally, for the prior distribution, the clean TIMIT database was used to estimate the hyper-parameters. In this research, the identity matrices were set for Φ_m and Ω_m, and the constant parameter c was heuristically handled. The hyper-parameter M_m defined in Eq. (55) was estimated by the mean of the MLLR transforms obtained from all different speakers in a set of the clean training database. This prior distribution is used in both MAPLR and RMCELR.

One main goal of the experiments in this section is to investigate the adaptation capability and the generalization effect of RMCELR compared to MLLR, MAPLR, and MCELR when the amount of adaptation data is extremely limited (less than 10 seconds of speech). In all adaptation experiments, only the mean vectors of the Gaussian components were adapted by using linear transformations. On the core testing set of the TIMIT HP database,

Table 14: Adaptation Performance Comparison in Phone Accuracy Rate (%) for a Rapid Adaptation Task.

Adaptation utterances (seconds)	2 (4.32s)	4 (8.46s)
MLLR	42.46	44.13
MCELR	42.53	44.24
MAPLR	42.91	44.57
RMCELR	**44.95**	**45.71**

the clean baseline yields a phone accuracy rate (PAR) of 33.38% because of the environmental mismatch while it shows a PAR of 69.12% on the clean core testing set.

An overall performance comparison of MLLR, MAPLR, MCELR and the proposed RMCELR with regard to the PAR (%) for various amounts of adaptation data is shown in Figure 12. As previously discussed, when the amount of adaptation data is severely limited, it is clear that MCELR is faced with the generalization problem and thus yields very minor gains over MLLR as reported in several other studies [24, 57, 64]. On the other hand, MAPLR gives slightly better adaptation performance than MLLR and MCELR in this rapid adaptation scenario. The better performance of MAPLR over MLLR and MCELR is attributed to the exploitation of the prior information in the regression parameter estimation. It is demonstrated that the MAP criterion is better than ML and MCE criteria in case of very limited adaptation data. Nevertheless, MAPLR still finds the transform parameters through the optimal distribution estimation as discussed and thus leads to the limited performance improvement [90, 91].

However, the proposed RMCELR adaptation method significantly outperforms MLLR, MAPLR and MCELR in the rapid adaptation scenario. It is mainly because RMCELR iteratively improves the generalization capability and the discriminability with the help of the regularized discriminative estimation instead of the distribution estimation in MAPLR. The proposed RMCELR adaptation method is inherent in the Bayesian property where the

prior information is involved such as MAPLR and in the discriminative nature where the competing hypotheses are applied like MCELR. As a consequence, the prior information as additional assumptions on the transformation parameters efficiently constrains and interpolates the MCELR loss as defined in Eq. (53) for rapid adaptation. In addition, as usual in the Bayesian analysis, as the amount of adaptation data increases, MAPLR and RMCELR converge to MLLR and MCELR, respectively, because of the small impact of prior information in the relatively large amount of adaptation data.

The detailed comparison results, where the amount of adaptation data is fewer than 10 utterances, are summarized in Table 14. As can be seen, the proposed RMCELR considerably outperforms MLLR, MAPLR and MCELR for rapid adaptation. In particular, it is interesting that RMCELR with "two utterances" outperforms MLLR, MAPLR and MCELR with "four utterances." This finding implies that the proposed adaptation method can have a superior effect at half the cost of MLLR, MAPLR and MCELR in this rapid adaptation scenario. Furthermore, although the proposed RMCELR method demands higher computational cost than MLLR and MAPLR, it has the same complexity as MCELR where the hyper-parameters of a prior distribution are obtained.

The influence of the scaling factor c in RMCELR with the smallest amount of adaptation data (two utterances/4.32 seconds) is illustrated in Figure 13. The scaling factor c is used to weigh prior information. If c is given as a small value such as 50 shown in Figure 13, the influence of the prior information is trivial, and thus RMCELR brings about a small gain. On the other hand, if c is set too large such as 300 shown in Figure 13, the transformations are misguided to the adaptation data as the number of iterations increases. Therefore, the adjustment of the scaling factor c is another important issue, as well as the hyper-parameter estimation of the prior density in the proposed RMCELR adaptation.

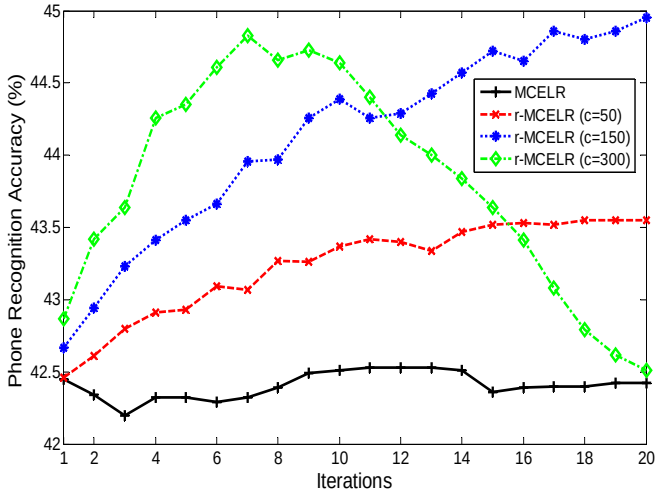


Figure 13: Sensitivity of RMCELR on the different values of the scaling factor c over iterations when the number of adaptation utterances is fixed to two utterances.

4.3 Structuring Framework to Prior Density Estimation for RMCELR

In the previous chapter, the proposed RMCELR adaptation method was trying to find the regularized MCE estimates of linear regression models by combining the prior, $\varphi = \{M_m, \Phi_m, \Omega_m\}$. The hyper-parameters, $\{M_m, \Phi_m, \Omega_m\}$, of the prior distribution are estimated from the training data in an empirical Bayes manner [98]. Although highly improved generalization and adaptation capability have been obtained by RMCELR over MLLR, MAPLR, and MCELR in a rapid adaptation task as reported in Chapter 4.2.4, this limited approach in estimating the hyper-parameters may not be reliable or accurate due to a mismatch between the training and testing conditions.

An estimate of the prior distribution from the training data or the speaker independent model such as RMCELR and MAPLR cannot directly represent the characteristics of the testing condition. A better solution is therefore to estimate the prior distribution directly from the adaptation data in association with RMCELR adaptation. In addition, a regression tree in MAPLR and RMCELR has been used to correlate model parameters and transform matrices W. However, it is necessary to cluster and estimate the prior densities in the given tree structure along with the transform estimation. For example, a prior evolution scheme can be incorporated in each level of the tree as the amount of adaptation data increases.

A structural maximum *a posteriori* (SMAP) adaptation framework was proposed in [99, 100] to provide a proper structuring of model parameters and prior densities. As a transformation-based approach like MLLR, SMAP organizes HMM mean vectors in a tree containing all the Gaussian distributions. Each leaf node in the tree is estimated using the MAP criterion where the prior densities at the leaf nodes are defined as the posterior densities of their parent nodes. This prior/posterior structural information is propagated from the root node down to the leaf nodes of a context decision tree.

A natural extension of the SMAP approach to the linear transform estimation is called SMAP linear regression (SMAPLR) [91, 101]. In SMAPLR, the transform matrices are estimated using a MAP criterion where prior densities for the transform matrices are hierarchically structured in a tree as proposed in SMAP. This hierarchical structure to the priors supports a better use of the adaptation data for the whole estimation process and thus provides more robust estimates to efficiently prevent over-fitting the adaptation data.

We motivate the use of the SMAP technique for the hyper-parameter estimation in RMCELR adaptation. In particular, we propose to add a hierarchical structure to the priors in the proposed RMCELR framework shown in the previous section. The prior densities for the transform matrices are hierarchically structured in a context decision tree according to the amount of the adaptation data available. Then, the transform matrices are derived using the regularized MCE criterion. For this reason, we call the proposed approach in this chapter structural regularized MCELR (SRMCELR). It is expected that more robust estimates can be obtained through the structured priors for rapid adaptation. Also, we expect that SRMCELR outperforms the MAP-based estimation of the transform matrices, such as MAPLR and SMAPLR, and RMCELR using the subjective priors.

4.3.1 Structured Prior Evolution for RMCELR

The structured priors, as a prior evolution scheme, to the linear transform matrices have been used in SMAPLR [91] and EMAPLR [102]. A key idea is that a transform estimated at a certain node can provide some useful information to constrain the estimation of its

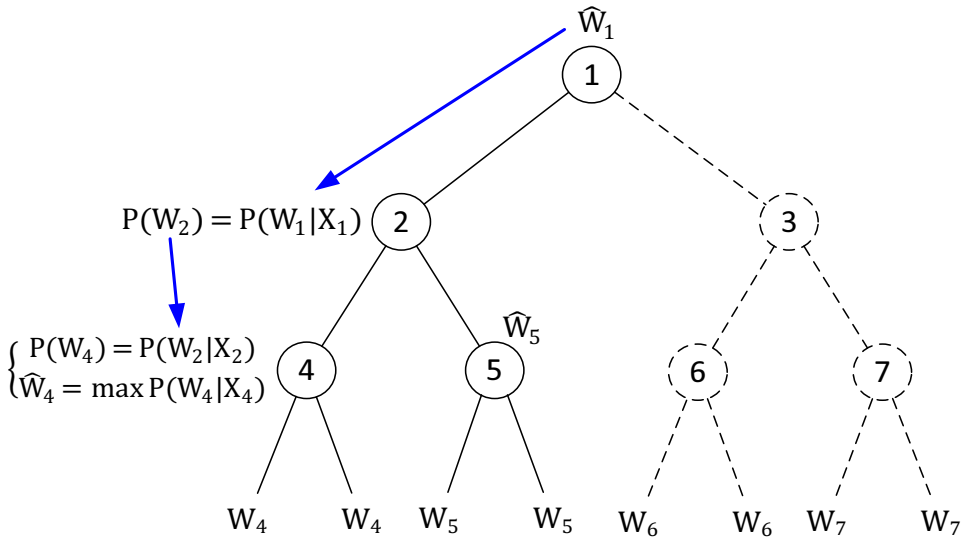


Figure 14: Tree-based SMAPLR Algorithm.

child nodes. Based on the MAP criterion, the posterior distribution at the parent node can be used as the prior distribution for the child nodes. In SMAPLR and EMAPLR, this process is propagated from the root node down to the leaf nodes as illustrated in Figure 14.

Figure 14 depicts an SMAPLR algorithm in a regression tree structure. Suppose nodes 1, 2, 4, and 5 are only valid for estimation based on the leaf occupation count threshold. Transform matrices W_6 and W_7 choose $\widehat{W}_1$ as an optimal solution. On the other hand, to estimate $\widehat{W}_4$ in node 4, the prior distribution $P(W_4)$ can be defined as the posterior distribution in its parent node. It can be derived as $P(W_4) = P(W_2|X_2)$. In the same way, the posterior distribution $P(W_1|X_1)$ in node 1 can be used as the prior distribution $P(W_2)$ in node 2. Such a structural constraint can be also hierarchically derived in a large regression tree. It has been shown that this hierarchical prior/posterior propagation can efficiently reduce the risk of over-fitting the adaptation data and thus improve the adaptation performance especially for very small amount of adaptation data available. SMAPLR and EMAPLR generally outperform MLLR and MAPLR.

In this thesis, we propose to apply the hierarchical prior structure into the RMCELR framework. Similar to SMAPLR/EMAPLR, in structural RMCELR (SRMCELR), the prior densities for the transform matrices are hierarchically structured in a tree based on the

amount of the adaptation data available. However, several issues have to be carefully taken into account to implement SRMCELR. First, the posterior distribution of the parent node which is propagated down to its child nodes as the prior in SMAPLR/EMAPLR is replaced by the approximation of the RMCELR solution in SRMCELR. Then, the approximation of the RMCELR solution of the parent node is propagated down to the child nodes and used as the mode of the prior distribution for the child nodes. Finally, based on the regularized MCE criterion, the prior approximation penalizes the MCELR solution in the child nodes.

In Figure 14, to estimate $\widehat{W}_4$ in node 4, the prior distribution $P(W_4)$ can be approximated by the RMCELR solution in its parent node and incorporated into the RMCELR objective function. Suppose i is an index of a certain node in a tree and $i-1$ is an index of its predecessor node, which means the parent node of the child node i. Then, the prior approximation in SRMCELR can be derived as follows:

$$\begin{aligned} P(W_i) &\approxeq \widehat{W}_{i-1} \\ M_i &= \widehat{W}_{i-1}. \end{aligned} \tag{62}$$

where $\widehat{W}_{i-1}$ is the RMCELR solution in the parent node and M_i is the mode of the prior distribution in the child node. Then, the RMCELR update equation shown in Eq. (60) can be rewritten as

$$\begin{aligned} \widehat{W}_i &= W_i^{MCELR} + \epsilon\zeta \log P(W_i) \\ &= W_i^{MCELR} - \epsilon\zeta c\left(W_i^{MCELR} - M_i\right) \\ &= W_i^{MCELR} - \epsilon\zeta c\left(W_i^{MCELR} - \widehat{W}_{i-1}\right) \end{aligned} \tag{63}$$

where $\widehat{W}_i$ is the SRMCELR solution in the child node i and ϵ, ξ, and c are the GPD step size, the regularization factor, and the scaling coefficient, respectively. In summary, the prior distribution at each node is taken to be the approximation of the RMCELR solution of its parent node. This process is propagated from the root node down to the leaf nodes in a tree structure.

4.3.2 A Comparison Study on DLT-based Adaptation Methods

4.3.2.1 MLLR and MAPLR/SMAPLR

In chapter 2.2, Maximum Likelihood Linear Regression (MLLR) was discussed and derived in detail. The objective function and the final update equation were defined by

$$\hat{\mathrm{W}}_{MLLR} = \arg\max_{\mathrm{W}} P(X|\lambda, \mathrm{W}). \tag{64}$$

and

$$w_{ri}^{ML} = \left(\sum_{m \in M_r} \sum_{t} \gamma_m(t) x_i(t) \xi_m^T \frac{1}{\sigma_{mi}^2} \right) \left(\sum_{m \in M_r} \sum_{t} \gamma_m(t) \xi_m \xi_m^T \frac{1}{\sigma_{mi}^2} \right)^{-1} \tag{65}$$

where w_{ri} is the i-th row of W_r, and $\gamma_m(t)$, $x_i(t)$, and σ_{mi}^2 were described in chapter 2.2. As discussed, when the amount of adaptation data is very sparse, MLLR is faced with the generalization problem and thus W^{ML} is poorly estimated. To overcome the limitation of the MLLR adaptation method, maximum a posteriori linear regression (MAPLR) was proposed using the MAP criterion as follows:

$$\begin{aligned} \hat{\mathrm{W}}_{MAPLR} &= \arg\max_{\mathrm{W}} P(\mathrm{W}|X, \lambda) \\ &= \arg\max_{\mathrm{W}} P(X|\mathrm{W}, \lambda) P(\mathrm{W}). \end{aligned} \tag{66}$$

In this criterion, the prior distribution of transform matrices W is defined by a matrix variate normal density shown in Eq. (55) same as RMCELR and SRMCELR. Given hyperparameters $\varphi = \{M_m, \Phi_m, \Omega_m\}$ of the prior distribution and assuming Φ_m is set to the identity matrix, the final update equation in MAPLR can be defined as

$$w_{ri}^{MAP} = \left(\sum_{m \in M_r} \sum_{t} \gamma_m(t) x_i(t) \xi_m^T \frac{1}{\sigma_{mi}^2} + \mathrm{m}_{ri} \Sigma_{ri}^{-1} \right) \left(\sum_{m \in M_r} \sum_{t} \gamma_m(t) \xi_m \xi_m^T \frac{1}{\sigma_{mi}^2} + \Sigma_{ri}^{-1} \right)^{-1} \tag{67}$$

where m_{ri} and Σ_{ri}^{-1} are mean vector and covariance matrix of the prior distribution for regression row vector w_{ri}. The key difference between MLLR and MAPLR is the use of the prior information which is defined by the matrix variate normal density with hyperparameters, M_m and Ω_m. As seen in Eq. (67), the additional terms, $\mathrm{m}_{ri}\Sigma_{ri}^{-1}$ and Σ_{ri}^{-1}, in both brackets are added in the MLLR solution shown in Eq. (65). These additional terms can

serve as constraints to the estimation of the adaptation data and thus better performance can be obtained for the small amount of adaptation data available.

SMAPLR follows the same update rule as MAPLR shown in Eq. (67), but has a different prior estimation and evolution scheme as explained in the previous section. In estimating the hyper-parameters, especially m_{ri} and Σ_{ri}^{-1}, the structured priors are hierarchically derived in a regression tree structure and estimated directly from the adaptation data available.

4.3.2.2 MCELR and RMCELR/SRMCELR

In MCELR, the MCE criterion is employed to estimate a set of discriminative linear transformations (DLTs), $\hat{W}^{MCE}$, which achieve the smallest empirical average loss with the given adaptation data. The final update equation for regression row vector w_{ri}^{MCE} can be written as follows:

$$\begin{aligned} w_{ri}^{(k+1)} &= w_{ri}^{(k)} + \epsilon_k \alpha \ell(X; w_{ri}) \left(1 - \ell(X; w_{ri})\right) \left(-\frac{\partial g}{\partial w_{ri}} + \frac{\partial G}{\partial w_{ri}} \right) \Bigg|_{w_{ri}=w_{ri}(k)} \\ &= w_{ri}^{(k)} + \epsilon_k \alpha \ell(X; w_{ri}^{(k)}) \left(1 - \ell(X; w_{ri}^{(k)})\right) \\ &\quad \times \Bigg[- \left\{ \sum_t \sum_{m \in M_r} \gamma_m(t) \left(\frac{x_i(t) - w_{ri}^{(k)} \xi_m}{\sigma_{mi}^2} \right) \xi_m^T \right\} \\ &\quad + \left\{ \sum_t \sum_{n \in M_r} \gamma_n(t) \left(\frac{x_i(t) - w_{ri}^{(k)} \xi_n}{\sigma_{ni}^2} \right) \xi_n^T \right\} \Bigg]. \end{aligned} \tag{68}$$

where the HMM labels m belong to the correct transcription of X and the labels n are associated in the competing sequences of X obtained by an N-best list or a phone/word lattice which is not equal to the transcription. From the above update equation, we can see that the adjustment of MCE-based transform matrices $w_{ri}^{(k)}$ is determined by the discrimination of the correct labels against the competing labels. Therefore, the classification error in association with the adaptation data can be efficiently minimized by the MCE-based discriminative adaptation approach.

However, when the severely limited adaptation data is available, the MCE-based DLTs

are generally over-trained to the adaptation data and thus the adjustment of $w_{ri}^{(k)}$ is misguided. To overcome the problem for the very limited adaptation data available, in this thesis we propose regularized MCELR which effectively constrains the MCE-based adjustment by using the prior information to the transform matrices $w_{ri}^{(k)}$. The final update equation of RMCELR can be written as follows:

$$\begin{aligned} w_{ri}^{(k+1)} &= w_{ri}^{(k)} + \epsilon_k \alpha \ell(X; w_{ri}^{(k)}) \left(1 - \ell(X; w_{ri}^{(k)})\right) \\ &\times \Bigg[- \left\{ \sum_t \sum_{m \in M_r} \gamma_m(t) \left(\frac{x_i(t) - w_{ri}^{(k)} \xi_m}{\sigma_{mi}^2} \right) \xi_m^T + \zeta c \left(w_{ri}^{(k)} - \mathrm{m}_{ri}^{(m)} \left(\Sigma_{ri}^{(m)} \right)^{-1} \right) \right\} \\ &+ \left\{ \sum_t \sum_{n \in M_r} \gamma_n(t) \left(\frac{x_i(t) - w_{ri}^{(k)} \xi_n}{\sigma_{ni}^2} \right) \xi_n^T - \zeta c \left(w_{ri}^{(k)} - \mathrm{m}_{ri}^{(n)} \left(\Sigma_{ri}^{(n)} \right)^{-1} \right) \right\} \Bigg]. \end{aligned} \quad (69)$$

where m and n are the correct and competing sequences for X, respectively, same as in MCELR. As can be seen, two different types of additional terms, which are $\zeta c \left(w_{ri}^{(k)} - \mathrm{m}_{ri}^{(m)} \left(\Sigma_{ri}^{(m)} \right)^{-1} \right)$ and $\zeta c \left(w_{ri}^{(k)} - \mathrm{m}_{ri}^{(n)} \left(\Sigma_{ri}^{(n)} \right)^{-1} \right)$ serve as constraints to the MCE adaptation. In particular, the gradient of the correct labels m is constrained by $+\zeta c \left(w_{ri}^{(k)} - \mathrm{m}_{ri}^{(m)} \left(\Sigma_{ri}^{(m)} \right)^{-1} \right)$. On the other hand, the gradient of the competing labels n is constrained in the opposite direction by $-\zeta c \left(w_{ri}^{(k)} - \mathrm{m}_{ri}^{(n)} \left(\Sigma_{ri}^{(n)} \right)^{-1} \right)$. Therefore, the adjustment of $w_{ri}^{(k)}$ in RMCELR can be more robust and accurate than MCELR, by the additional constraints which can sequentially adjust the MCE estimation.

SRMCELR follows the same update rule as RMCELR shown in Eq. (69). However, the different prior estimation and evolution scheme are applied into SRMCELR. As explained in Section 4.3.1, the hierarchical priors are embedded into the tree structure and derived in a prior evolution scheme based on the adaptation data available. Hence, the hyperparameters m_{ri} and Σ_{ri}^{-1} in SRMCELR can be more robust than RMCELR because of the enhanced prior selection and estimation scheme.

4.3.2.3 MAPLR/SMAPLR and RMCELR/SRMCELR

We have studied different DLT-based methods: MLLR, MAPLR, SMAPLR, MCELR, RMCELR, and SRMCELR. From the Eqs (64–69), we can see the clear difference among the

methods. As a conclusion, we summarize some critical difference between MAPLR/SMAPLR and RMCELR/SRMCELR, of which all use the prior information for adaptation.

First, these adaptation methods choose different objective criteria, MAP and RMCE, for optimization. The RMCE criterion takes advantage of both the MCE and MAP criteria. It has been shown in several studies [24, 103, 64] that the MCE criterion is superior to the MAP criterion given larger amount of adaptation data. The MCE discriminative criterion can achieve the optimal model by efficiently minimizing the empirical error in the given adaptation data. However, when the adaptation data is severely limited, the MAP-based adaptation is better than the MCE-based adaptation because of the use of the prior information. In the RMCE criterion, the prior information as a regularization form is incorporated into the MCE criterion. Hence, RMCE overcomes the limitation of MCE for rapid adaptation while keeping an intrinsic nature of the discriminative criterion. It is thus expected that RMCE is better than MAP for rapid adaptation.

Second, the use of the competing hypotheses play a key role in not only training, but also adaptation. As discussed, the adjustment of the transform matrices in RMCELR/SRMCELR is composed of the contributions from both the correct and competing hypotheses. On the other hand, MAPLR/SMAPLR concentrate on optimizing the model where the corresponding labels are introduced in the given correct transcription. This difference in the hypothesis utilization has a great effect on both the prior estimation and adaptation process.

Finally, a critical difference between MAPLR/SMAPLR and RMCELR/SRMCELR is the way in utilizing the prior information. From the final update rules of these methods shown in Eq. (67) and Eq. (69), the difference can be analyzed as follow:

$$\begin{aligned} &\text{MAPLR/SMAPLR:} && \mathrm{m}_{ri}\Sigma_{ri}^{-1} \\ &\text{RMCELR/SRMCELR:} && \begin{cases} \zeta c\left(w_{ri}^{(k)} - \mathrm{m}_{ri}^{(m)}\left(\Sigma_{ri}^{(m)}\right)^{-1}\right) & m \in \mathit{correct} \\ -\zeta c\left(w_{ri}^{(k)} - \mathrm{m}_{ri}^{(n)}\left(\Sigma_{ri}^{(n)}\right)^{-1}\right) & n \in \mathit{competing} \end{cases} \end{aligned} \tag{70}$$

where m_{ri} and Σ_{ri}^{-1} are mean vector and covariance matrix of the prior distribution for regression row vector w_{ri}. We can see that the constraints $\mathrm{m}_{ri}\Sigma_{ri}^{-1}$ from the prior distribution

is added as a *simple linear combination* in MAPLR/SMAPLR. On the other hand, in RMCELR/SRMCELR, two types of constraints, $\mathrm{m}_{ri}^{(m)}\left(\Sigma_{ri}^{(m)}\right)^{-1}$ and $\mathrm{m}_{ri}^{(n)}\left(\Sigma_{ri}^{(n)}\right)^{-1}$, from the correct and competing labels, respectively, are first *subtracted* from the MCELR solution $w_{ri}^{(k)}$. The differences are then *weighted* by ζc and the weighted differences are finally incorporated into the MCELR estimation as a form of a *weighted linear combination*. Separately from the objective criterion difference, this sophisticated utilization of the prior information in RMCELR/SRMCELR establishes the clear superiority against MAPLR/SMAPLR, and thus directly leads to better adaptation and generalization capability for rapid adaptation.

4.3.3 An Overall Comparison on Rapid Adaptation Experiments

Rapid adaptation experiments are conducted on the same database as shown in Section 4.2.4. Moreover, all experimental setups are identical with the section. Experiments reported in this section can be viewed as the additional comparison study including SMAPLR and SRMCELR. An overall comparison between the various DLT-based adaptation methods studied in the previous section can be also investigated in detail. We will present the results by MLLR, MCELR, MAPLR, SMAPLR, RMCELR, and SRMCELR, respectively.

As discussed, SMAPLR and SRMCELR follow the update rule of MAPLR and RMCELR, respectively. However, they have the different prior selection and estimation scheme. As shown in section 4.3.2 in detail, the structured priors in SMAPLR and SRMCELR are hierarchically derived in the tree and evolved based on the adaptation data available. In particular, SRMCELR performs the iterative prior estimation where the hyper-parameters are re-estimated at every iteration. In addition, the model structure and initial parameter setups are also the same as in Section 4.2.4.

A goal of the experiments in this section is to investigate the adaptation capability and the generalization effect of SMAPLR and SRMCELR compared to other DLT-based methods when the amount of adaptation data is extremely limited (less than 10 seconds of speech). A direct comparison between SMAPLR and SRMCELR can be drawn as done between MAPLR and RMCELR in Section 4.2.4.

Table 15: Rapid Adaptation Performance Comparison in Phone Accuracy Rate (%) on Various Adaptation Methods

Adaptation utterances (seconds)	2 (4.32s)	4 (8.46s)
MLLR	42.46	44.13
MCELR	42.53	44.24
MAPLR	42.91	44.57
SMAPLR	43.60	45.09
RMCELR	**44.95**	**45.71**
SRMCELR	**45.30**	**46.03**

The results by various DLT-based adaptation methods for a rapid adaptation task are summarized in Table 15. Also, Figure 15 depicts a graphical comparison on the methods for the proposed experimental setup. Adaptation results except for SMAPLR and SRMCELR are adopted from Section 4.2.4. We can still maintain the experimental observations and conclusions in Section 4.2.4.

Additionally, first we can see that SMAPLR which utilizes the structured prior estimation outperforms MAPLR with the limited approach in estimating the prior distribution. It is obvious that the structural and hierarchical prior estimation can provide a better use of the adaptation data. The enhanced prior estimation approach directly leads to a better adaptation performance by efficiently constraining the transform parameters for rapid adaptation.

However, SMAPLR yielded a smaller amount of improvement compared to the proposed RMCELR and SRMCELR methods. It can be explained that the MAP criterion is less effective than the RMCE criterion for rapid adaptation. As discussed, the RMCE criterion takes advantage of both the MCE and MAP criteria, by incorporating the prior distribution as a regularization term into the MCE criterion. Moreover, more sophisticated utilization of the prior information compared in Eq. (70) directly makes the clear superiority against SMAPLR.

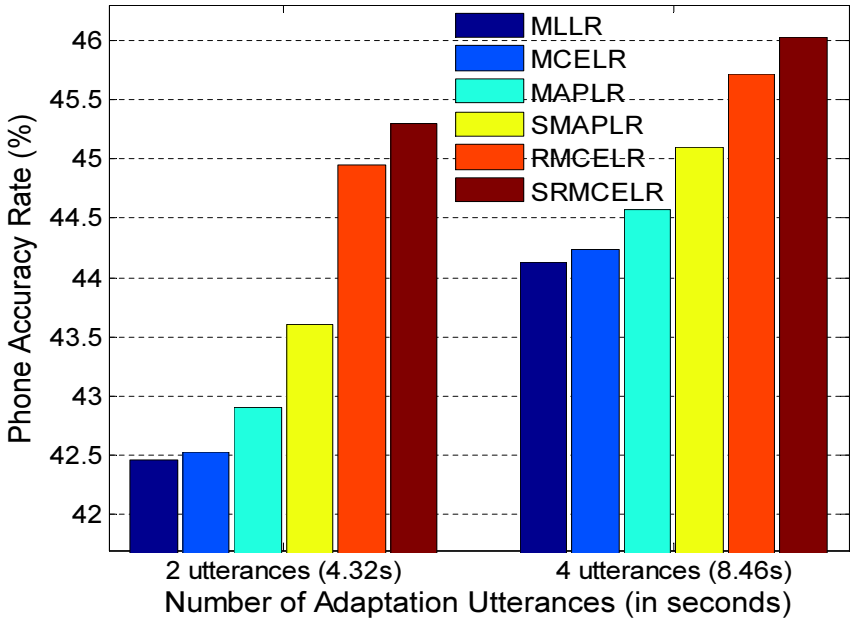


Figure 15: A Graphical Comparison on Various Adaptation Methods for Rapid Adaptation Experiments

In particular, SRMCELR is better than RMCELR for a rapid adaptation task. For the extremest setup (2 utterances/4.32 seconds of adaptation speech available), a PAR of 45.30% was obtained by SRMCELR while RMCELR yielded a PAR of 44.95% in the same setup. The performance improvement by SRMCELR mainly comes from the use of the structured priors and their evolutive estimation scheme described in Section 4.3.1. In the end, SRMCELR outperforms all other DLT-based adaptation methods for rapid adaptation because of the superior objective criterion, RMCE, and the structured framework to the prior density estimation.

Finally, we provide a cross validation study of MLLR, SMAPLR and SRMCELR so as to statistically demonstrate the effectiveness of the proposed SRMCELR adaption method over SMAPLR which has shown the most successful performance among the current linear transform-based adaptation methods. Specifically, we chose a single male speaker (mtcs0) in the noisy TIMIT database for a k-fold cross validation. In the TIMIT dataset, each speaker read a different set of 10 sentences. We excluded two dialect sentences (the SA sentences) which are meant to expose the dialectal variants of the speakers.

Table 16: Adaptation Performance (in PAR %) of MLLR, SMAPLR, and SRMCELR using a Four-fold Cross Validation Procedure

CV-subset # (seconds)	CV-subset 1 (5.46s)	CV-subset 2 (5.10s)	CV-subset 3 (6.91s)	CV-subset 4 (6.29s)
MLLR	50.22	47.52	46.32	50.00
SMAPLR	51.98	50.83	49.78	51.46
SRMCELR	55.95	53.72	51.52	57.28

Upon the eight utterances for the chosen male speaker, a four-fold cross validation technique is used to evaluate the rapid adaptation performance of MLLR, SMAPLR, and SRMCELR. Each validation fold consists of two utterances. Then, four adaptation experiments are conducted, with one of the folds used for adaptation and the remaining three folds for testing. To ensure fairness, the adaptation and the testing datasets are the same for all three adaptation methods. Other experimental setups such as the baseline model and hyper-parameter initialization are identical to the setup described in the previous adaptation experiment.

In this experiment, the phone accuracy rates (PARs) of MLLR, SMAPLR, and SRMCELR are evaluated using the four-fold cross validation procedure. The PARs of these three adaptation methods using each cross validation subset are tabulated in Table 16. As we can see, SMAPLR leads to slightly better performance than MLLR when using the validation subset 1 and 4 while much notable improvements are obtained by SMAPLR using the validation subset 2 and 3. On the other hand, the proposed SRMCELR method significantly outperforms both MLLR and SMAPLR on all validation subsets. In particular, an average absolute gain of 6.11% is achieved by SRMCELR over MLLR while SMAPLR yields an average absolute gain of 2.49% over MLLR. It is demonstrated that SRMCELR produces a consistent and significant performance enhancement on any given validation subset. As a result, these cross validation experimental results lead us to claim that the proposed SRMCELR method has more robust and effective adaptation capability than SMAPLR for rapid

Table 17: A Statistical Significance Testing using a p-value on the Cross Validation Experimental Results

	p-value
SMAPLR over MLLR	0.05743
SRMCELR over MLLR	0.00848

adaptation.

Furthermore, we conduct a statistical significance test on the cross validation experimental results in Table 16. In many statistical tests, a p-value is commonly used as the probability of the difference in a dataset being due to sampling error. It is well known that if a p-value is less than the pre-determined significance level which is often 0.05, then one can determine that a test statistic is statistically significant.

Table 17 shows the p-values of SMAPLR and SRMCELR over MLLR. A p-value of 0.05743 is obtained by SMAPLR over MLLR and exceeds 0.05 which is normally determined as a significance level. Hence, the results of SMAPLR are not statistically significant over MLLR in this cross validation study. On the other hand, a p-value of 0.00848 is obtained by SRMCELR over MLLR. This p-value is extremely lower than a typical significance level (0.05) and the p-value of SMAPLR. From this statistical significance test, we can see that SRMCELR is statistically much significant than SMAPLR over MLLR.

4.4 Chapter Summary

In this chapter, we proposed several novel discriminative linear transform (DLT) based adaptation methods using MVE and MCE criteria for speech detection and recognition. First, we proposed the MVE linear regression (MVELR) adaptation method which estimates a set of DLTs within the MVE criterion. The proposed MVELR method directly minimizes the total verification error with the given adaptation data and thus yields better estimations to the linear transforms compared to the conventional ML-based adaptation

approach (MLLR). Experimental results confirmed that the proposed MVELR method significantly reduces the total error rate over all three phonetic categories of the detectors compared to MLLR.

Furthermore, the limitations of the DLTs for rapid adaptation were addressed and the regularized MCE (RMCE) criterion was proposed to effectively utilize the MCE-based DLTs for rapid adaptation. The RMCE criterion was formulated by introducing the *a priori* distribution as a regularization term to the original MCE empirical risk. This RMCE criterion was applied to estimating the DLTs and the RMCE linear regression (RMCELR) adaptation method was proposed for rapid adaptation. In addition, structural RMCELR (SRMCELR), in which the prior densities for the transform matrices are hierarchically structured in a tree according to the amount of the adaptation data available, was proposed. The proposed RMCELR and SRMCELR adaptation methods take advantage of both the Bayesian perspective and DLT-based adaptation. Therefore, more robust estimates and improved generalization capability can be secured for rapid adaptation. Extensive rapid adaptation experiments were carried out to validate the effectiveness of the proposed methods. The experimental results revealed that the proposed RMCELR and SRMCELR methods significantly outperform all current linear transform-based adaptation methods in a rapid adaptation scenario.

CHAPTER 5

CONSTRAINED DISCRIMINATIVE TRAINING FOR RECEIVER OPERATING CHARACTERISTIC OPTIMIZATION

5.1 Motivation

Many machine learning algorithms have been developed to achieve the optimal performance of various classification and verification applications. One critical issue is how to optimize a model to meet a pre-set performance objective. In many real-world applications, the model performance can be summarized by the false rejection (FR) and false alarm (FA) rates. Minimizing FRR and FAR has been thus chosen as an *overall objective* in a wide variety of machine learning methods [104, 37, 105, 36].

As recent pattern recognition applications become diverse, it has been increasingly recognized that many realistic applications often require an optimal solution that meets a particular operating target. In applications that involve detection, this particular operating target may be a specific FRR (or FAR), say 0.1%, and the design objective is to optimize model parameters to minimize FAR (or FRR). For example, an automatic teller machine (ATM) with voice authentication may demand a very low FAR because a banking business needs to avoid excessive inconveniences imposed on its clients for fear of driving the clients away. In contrast, a language learning application may allow a relatively high FAR, but requires a very low FRR because a good student may reject the learning aid when he or she is unjustly graded by the system. Therefore, it is desirable that a novel system design methodology be developed to allow model optimization at any given operating point, ultimately forming a new ROC curve, at every point of which there exist a pair of models, the target and the alternative, that achieve an optimal performance at that specific operating point.

Figure 16 shows two receiver operating characteristic (ROC) curves achieved by two sets of system models, respectively. As shown in the figure, the two ROC curves have equal

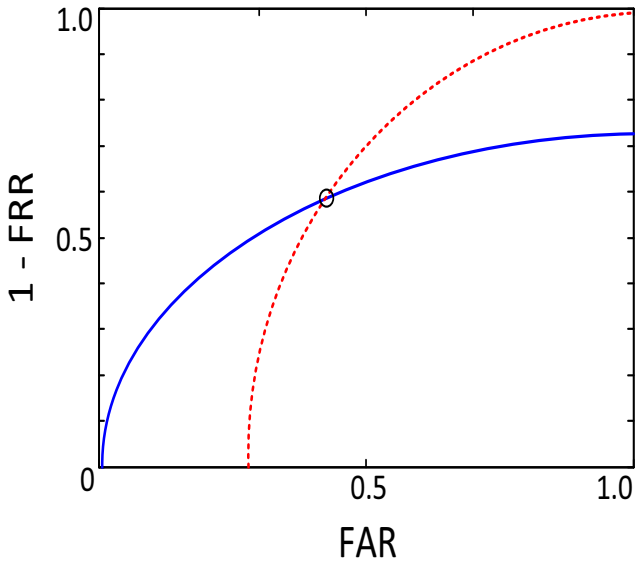


Figure 16: Two ROC Curves with Same AUC and EER Values

areas under the ROC curve (AUC) and equal error rate (EER) values. Without referencing to a particular operating point, these two systems may be considered equally valuable. However, if a low FAR is required at a fixed 30% FRR point, the system with a red dotted ROC curve is much preferred. In a reversed condition, it is clear that the system with a blue ROC curve is much more valuable. As a consequence, it is desirable to formulate a training method which can directly optimize a system for such particular operating needs.

In this chapter, we will propose a *constrained optimization formulation* of minimum verification error (MVE) [63, 64, 106] training for particular operating characteristic optimization. Suppose there are two conflicting objectives such as FAR and FRR as mentioned above. One goal is to construct a constrained objective function in which one objective, e.g., FAR, is minimized with a specified constraint on another *target* objective, e.g., FRR at 1%. It is necessary to provide an analytical solution in formulating the constrained objective function and the corresponding optimization procedure. In this research, we propose to apply *augmented Lagrange multiplier* (ALM) [107, 108, 109] into the MVE framework for a constrained objective function formulation. In the following sections, a derivation of the constrained MVE objective function and a systematic learning/optimization procedure

are described in detail.

5.2 Receiver Operating Characteristic (ROC) Definition and Optimization

The receiver operating characteristic (ROC) curve [36, 110] has been widely used in evaluating the reliability of diverse pattern recognition applications such as text categorization [111, 112] and speaker verification/identification [31, 32, 113]. In binary classification, a traditional ROC curve is plotted by varying a threshold on positive and negative sample scores. This is thus a plot of a false positive rate against a true positive rate to all possible operating points without changing the system model parameters. The ROC metric generally provides more useful information than a single error measurement. Among a set of different models, it allows one to evaluate a performance tradeoff of a certain model with respect to a wide variety of operating conditions so as to select the model that better suits a particular operating need.

A detection error tradeoff (DET) curve [114, 115] as a variant of the ROC curve has been found useful in speech applications. The DET curve plots a false negative rate on the Y axis instead of a true positive rate in ROC. Another key difference is that DET graphs are log scaled on both axes. Therefore, the area of the lower left part of the curve is expanded and a curve shape is almost linear. The DET curve makes it easier to read the lower left part and determine a tradeoff between FAR and FRR than the ROC curve when well-performing models are evaluated.

The area under the ROC curve (AUC) [104, 37] is also an important factor as a measure of model performance. A single AUC value can be found by calculating the area under the ROC curve. It is closely related to the quantity of ranking performance defined as the probability of the positive samples ranked higher than the negative samples. The AUC value is thus equivalent to Wilcoxon-Mann-Whitney statistic [116]. The AUC metric has been mostly used as a ranking performance measure [117, 118, 119].

In the machine learning literature, there has been a lot of effort in optimizing ROC and AUC for diverse applications. In [120], the AUC metric is directly used as an objective function to be optimized for the classifier learning in the area of information retrieval. Herschtal and Raskutti proposed a RankOpt algorithm [105] to optimize a linear binary classifier by using the AUC rank statistic as an objective function and gradient descent-based optimization. Similarly, ROC–AUC optimization methods have also been proposed for neural networks [121] and SVMs [122, 38]. All these methods aim at a perfect ranking on the sample instances, e.g., AUC=1, and thus directly lead to maximizing the AUC value. However, as discussed already, two ROC curves with the same AUC value can be very different at certain operating points.

Several studies have shown that the ROC–AUC criterion is not rich and flexible enough in optimizing a model for diverse operating needs. To address this issue, multi-objective optimization (MOO) [123, 124, 109] in which many different objectives can be optimized simultaneously has been investigated in machine learning communities. The earliest work was founded on the Pareto optimality [124]. Some of successful methods to solve the MOO problem include multi-objective evolutionary algorithms (MOEA) [125, 126, 127], goal programming (GP) [128, 129], and a classifier combination approach [130, 131]. These methods have achieved great success in a wide range of applications, for example, text categorization [132, 133] and biometric systems [134, 135]. However, most of the MOO-based methods aim at balancing conflicting objectives or an error tradeoff and thus can be considered as an ensemble solution.

In the ASR literature, iterative constrained optimization (ICO) based on the MOO framework was proposed in [69] for finding compromise solutions that are satisfactory for each of multiple competing performance criteria. The proposed ICO approach was applied to an automatic language identification (LID) task and resulted in a good balance among the many competing objectives. However, it is still viewed as an ensemble solution, which is not directly related to a particular operating need.

In [136], a constrained optimization problem for a specific operating point was addressed in the context of utterance verification. The minimum verification error rate constrained optimization (MVER-CO) training was proposed by utilizing a penalty function approach in the constrained optimization literature [107, 123]. However, the penalty parameter which determines the quality and convergence of constrained optimization was not tuned during training. Furthermore, decision threshold variations at a given particular operating point over learning were not taken into account as well. It is still necessary to design a direct and effective solution for particular operating point optimization. In this thesis, we make use of the augmented Lagrange multiplier (ALM) framework and empirical threshold variation over the minimum verification error (MVE) criterion.

5.3 Limitations of MVE Criterion for Particular Operating Point Optimization

As discussed in Section 3.2.3.1, MVE training was proposed to optimize detectors or verifiers by directly minimizing the empirical detection/verification error given labeled training data. The empirical error is approximated by a loss function which is continuous and differentiable function, for example, a sigmoid function. The smoothed MVE loss constitutes an unconstrained objective function, typically using a sum of the two types (I and II) of errors as the default optimization objective.

Although the empirical error may be measured at diverse operating points to meet some particular design requirements, it is normally assigned as the number of FR and FA given labeled transcription without taking into accounts any operating points. Therefore, it can be viewed as the error counting function depending on the number of positive and negative samples in the training data. In real-world applications, the dataset for learning is often imbalanced where the number of observations belonging to each of positive and negative classes is different. In this case, the majority class, e.g., negative class, is dominantly trained and the class-dependent error type, here FAR, is thus further reduced. The current

MVE framework constructs an objective function as a combination of errors regardless of the between-class imbalance problem.

According to the MVE objective function, MVE results directly in reducing the total number of errors. For this reason, the minimum total error rate (MTER) [59, 64] has been used as a direct measure in evaluating a model trained by the MVE method. However, in designing and evaluating a model for a particular operating point, AUC, EER, and the whole ROC curve as well as MTER are not directly related to a point-wise performance metric. Therefore, to enable the particular operating point optimization in the current MVE framework, the MVE objective function has to be redesigned by embedding the specific operating need into the empirical error estimates.

In addition, a mis-verification measure in the MVE framework assumes that the decision threshold is always zero or pre-set as a median value heuristically found from a development set. It is hard to tune the decision threshold over learning. If one inexplicitly adjusts the decision threshold, some important samples may be regarded as outliers and thus would not be involved in learning. When the learning objective is targeted at a particular operating point, the decision threshold should shift onto the operating point and assume an important role in model optimization. Then, those samples near the adjusted threshold can be appropriately treated and the discrimination of the samples would be intensively enhanced. This gives rise to the new concept of ROC optimization, aiming at obtaining optimized models at every point on the optimized ROC curve.

To address this need, we propose to apply augmented Lagrange multiplier (ALM) into the MVE framework and utilize threshold variation at the particular operating point over learning. The ALM framework makes the MVE objective function flexible enough to embed a particular operating need into the empirical error estimates. Furthermore, the ALM parameters associated with the empirical error variation efficiently handle the constrained MVE optimization over iterations. Meanwhile, the required decision threshold at an operating point is searched at every iteration and directly used in a mis-verification measure so

as to shift a critical decision boundary onto the point to be optimized.

5.4 Constrained Scenarios in Speech Detection and Verification

In speech detection and verification, a hypothesis H_0 with speech segment X_k is claimed to have come from a target model λ_t. To verify the claim, a model λ_a for the alternative hypothesis H_a is used to generate the log-likelihood ratio (*LLR*) as follows:

$$LLR(X_k|H_0, \lambda_t, H_a, \lambda_a) = \frac{1}{f_k} \log \left[\frac{P(X_k|H_0, \lambda_t)}{P(X_k|H_a, \lambda_a)} \right] \gtrless \theta_k \text{ ; Accept or Reject.} \tag{71}$$

where f_k is the total number of frames in speech segment X_k and a decision is made by comparing *LLR* with a threshold θ_k. Let the whole training set X be partitioned into two parts according to correct labels:

$$X = \{X_{pos}, X_{neg}\} = \{x_m^+, x_n^- \in R^D | 1 < m < M, 1 < n < N\}. \tag{72}$$

Then, a set of *LLR* scores, S^+ and S^- for positive and negative tokens x_m^+ and x_n^-, can be found by Eq. (71):

$$S = \{S_i^+, S_j^- \in R | 1 < i < M, 1 < j < N\}. \tag{73}$$

Finally, a given threshold θ_k determines a *FAR* and *FRR* for a particular operating point at θ_k as follows:

$$FRR_k = \frac{1}{M} \sum_{i=1}^{M} I(S_i^+ < \theta_k) \tag{74}$$

$$FAR_k = \frac{1}{N} \sum_{j=1}^{N} I(S_j^- > \theta_k). \tag{75}$$

where $I(\cdot)$ is an indicator function.

As discussed, there are two types of constrained scenarios commonly adopted in detection/verification as follows:

$$\min FRR \quad \text{subject to} \quad FAR = \alpha, \quad 0 \leq \alpha \leq 1 \tag{76}$$

$$\min FAR \quad \text{subject to} \quad FRR = \beta, \quad 0 \leq \beta \leq 1. \tag{77}$$

The expressions above are defined at a particular operating condition, either α or β, which can be set to any value between 0 and 1. There exist a wide variety of potential operating conditions depending upon an application specification or requirement. Obviously, we can sample every operating conditions with their corresponding thresholds as follows:

$$\{FAR_i,\ \theta_{FAR_i} | 0 \leq i \leq \infty,\ 0 \leq FAR_i \leq 1\} \ \rightarrow\ \{FAR_i = \alpha_i,\ \theta_{\alpha_i}\} \tag{78}$$

$$\{FRR_j,\ \theta_{FRR_j} | 0 \leq j \leq \infty,\ 0 \leq FRR_j \leq 1\} \ \rightarrow\ \{FRR_j = \beta_j,\ \theta_{\beta_j}\}. \tag{79}$$

As can be seen, if the number of conditions, I and J, is large enough (∞), we can sample every operating points representing the entire ROC behavior. As a result, we can have a set of information for every operating points when I and J are set to be large enough. If these objectives are incorporated into model training, basically we will have $I + J$ sets of models. It would directly lead to huge computational complexity and memory capacity. Thus, an appropriate number of $I + J$ should be carefully chosen for tradeoff between optimal performance and computation. Finally, by combining or fusing a set of information from the $I + J$ sets of models, we may form a new ROC curve optimized at every operating point, of which there exist a pair of models. This new type of ROC formation is out of scope of this thesis. Furthermore, in order to obtain the new ROC curve, a system design methodology that allows model optimization at any given operating point has to be taken into account first. In this thesis, we focus on optimizing a system model for a particular operating condition.

For a particular operating point with the corresponding decision threshold, Eqs. (76) and (77) can be expressed as

$$\min FRR \quad \text{subject to} \quad FAR = \alpha_i \ \text{ at } \ \theta_{\alpha_i} \tag{80}$$

$$\min FAR \quad \text{subject to} \quad FRR = \beta_j \ \text{ at } \ \theta_{\beta_j} \tag{81}$$

where α_i and β_j are particular operating points with the corresponding thresholds θ_{α_i} and θ_{β_j}, respectively. The key challenge here is how to translate an operating point requirement

(e.g., FRR=0.01) to the corresponding threshold for testing and further to the error objective function, which properly combines the two types of errors and is the target to be minimized.

Unfortunately, the classical MVE training cannot explicitly cope with the scenarios such as Eqs. (80) and (81) because of the constraints and non-linearity. To embed these constrained scenarios into an MVE training framework, a theory of constrained and non-linear optimization should be considered. Furthermore, a systematic learning procedure with constrained objective functions have to be investigated thoroughly. We will review the fundamental principles of the constrained optimization in the next section.

5.5 Constrained Optimization Techniques

The standard structure of most constrained optimization problems [107, 108, 137] is essentially contained in the following:

$$\begin{aligned} &\text{minimize} && f(\lambda) \\ &\text{subject to} && g_i(\lambda) = 0, \quad i = 1, \ldots, p \end{aligned} \tag{82}$$

where λ has dimensions $n \times 1$, $\lambda \in \mathbb{R}^n$, $f(\lambda)$ is the objective function to be minimized, and $g(\lambda)$ are a set of equality constraints. The most simple and straightforward approach to handling constrained problems of the above form is to apply a suitable unconstrained optimization algorithm.

5.5.1 Penalty Function Approach

Historically, the earliest development to solve the constrained optimization problem, Eq. (82), is a *sequential minimization method* based on the use of *penalty* or *barrier* functions. This is referred to as a *sequential penalty function technique*. For the equality problem such as Eq. (82), the quadratic penalty function is defined as

$$\phi(\lambda, \rho) = f(\lambda) + \frac{\rho}{2} \sum_{i}^{P} \{g_i(\lambda)\}^2 \tag{83}$$

where ρ is the penalty parameter, $\rho \gg 0$. The penalty is formed from a sum of squares of constraint violations and the parameter ρ determines the amount of the penalty. By making this penalty parameter larger, we penalize constraint violations more severely, thereby forcing the minimizer of the penalty function closer to the feasible region for the constrained problem. For example, we can choose a fixed sequence $\left\{\rho^{(k)}\right\} \to \infty$, typically $\left\{1, 10, 10^2, 10^3, \ldots\right\}$ and then find a local minimizer for each $\rho^{(k)}$ [137, 108].

5.5.2 Lagrange Multiplier

In mathematical optimization, Lagrange multiplier provides a strategy for finding the local minima of a function subject to equality constraints such as Eq. (82). We introduce a new variable c called a Lagrange multiplier [107] and study the Lagrange function defined by

$$\phi(\lambda, c) = f(\lambda) - \sum_{i}^{P} c_i g_i(\lambda) \tag{84}$$

In general, we can set the partial derivatives to zero to find the minimum:

$$\nabla_\lambda \phi(\lambda^*, c^*) = 0 \tag{85}$$

$$\nabla_c \phi(\lambda^*, c^*) = 0 \tag{86}$$

where λ^* is the minimum solution and c^* is the set of associated Lagrange multiplier. This means that $\nabla_\lambda f$ and $\nabla_\lambda g$ must be parallel [107, 108]. That is, there exists some $c \in \mathbb{R}$ such that

$$\nabla_\lambda f - c\nabla_\lambda g = 0 \quad \to \quad \nabla_\lambda f = c\nabla_\lambda g. \tag{87}$$

5.5.3 Augmented Lagrange Multiplier

The penalty function and Lagrange multiplier methods suffer from some computational disadvantages and are not entirely efficient. The augmented Lagrange multiplier (ALM) method combines the classical Lagrange method with the penalty function approach. The ALM [107, 108] for the equality constrained problem defined in Eq. (82) is introduced as

$$\phi(\lambda, c, \rho) = f(\lambda) - \sum_{i}^{P} c_i g_i(\lambda) + \frac{\rho}{2} \sum_{i}^{P} \{g_i(\lambda)\}^2 \tag{88}$$

where c_i are the Lagrange multipliers and ρ is the adjustable penalty parameter. If all the multipliers c_i are chosen to be identically zero, this becomes the usual penalty function approach as described in Section 5.5.1. On the other hand, if all stationary values c_i^* are available, then it can be shown [108] that for any positive value of ρ, the minimization of $\phi(\lambda, c, \rho)$ with respect to λ gives the solution λ^* to problem of Eq. (82).

We now design an algorithm that fixes the penalty parameter ρ to some value $\rho_k > 0$ at its k-th iteration, fixes c at the current estimate c^k, and performs minimization with respect to λ. Using λ_k to denote the approximate minimizer of $\phi(\lambda, c, \rho)$, we have by the optimality conditions for unconstrained minimization [137] that

$$0 \approx \nabla_\lambda \phi(\lambda, c, \rho) = \nabla f(\lambda) - \sum_i^P \left[c_i^k - \rho_k g_i(\lambda)\right] \nabla g_i(\lambda) \tag{89}$$

By comparing with the optimality condition Eq. (87) for Eq. (82), we can set

$$c_i^{k+1} = c_i^k - \rho_k g_i(\lambda) \tag{90}$$

With this setup and sufficiently large ρ, the minimizer λ^* can be iteratively searched.

5.6 Constrained MVE Training using Augmented Lagrange Multiplier

5.6.1 Preliminaries

Before we derive a constrained MVE formulation, we will re-interpret the constrained scenarios discussed in Section 2 through a MVE training perspective.

First, the *FRR* and *FAR* shown in Eqs (74-75) can be written in a MVE framework as follows:

$$FRR_k = \frac{1}{M}\sum_{i=1}^{M} I(S_i^+ < \theta_k) \approx \frac{1}{M}\sum_{i=1}^{M} \{1 - \ell(S_i^+, \theta_k)\} = \frac{1}{M}\sum_{i=1}^{M} \frac{1}{1 + \exp\left(-\gamma d_I(S_i^+, \theta_k)\right)} \tag{91}$$

$$FAR_k = \frac{1}{N}\sum_{j=1}^{N} I(S_j^- > \theta_k) \approx \frac{1}{N}\sum_{j=1}^{N} \ell(S_j^-, \theta_k) = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{1 + \exp\left(-\gamma d_{II}(S_j^-, \theta_k)\right)} \tag{92}$$

where $I(\cdot)$ is an indicator function, $\ell(\cdot)$ is a sigmoid function, and $d(\cdot)$ is a mis-verification

measure function. $d_I(S_i^+, \theta_k)$ and $d_{II}(S_j^-, \theta_k)$ are two different types of mis-verification measures for Type I (FR) and Type II (FA), respectively. They are defined by

$$d_I(S_i^+, \theta_k) = -g_t(X_{pos}|\lambda_t^i) + g_a(X_{pos}|\lambda_a^i) + \theta_k \tag{93}$$

$$d_{II}(S_j^-, \theta_k) = +g_t(X_{neg}|\lambda_t^j) - g_a(X_{neg}|\lambda_a^j) - \theta_k. \tag{94}$$

where g_t and g_a are the normalized log likelihoods, and λ_t and λ_a are the parameter sets of the target model and the anti-model [63, 106] for the given speech segment, respectively.

Now recall one of the constrained scenarios that we are interested:

$$\min FRR \quad \text{subject to} \quad FAR = \alpha_i, \quad \text{at} \ \ \theta_{\alpha_i}. \tag{95}$$

This can be written as the constrained optimization problem form defined in Eq. (82) as follows:

$$\begin{aligned} \text{minimize} \quad & f(\lambda) = FRR(\theta_{\alpha_i}) \\ \text{subject to} \quad & g_i(\lambda) = FAR(\theta_{\alpha_i}) - \alpha_i = 0. \end{aligned} \tag{96}$$

Similarly, we can define the constrained form for Eq. (81):

$$\begin{aligned} \text{minimize} \quad & f(\lambda) = FAR(\theta_{\beta_j}) \\ \text{subject to} \quad & g_j(\lambda) = FRR(\theta_{\beta_j}) - \beta_j = 0. \end{aligned} \tag{97}$$

Above two equations are our target problem definitions that will be embedded into a constrained MVE objective function.

5.6.2 Constrained MVE Objective Function

As shown in [63, 106, 64], the classical MVE objective function is defined by

$$\begin{aligned} L(X|\lambda) &= \frac{1}{M}\sum_{m=1}^{M} \ell(d_I(X_m^+|\lambda)) + \frac{1}{N}\sum_{n=1}^{N} \ell(d_{II}(X_n^-|\lambda)) \\ &= FRR(\theta) + FAR(\theta) \qquad ; \quad \theta = 0 \end{aligned} \tag{98}$$

From the above definition, it is clear that the original MVE does not assume any threshold preference and does assume equal importance (no constraint) between the *FAR* and *FRR*. Thus, it can be viewed as a simple combination (*total*) of empirical error of the *FAR* and *FRR* without threshold concerns.

Suppose that given a constraint α_i with respect to *FAR* we are going to minimize *FRR* as described in Eq. (96). This objective can be incorporated into MVE by using the penalty function method:

$$L(X|\lambda,\rho) = FRR(\theta_{\alpha_i}) + \frac{\rho}{2}\left\{FAR(\theta_{\alpha_i}) - \alpha_i\right\}^2. \tag{99}$$

Similarly, the Lagrange multiplier method introduced in Section 3.2. can be used to constrain $FAR = \alpha_i$ and the constrained MVE objective function using the Lagrange multiplier method is defined by:

$$L(X|\lambda,c) = FRR(\theta_{\alpha_i}) - c\left\{FAR(\theta_{\alpha_i}) - \alpha_i\right\}. \tag{100}$$

Finally, the augmented Lagrange multiplier (ALM) method can assign the constrained MVE objective function as follows:

$$L(X|\lambda,c,\rho) = FRR(\theta_{\alpha_i}) - c\left\{FAR(\theta_{\alpha_i}) - \alpha_i\right\} + \frac{\rho}{2}\left\{FAR(\theta_{\alpha_i}) - \alpha_i\right\}^2. \tag{101}$$

In the same ALM structure, we can define the constrained MVE objective function for the scenario described in Eq. (97):

$$L(X|\lambda,c,\rho) = FAR(\theta_{\beta j_i}) - c\left\{FRR(\theta_{\beta_j}) - \beta_j\right\} + \frac{\rho}{2}\left\{FRR(\theta_{\beta_j}) - \beta_j\right\}^2. \tag{102}$$

We have defined three constrained MVE objective functions: $L(X|\lambda,\rho)$, $L(X|\lambda,c)$, and $L(X|\lambda,c,\rho)$. Among them, we will adopt the form $L(X|\lambda,c,\rho)$ which is defined by the ALM method because of its effective optimality property.

5.6.3 Derivation of the Training Procedure

The constrained MVE objective function, either Eq. (101) or (102), can be minimized by the generalized probabilistic descent (GPD) algorithm as the classical MVE adopts:

$$\lambda_{k+1} = \lambda_k - \epsilon_k \nabla L(X|\lambda,c,\rho)\Big|_{\lambda=\lambda_k} \tag{103}$$

where ϵ_k is the learning rate, and k is the cumulative number of the processed training samples. A derivative of the objective function $L(X|\lambda, c, \rho)$ can be written as

$$\nabla_\lambda L(X|\lambda, c, \rho) = \nabla_\lambda FRR(\theta_{\alpha_i}) - \left[c - \rho\left\{FAR(\theta_{\alpha_i}) - \alpha_i\right\}\right] \nabla_\lambda FAR(\theta_{\alpha_i}). \tag{104}$$

$\nabla_\lambda FRR(\theta_{\alpha_i})$ and $\nabla_\lambda FAR(\theta_{\alpha_i})$ can be expressed as

$$\nabla_\lambda FRR(\theta_{\alpha_i}) = \frac{1}{M}\sum_{m=1}^{M} \gamma\ell(d_I)\left\{1 - \ell(d_I)\right\} \frac{\partial}{\partial\lambda}\left\{d_I(S_m^+, \lambda^m, \theta_k)\right\} \tag{105}$$

$$\nabla_\lambda FAR(\theta_{\alpha_i}) = \frac{1}{N}\sum_{n=1}^{N} \gamma\ell(d_{II})\left\{1 - \ell(d_{II})\right\} \frac{\partial}{\partial\lambda}\left\{d_{II}(S_n^-, \lambda^n, \theta_k)\right\} \tag{106}$$

where $d_I(\cdot)$ and $d_{II}(\cdot)$ are mis-verification measure functions defined in Eqs. (93-94). As discussed in Section 3.3., the optimality condition in Eq. (90) suggests the following update rule for Lagrange multiplier c at iteration l:

$$c_{l+1} = c_l - \rho_l\left\{FAR(\theta_{\alpha_{i(l)}}) - \alpha_i\right\}. \tag{107}$$

Meanwhile, ρ_l is fixed at its current iteration l and then increased at the next iteration $l + 1$ such as

$$0 < \rho_l \leq \rho_{l+1} \qquad \forall l \tag{108}$$

where $\left\{\rho^{(l)}\right\} > 0$.

In this research, to reasonably adjust the penalty parameters, we increase ρ_l by multiplication with a factor $\eta > 1$ only if the constraint violation as measured by $|FAR(\theta_{\alpha_i}) - \alpha_i|$ is not decreased by a factor $\xi < 1$ over the previous minimization. For example,

$$\rho_{l+1} = \begin{cases} \eta\rho_l & \text{if } |FAR(\lambda_l, \theta_{\alpha_{i(l)}}) - \alpha_i| > \xi|FAR(\lambda_{l-1}, \theta_{\alpha_{i(l-1)}}) - \alpha_i|, \\ \rho_l & \text{else.} \end{cases} \tag{109}$$

In our implementation, we set $\eta = 10$ and $\xi = \frac{1}{4}$ as typically recommended [107, 108].

The proposed training procedure is iteratively performed when the following stopping conditions are not met:

$$L(X|\lambda_l, c_l, \rho_l) \leq L(X|\lambda_{l+1}, c_{l+1}, \rho_{l+1}) \qquad \forall x^+, x^- \in X \tag{110}$$

$$\|\nabla_{\lambda} L(X|\lambda, c, \rho)\| \leq \delta \tag{111}$$

where δ is a tolerance level, generally set as very small value. It means that the proposed iterative minimization method is terminated when the constrained objective function value is not decreased in comparison with the previous iteration and the gradient of the objective function is sufficiently small, but not necessarily zero. In summary, an algorithm description of constrained MVE for particular operating point optimization is presented in Table 18.

5.7 Experiments

In order to study the impact of the proposed method for particular operating point optimization in the ROC space, we have conducted a series of experiments on the standard TIMIT database [70]. We chose 6-class broad phonetic class (BPC) [1] detection for an evaluation task. This BPC category is based on the articulatory manner [83] and mapped from 48 monophones into 6 articulatory features as shown in Table 9. Among the 6 classes, we exclude the silence class since the baseline ML-trained model on this class already shows very low FAR or FRR at any given constraints. For example, an FAR of 0.3% is observed at a 2% FRR constraint while an FRR of 0.7% is observed at a 2% FAR constraint.

Table 19 gives the numbers of positive and negative segments for each BPC sub-class except the silence class in the TIMIT training dataset. As can be seen from the table, the negative segments are much more dominant than the positive segments in the TIMIT training set. In many real-world applications, we can see a similar class imbalance problem between positive and negative samples.

In all experiments, the input speech is represented by the common 39 dimensional feature vectors with 12MFCC, 12Δ, 12$\Delta\Delta$, and three log-energy values. Based on the mapping rule in Table 9, the target models and anti-models in all sub-class detectors are constructed by three-state strict left-to-right HMMs and 16-component Gaussian-mixture density with diagonal covariance matrices in each HMM state. Given sample observations

Table 18: An Algorithm Description of Constrained Minimum Verification Error Training Given FAR=α_i Constraint

I. Initialization

1. A set of training samples; $x^+,\ x^- \in X$.
2. Calculate *LLR* scores, $S^+,\ S^- \in S$, using a set of initial (ML) model λ_0.
3. Input a particular operating condition α_i.
4. Find the corresponding threshold $\theta_{\alpha_{i(0)}}$.
5. Choose $\gamma = 1.0,\ \epsilon \ll 1.0,\ \eta = 1.5,\ \xi = \frac{1}{4}$ and $\delta = 10^{-3}$.
6. Initialize $c_0 = 0$ and $\rho_0 = 1$.

II. Repeat *L* iterations (e.g., 10 iterations, $l \in [1, 2, \ldots, 10]$)

1. Do constrained MVE learning given K training samples.

   for $k \in 1, 2, \ldots, K$

   Calculate the gradient $\nabla_\lambda L(X_k|\lambda, c, \rho)$

   Update λ by GPD: $\lambda_{k+1} = \lambda_k - \epsilon_k \nabla L(X_k|\lambda, c, \rho)\Big|_{\lambda=\lambda_k}$

   end for

2. Update the *LLR* scores and thresholds by the current iteration model.

   Calculate *LLR* scores, $S^+,\ S^- \in S$, using the updated model λ_l

   Find new $\theta_{\alpha_{i(l)}}$ at the input constraint α_i

3. Update the Lagrange multiplier c and penalty parameter ρ.

   $c_{l+1} = c_l - \rho_l \left\{ FAR(\theta_{\alpha_{i(l)}}) - \alpha_i \right\}$

   if $|FAR(\lambda_l, \theta_{\alpha_{i(l)}}) - \alpha_i| > \xi |FAR(\lambda_{l-1}, \theta_{\alpha_{i(l-1)}}) - \alpha_i|$

   $\rho_{l+1} = \eta \rho_l$

   else

   $\rho_{l+1} = \rho_l$

   end if

4. See if the stopping conditions are satisfied.

   if $L(X|\lambda_{l-1}, c_{l-1}, \rho_{l-1}) \leq L(X|\lambda_l, c_l, \rho_l)$ and $\|\nabla_\lambda L(X|\lambda_l, c_l, \rho_l)\| \leq \delta$

   Stop iteration

   end if

Table 19: Numbers of Positive and Negative Segments for each BPC Sub-class in the TIMIT Training Set.

segments	Positive	Negative
fricatives	20,271	96,601
vowels	46,471	70,401
nasals	12,224	104,648
stops	15,741	101,131
others	13,896	102,976

on the training set as shown in Table 19, all models are first estimated by the conventional maximum likelihood (ML) method using the EM algorithm as a baseline. Then, the ML baseline models are further trained by the MVE method and these MVE-trained models are compared with the proposed constrained MVE (CMVE) method. Same as MVE, CMVE is applied to the ML models and thus we can see a direct comparison between MVE and CMVE over ML.

In MVE training, we follow the conventional setup that constructs an overall objective function as the total number of errors and assumes that thresholds are equals to zero for all classes and every iteration. Unlike the conventional MVE setup, the proposed constrained MVE method constitutes a constrained objective function formulated by the ALM framework as described in Chapter 5.6.2. In this objective function formulation, the FAR and FRR are first approximated as smooth functions of the detectors and then one of the error types is constrained with an input particular operating point. The derivation of the constrained objective function and the update rule over iterations is described in Chapter 5.6.3. In the proposed method, the input operating constraint is required prior to the learning. In addition, associate hyper-parameters such as the GPD parameters and ALM parameters require experimental investigation for proper initialization.

In our experiments, a sigmoid slope parameter γ is fixed at 1 and a GPD learning rate ϵ

has a small value less than 1.0 to avoid the over-training problem. For the ALM parameter initialization, we simply set $c_0 = 0$ and $\rho_0 = 1$ while η and ξ are heuristically adjusted for each sub-class. The threshold values corresponding to the input operating point are updated every iteration using the current iteration model. In the following experiments, all reported results are observed at 6-th iteration for both MVE and CMVE.

In the evaluation, either FAR or FRR at a particular operating constraint is mainly concerned instead of AUC, EER, and MTER. As discussed, in evaluating a model for a particular operating point, overall performance measures such as AUC and EER are not directly related to a point-wise performance metric. Hence, we directly measure a target error rate at a given particular constraint. In this thesis, we choose two different types of constraints, a 2% FRR and a 2% FAR to illustrate the design methodology without loss of generality. A main goal of the experiments is to investigate the tradeoff at a particular operating condition, not on the overall ROC space. We report the performance at a particular operating condition and also provide the DET curves to see the difference between ML, MVE, and CMVE on the entire ROC space.

5.7.1 Minimize FAR at 2% FRR Constraint

The first experiment scenario is to minimize the FAR at a given 2% FRR constraint. First of all, the performance of the two learning algorithms, MVE and CMVE, with increasing number of iterations is provided to demonstrate the effectiveness and convergence property of the proposed CMVE method over MVE. The performance is measured by either the minimum total error rate (MTER) or the FAR at a 2% FRR constraint. Figure 17 shows the MTERs of MVE and CMVE on a fricative class over 10 iterations. The MTER of MVE was consistently decreased as the number of iterations increases. As discussed, since the MVE method aims at minimizing the total number of errors, the MTER can be directly minimized by MVE. Meanwhile, the proposed CMVE method is not directly related to the MTER metric and hence the inconsistent error reduction was observed by CMVE. Furthermore, a relatively small amount of gain in minimizing the MTER was achieved

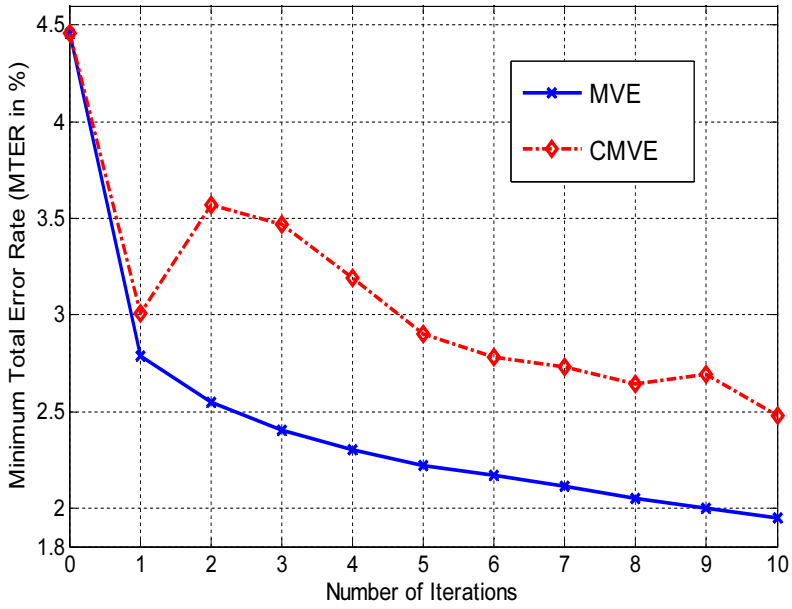


Figure 17: The Minimum Total Error Rates (MTERs in %) of the Traditional MVE Method and the Proposed CMVE Method over 10 Iterations: On a Fricative-Class in Training Set.

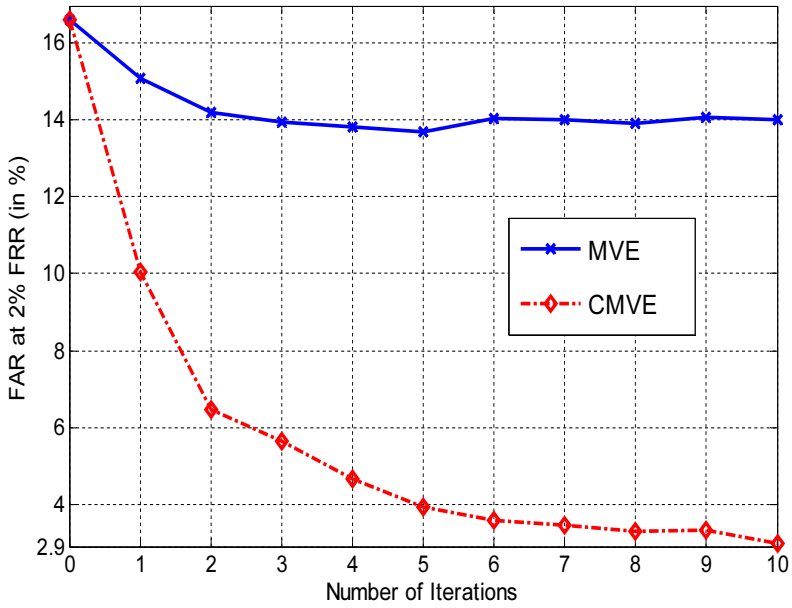


Figure 18: The False Alarm Rates (FARs) of the Traditional MVE Method and the Proposed CMVE Method at a 2% False Rejection Rate (FRR) Point over 10 Iterations.: On a Fricative-Class in Training Set.

by CMVE when compared to MVE.

However, a goal of the experiments in this research is to investigate the tradeoff at a particular operating condition, not on the overall error metric such as the MTER. Figure 18 presents the FARs of MVE and CMVE at a 2% FRR constraint on the same class over 10 iterations. The proposed CMVE method significantly reduced the FAR at a 2% FRR as the iteration proceeds. As the objective of CMVE is to directly minimize a target error rate at a given particular constraint, a considerable amount of FAR reduction at the given 2% FRR constraint was achieved by CMVE. In addition, the FAR of CMVE was consistently decreased over the iterations and converged after 5 iterations. On the contrary, the MVE method yielded very limited FAR reduction because of the inconsistent optimization criterion in this constraint scenario.

The results on each BPC sub-class by ML, MVE, and CMVE are summarized in Table 20. At a 2% FRR point, the FARs of all sub-classes by the ML method represents very high error rate. It is obvious that the ML method cannot optimize the detection performance at a particular operating need. The ML criterion aims at the optimal distribution estimation by maximizing the likelihood of the model parameters with the given set of observations. However, maximizing the likelihood does not guarantee a minimum error rate both in total and at a certain point because of the inconsistent criteria between detector training and performance evaluation.

On the other hand, the FARs of most sub-classes at the 2% FRR constraint are substantially reduced by the MVE method. For example, the FARs of vowels and stops are reduced, from 11.96% to 3.77% and from 14.15% to 5.97%, respectively. However, the FARs of fricatives and others still have very high error rate. From Figure 19, we see that the MVE-trained model shows only moderate performance improvement over the ML model at the 2% FRR point. In contrast, the FRR of the MVE-trained model at a the 2% FAR is significantly reduced over ML. It is because the MVE method mainly handles the negative samples during learning and thus the errors around a low FAR is dominantly reduced. As

Table 20: False Alarm Rate (%) at 2% False Rejection Rate Constraint on Training Set

	ML	MVE	Constrained MVE	Relative Improvement over MVE
fricatives	16.61	14.03	3.59	74.41
vowels	11.96	3.77	2.68	28.91
nasals	9.69	5.67	1.10	80.60
stops	14.15	5.97	2.44	59.13
others	36.09	21.90	6.75	69.18
Average	**17.70**	**10.27**	**3.31**	**62.45**

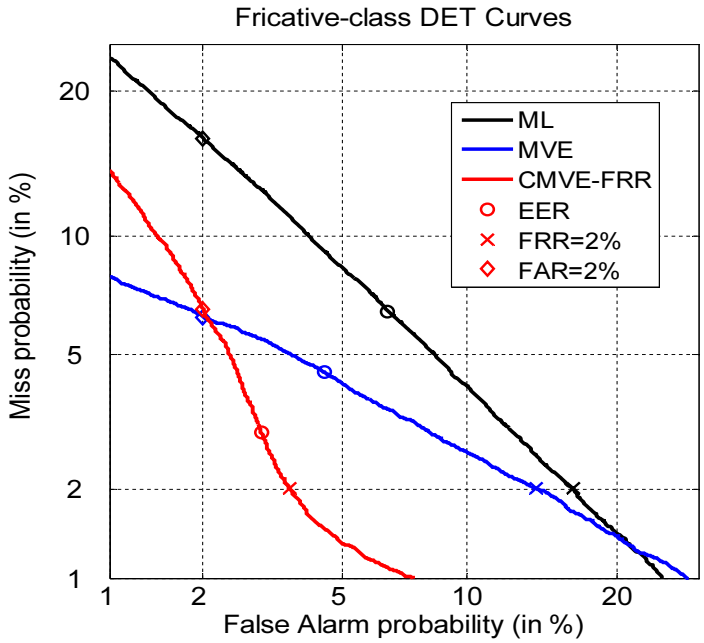


Figure 19: DET Analysis of fricative-class by ML, MVE, and CMVE at 2% False Rejection Rate Constraint

intended in the MVE criterion which aims at minimizing the total number of errors, MVE training has a much greater impact on the negative samples which directly lead to a low FAR.

Finally, the proposed CMVE method dramatically reduces the FARs of all sub-classes at the 2% FRR constraint over ML and MVE. The relative improvement of the proposed CMVE method over MVE on the training dataset is tabulated in the last column of Table 20. On a direct comparison of CMVE over MVE, an average relative improvement of 62.45% is obtained. In particular, two tricky classes, fricatives and others, which have been observed with the limited improvement by MVE, have been substantially enhanced by CMVE with a relative improvement of 74.41% and 69.18%, respectively. Meanwhile, the CMVE-trained models on nasals and stops yielded an extremely low FAR of 1.10% and 2.44%, respectively. The CMVE-trained model on vowels class showed a moderate performance improvement compared to other classes. Unlike other classes, statistics of the positive samples in the vowels class represent almost zero mean and small variance. It means that the threshold at the 2% FRR point is not far away from the zero that is normally set for a threshold of a mis-verification measure in the conventional MVE training. Therefore, the main gain by the CMVE method on this class comes from the ALM-based weighting rather than the threshold shifting.

Table 21 shows the results of the same models on the testing set. Experimental observations on the training set also hold for the testing set. The FARs of all sub-classes by the CMVE-trained models at a 2% FRR are considerably reduced when compared to ML and MVE. Although an average relative improvement from the training set to the testing set is reduced from 62.45% to 32.31%, the CMVE method still demonstrates its effectiveness for particular operating point optimization. Since the CMVE method makes use of the empirical weighting and threshold shifting on the training data over the iterations, it often gives rise to the over-fitting problem onto the training set. To prevent this over-fitting problem, regularization techniques [85, 122, 84] in the machine learning literature would be helpful.

Table 21: False Alarm Rate (%) at 2% False Rejection Rate Constraint on Testing Set

	ML	MVE	Constrained MVE	Relative Improvement over MVE
fricatives	15.84	14.56	7.25	50.21
vowels	12.59	5.45	4.96	8.99
nasals	9.95	5.69	4.06	28.65
stops	13.75	6.64	4.37	34.19
others	34.79	21.06	12.74	39.51
Average	**17.38**	**10.68**	**6.68**	**32.31**

This issue is currently out of scope of this dissertation. However, we expect that it would be a promising extension to the proposed CMVE method.

5.7.2 Minimize FRR at 2% FAR Constraint

The second experiment scenario is to minimize the FRR at a given 2% FAR constraint. The results on the BPC detection task by ML, MVE, and CMVE are summarized in Table 22. Similar to the first experiment results, the target error of all sub-classes by the ML method represents very high error rate at a given operating constraint. As discussed, the ML criterion cannot lead to the optimal performance for particular operating point optimization.

On the other hand, the FRRs of most sub-classes at the 2% FAR constraint are substantially reduced by the MVE method. For example, the FRRs of vowels and stops are reduced, from 33.32% to 3.46% and from 22.90% to 5.45%, respectively. In particular, for all sub-classes the MVE method provides much better results at an FAR constraint than at an FRR constraint. As seen in Table 19, the numbers of the negative segments are much more than those of the positive segments in the TIMIT training set. In this between-class imbalance dataset, the MVE method is more efficient at an FAR constraint so that the negative samples mostly contributes to the optimization. Therefore, by the MVE-trained models

Table 22: False Rejection Rate (%) at 2% False Alarm Rate Constraint on Training Set

	ML	MVE	Constrained MVE	Relative Improvement over MVE
fricatives	16.25	6.30	4.38	30.48
vowels	33.32	3.46	2.96	14.45
nasals	10.65	4.02	1.43	64.43
stops	22.90	5.45	3.18	41.65
others	44.83	13.10	9.60	26.72
Average	**25.59**	**6.47**	**4.31**	**35.54**

we can see an average FRR of 6.47% at a 2% FAR constraint, instead of an average FAR of 10.27% at a 2% FRR.

The proposed CMVE method still outperforms the MVE method at an FAR constraint. In the last column of Table 22, we can see that the relative improvement of the proposed CMVE method over MVE on the training dataset is 35.54%. At an FAR constraint, the CMVE method generally reduces the FRR in the ROC space when compared to MVE as shown in Figure 20. However, when compared to ML, the most reduced point lies on a 2% FAR as expected in its learning criterion. On the other hand, the results of the same models on the testing set are summarized in Table 23. Although experimental observations on the training set also hold for the testing set, we can also see the over-training problem as we have already seen in the first experiment scenario. An average relative improvement from the training set to the testing set is reduced from 35.54% to 11.09%. As discussed, it is necessary to apply some regularization techniques into the current CMVE learning framework so as to prevent the over-fitting problem.

In summary, Figure 21 shows an overall DET analysis by ML, MVE, and two different CMVE methods. It is evident that the CMVE methods result in mainly minimizing the target error at the given operating constraints, either at the 2% FRR or at the 2% FAR.

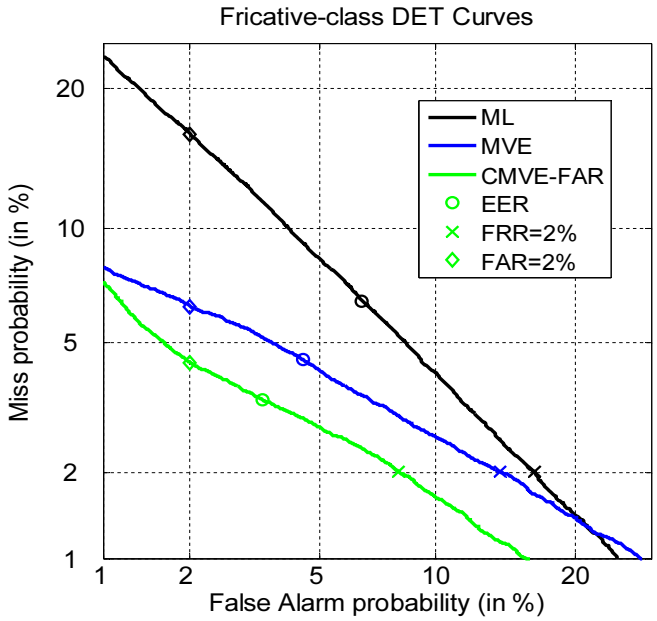
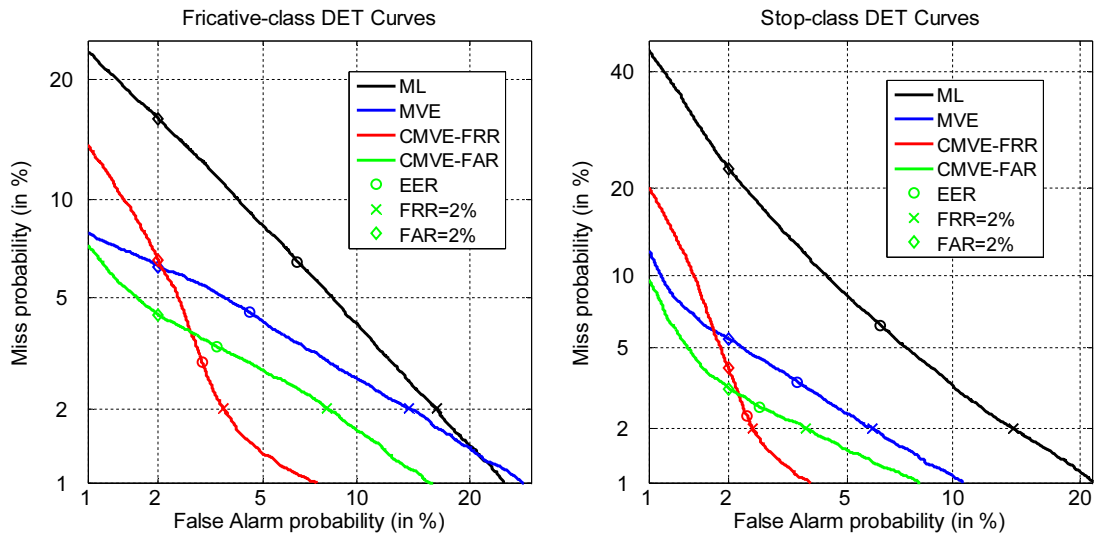


Figure 20: A DET Analysis of fricative-class by ML, MVE, and CMVE at 2% False Alarm Rate Constraint

Table 23: False Rejection Rate (%) at 2% False Alarm Rate Constraint on Testing Set

	ML	MVE	Constrained MVE	Relative Improvement over MVE
fricatives	16.60	7.36	6.56	10.87
vowels	35.90	6.11	5.33	12.77
nasals	11.13	4.65	4.00	13.98
stops	22.10	7.37	6.45	12.48
others	45.29	18.42	17.43	5.37
Average	**26.20**	**8.78**	**7.95**	**11.09**



(a) fricative-class. **(b) stop-class.**

Figure 21: An Overall DET Analysis by ML, MVE, and CMVE. (a) fricative-class. (b) stop-class.

Several conclusions in regard of a comparison study between MVE and CMVE are drawn from the DET curves in Figure 21 and the results in Tables 20–23:

- In the highly unbalanced dataset between the positive and negative samples, the MVE method yields *biased models* which were optimized toward the dominant-class samples although its objective criterion aims at minimizing an equally weighted sum of FAR and FRR.

- Given the highly imbalanced training samples, the proposed CMVE method directly minimizes the target error at any given operating constraints of the conflicting error, by making use of the ALM framework for an appropriate weighting mechanism.

- After MVE training, some sub-classes still show high error rates at a given low FAR or FRR. It it due to the rigid structure of the current MVE criterion using a fixed threshold over learning. Thus, some samples at a low FAR or FRR are regarded as *outliers* and would not be involved during learning.

- The decision threshold corresponding to a given operating point is iteratively updated

over learning in the proposed CMVE method. Through this iterative threshold shifting, the samples near the given operating point are appropriately taken into account during training optimization.

- Due to the use of the empirical weighting and threshold shifting during learning, the CMVE-trained models tend to be over-fitted to the training samples. To increase the generalization capability to the test samples, some regularization techniques, such as the margin concept in the SVM literature or the regularization techniques discussed in previous chapters, can be applied into the proposed CMVE learning framework.

5.8 Chapter Summary

In this chapter, we proposed new constrained discriminative training for particular operating characteristic optimization. In real-world applications, a model optimized at a particular operating scenario is often required. However, the conventional receiver operating characteristic (ROC) optimization methods cannot directly handle this practical requirement since they aim at overall performance optimization or a perfect ranking on the sample instances. To solve the problem of designing a constrained model optimized at a particular operating characteristic, we derived a constrained optimization formulation of MVE training by applying an augmented Lagrange multiplier (ALM) into the MVE criterion. The ALM framework makes the MVE objective function flexible enough to embed a particular operating need into the empirical error estimates. The proposed constrained MVE (CMVE) method directly minimizes the target error at any given operating constraint by making use of the ALM technique for an appropriate weighting mechanism. Meanwhile, the required decision threshold at a given operating point was searched at every iteration and directly used in a mis-verification measure so as to shift a critical decision boundary onto the point to be optimized. Through this iterative threshold shifting, sample data, evaluated to be near the threshold of the given operating point, are properly utilized in optimization.

We presented two sets of experimental results to demonstrate the effectiveness of our

approach. In particular, we chose two different types of constraints, a 2% FRR and a 2% FAR, to illustrate the design methodology without loss of generality. The main goal of the experiments was to investigate the tradeoff at a particular operating condition, not on the overall ROC space. Experimental results demonstrated that the proposed CMVE method results in mainly minimizing the target error at the given operating constraints, either at 2% FRR or at 2% FAR.

CHAPTER 6

CONCLUSION, CONTRIBUTIONS, AND FUTURE WORK

6.1 Conclusion and Contributions

In this thesis, we propose novel objective-driven discriminative training and adaptation frameworks, which are generalized from the minimum classification error (MCE) criterion, for various tasks and scenarios of speech recognition and detection. All proposed frameworks in this thesis were constructed to overcome the current limitations in utilizing the discriminative criteria for a task-specific goal or a particular scenario. Three task-specific requirements that many ASR applications often require in practice were addressed to formulate new objective-driven discriminative criteria. In this formulation, each objective required by an application or a developer is directly embedded into the learning criterion, thereby allowing system optimization to accomplish the desired performance. Through many mathematical derivations and experimental results, the proposed objective-driven discriminative training and adaptation frameworks are shown to accomplish theoretical optimality and encouraging results in various applications of speech recognition and detection. Major contributions of this thesis can be summarized as:

- Several novel discriminative criteria, generalized from the MCE criterion, are proposed beyond the general purposes of the current discriminative criteria.
- Each of the proposed discriminative criteria can optimize a system model for a task-specific goal or a particular scenario, which cannot be directly handled by the current discriminative criteria.
- A theoretical framework for optimal learning following the minimum error principle is provided in use of the proposed discriminative criteria.
- Extensive experimental validations are provided for various applications of speech recognition and detection.

6.1.1 Individual Error Minimization Learning Framework

A new discriminative training paradigm for direct minimization of each different type of the ASR errors was first proposed. We interpreted the commonly known three recognition error types, namely, insertion, deletion and substitution, from an event detection viewpoint and introduced an individual error minimization learning framework aiming at direct reduction of these individual errors. Specifically, three individual error minimization learning algorithms were proposed: MD(eletion)E, MI(nsertion)E, and MS(ubstitution)E, respectively. In addition, the adaptive utterance verification (UV) framework, in which MSE-trained models are utilized in both recognition and verification stages, was proposed to enhance the overall UV performance. The contributions in this topic are:

- New insights and ideas on how to interpret speech recognition and detection errors were provided. By re-interpreting the three types of recognition error in the context of a detection problem, the deletion, insertion, and substitution errors were respectively explained as miss, false alarm, and miss/false-alarm errors happening together.

- The re-interpretation of recognition and detection errors was directly embedded in formulating the individual error minimization learning framework. A theoretical framework was derived from the minimum verification error (MVE) criterion.

- An adaptive utterance verification (UV) framework was constructed by integrating the recognition and verification stages using the MSE-trained model thus overcoming several limitations of the conventional rigid two-stage UV.

6.1.2 New Approaches to Discriminative Linear Transform-based Adaptation using MCE and MVE Criteria

For this topic, we developed several novel discriminative linear transform (DLT) based adaptation methods for speech detection and recognition. The MVE linear regression (MVELR) was first proposed as a new discriminative adaptation method for speech detection and verification. Then, to deal with the generalization issue for rapid adaptation,

we proposed the regularized MCE linear regression (RMCELR) method. Furthermore, a structural framework for the prior density estimation in RMCELR was proposed. Extensive adaptation experiments were carried out on speech recognition and detection tasks to validate the effectiveness of the proposed methods. The contributions in this topic are:

- An objective function in the proposed MVELR adaptation was formulated as a way of keeping consistency between detector training and performance evaluation under a mismatched condition. This consistency directly led to the optimal detector performance in an adaptation scenario.

- The optimality of the DLTs can be ascertained with a sufficient amount of adaptation data. When data are severely limited, to overcome the limitation of the DLTs for rapid adaptation, a regularized MCE (RMCE) criterion was formulated by introducing the prior distribution as a regularization term to the original MCE empirical risk.

- Structural RMCELR (SRMCELR) was formulated, in which the prior densities for the transform matrices are hierarchically structured in a context decision tree according to the amount of the adaptation data available.

6.1.3 Constrained Discriminative Training for Particular Operating Point Optimization

A constrained optimization formulation of MVE training was proposed for operating characteristic optimization. Without loss of generality, we chose two different types of constraints, a 2% FRR and a 2% FAR, and investigated a design methodology which aims at minimizing the complementary type of errors. Experimental results demonstrate that the proposed constrained MVE methods result in mainly minimizing the target error at the given operating constraints, either at 2% FRR or at 2% FAR. The contributions in this topic are:

- A novel constrained discriminative training algorithm for operating characteristic optimization was formulated by applying an augmented Lagrange multiplier (ALM)

into the MVE criterion. An analytical solution and the corresponding optimization procedure were provided.

- Given the highly imbalanced training samples, the proposed CMVE method directly minimizes the target error at the given operating constraint, by making use of the ALM technique for an appropriate weighting mechanism.

- The decision threshold corresponding to a given operating point was iteratively updated over learning in the proposed CMVE method. Through this iterative threshold shifting, sample data, evaluated to be near the threshold of the given operating point, are properly utilized in optimization, resulting in substantial improvements in performance.

6.2 Future Work

We plan to explore more in extending the proposed objective-driven discriminative training and adaptation frameworks. First, for the individual error minimization learning, there are still many open issues. One of the important issue is context-dependent (CD) anti-subword modeling for improved discriminability during the DT phase. In this thesis, the anti-models for the limited context-independent (CI) monophones were employed with the CD target models in the DT phase. It is likely that the use of the CD anti-subword models discriminatively trained with the corresponding CD target models would lead to consistent and improved performance. Furthermore, the proposed individual error minimization learning can be further implemented on the weighted finite state transducer (WFST)-based hypotheses. The quality of the hypothesis space plays an important role in discriminative training. We believe that WFST-based approach would significantly improve the performance of the individual error minimization learning.

In addition, for the MCE-based DLTs in rapid adaptation, we plan to integrate a fully Bayesian treatment of the transformation matrices into the proposed RMCE objective function. The variational Bayes method would further improve the proposed discriminative

adaptation methods for rapid adaptation. Finally, in our particular operating point optimization experiments, we have seen that the CMVE-trained models tend to be over-fitted to the training samples. It is mainly due to the use of the empirical weighting and threshold shifting over training. To increase the generalization capability to the test samples, some regularization techniques, such as the margin concept in the SVM literature, can be incorporated into the proposed CMVE learning framework.

BIBLIOGRAPHY

[1] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA: Prentice Hall, 1993.

[2] X. Huang, A. Acero, and H. Hon, *Spoken language processing*. Upper Saddle River, NJ, USA: Prentice Hall, 2001.

[3] F. Jelinek, *Statistical methods for speech recognition*. MIT press, 1997.

[4] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[5] B. Juang and L. Rabiner, "Hidden markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.

[6] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition," *IEEE Signal Processing Magazine*, vol. 25, pp. 14–36, Sep. 2008.

[7] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," *Speech Communication*, vol. 34, no. 3, pp. 287–310, 2001.

[8] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Proc. Interspeech*, pp. 2133–2136, Sep. 2005.

[9] L. Bahl, P. Brown, P. De Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 49–52, Apr. 1986.

[10] Y. Normandin, *Hidden Markov models, maximum mutual information estimation, and the speech recognition problem*. McGill University, 1991.

[11] S. Kapadia, V. Valtchev, and S. Young, "Mmi training for continuous phoneme recognition on the timit database," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 491–494, Apr. 1993.

[12] B. Juang, W. Chou, and C. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 257–265, May 1997.

[13] B. Juang and S. Katagiri, "Discriminative learning for minimum error classification pattern recognition," *IEEE Transactions on Signal Processing*, vol. 40, pp. 3043–3054, Dec. 1992.

[14] W. Chou, "Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition," *Proceedings of the IEEE*, vol. 88, pp. 1201–1223, Aug. 2000.

[15] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 105–108, Jan. 2002.

[16] D. Povey, *Discriminative training for large vocabulary speech recognition*. PhD thesis, Cambridge, UK: Cambridge University, 2004.

[17] C. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proceedings of the IEEE*, vol. 88, pp. 1241–1269, Aug. 2000.

[18] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.

[19] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[20] S. Tsakalidis, V. Doumpiotis, and W. Byrne, "Discriminative linear transforms for feature normalization and speaker adaptation in hmm estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 367–376, May 2005.

[21] D. Povey, M. Gales, D. Kim, and P. Woodland, "Mmi-map and mpe-map for acoustic model adaptation," in *Proc. Eighth European Conference on Speech Communication and Technology*, pp. 1981–1984, 2003.

[22] P. Woodland, "Speaker adaptation for continuous density hmms: A review," in *Proc. ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, pp. 11–19, Aug. 2001.

[23] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *Proc. Eurospeech*, pp. 1203–1206, 2001.

[24] J. Wu and Q. Huo, "A study of minimum classification error (mce) linear regression for supervised adaptation of mce-trained continuous-density hidden markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 478–488, Feb. 2007.

[25] L. Uebel and P. Woodland, "Improvements in linear transform based speaker adaptation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 49–52, 2001.

[26] L. Wang and P. Woodland, "Mpe-based discriminative linear transforms for speaker adaptation," *Computer Speech and Language*, vol. 22, no. 3, pp. 256–272, 2008.

[27] E. McDermott, T. J. Hazen, J. Le Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 203–223, 2007.

[28] Y. Zhao, A. Ljolje, D. Caseiro, and B.-H. Juang, "A general discriminative training algorithm for speech recognition using weighted finite-state transducers," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4217–4220, IEEE, 2012.

[29] F. Ehsani and E. Knodt, "Speech technology in computer-aided language learning: Strengths and limitations of a new call paradigm," *Language Learning and Technology*, vol. 2, no. 1, pp. 45–60, 1998.

[30] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[31] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, 1995.

[32] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[33] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 695–707, Nov. 2000.

[34] T. J. Hazen, "A comparison of novel techniques for rapid speaker adaptation," *Speech Communication*, vol. 31, no. 1, pp. 15–33, 2000.

[35] A. Gunawardana and W. Byrne, "Discounted likelihood linear regression for rapid speaker adaptation," *Computer Speech and Language*, vol. 15, no. 1, pp. 15–38, 2001.

[36] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *Machine Learning*, vol. 31, pp. 1–38, 2004.

[37] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001.

[38] A. Rakotomamonjy, "Optimizing area under roc curves with svms," 2004.

[39] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[40] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, p. 1738, 1990.

[41] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, p. 1304, 1974.

[42] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[43] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, pp. 353–356, IEEE, 1996.

[44] L. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[45] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[46] V. Valtchev, J. Odell, P. Woodland, and S. Young, "Mmie training of large vocabulary recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303–314, 1997.

[47] P. Woodland and D. Povey, "Large scale discriminative training of hidden markov models for speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 25–47, 2002.

[48] B. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, vol. 5, no. 3, pp. 275–294, 1991.

[49] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[50] P. Gopalakrishnan, D. Kanevsky, A. Nádas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Transactions on Information Theory*, vol. 37, pp. 107–113, Jan. 1991.

[51] W. Chou, C. Lee, and B. Juang, "Minimum error rate training based on n-best string models," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 652–655, Apr. 1993.

[52] R. Schluter and W. Macherey, "Comparison of discriminative training criteria," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 493–496, May 1998.

[53] Q. Fu, X. He, and L. Deng, "Phone-discriminating minimum classification error (p-mce) training for phonetic recognition," in *Proc. Interspeech*, pp. 2073–2076, Aug. 2007.

[54] E. McDermott, S. Watanabe, and A. Nakamura, "Margin-space integration of mpe loss via differencing of mmi functionals for generalized error-weighted discriminative training," in *Proc. Interspeech*, pp. 224–227, Sep. 2009.

[55] X. He and W. Chou, "Minimum classification error linear regression for acoustic model adaptation of continuous density hmms," in *Proc. International Conference on Multimedia and Expo (ICME)*, vol. 1, pp. 397–400, July 2003.

[56] M. Gibson, *Minimum Bayes risk acoustic model estimation and adaptation*. PhD thesis, University of Sheffield, 2008.

[57] S. Matsuda, Y. Tsao, J. Li, S. Nakamura, and C. Lee, "A study on soft margin estimation of linear regression parameters for speaker adaptation," in *Proc. Interspeech*, pp. 1603–1606, Sep. 2009.

[58] S. Furui, "Generalization problem in asr acoustic model training and adaptation," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 1–10, Dec. 2009.

[59] Q. Fu, *A generalization of the minimum classification error (MCE) training method for speech recognition and detection*. PhD thesis, Georgia Institute of Technology, 2008.

[60] M. Rahim and C. Lee, "String-based minimum verification error (sb-mve) training for speech recognition," *Computer Speech and Language*, vol. 11, no. 2, pp. 147–160, 1997.

[61] A. Rosenberg, O. Siohan, and S. Parathasarathy, "Speaker verification using minimum verification error training," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 105–108, May 1998.

[62] Y. Liao, J. Tu, S. Chang, and C. Lee, "An enhanced minimum classification error learning framework for balancing insertion, deletion and substitution errors," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 587–590, Dec. 2007.

[63] Q. Fu and B.-H. Juang, "Segment-based phonetic class detection using minimum verification error (mve) training," in *Proc. Interspeech*, pp. 3029–3032, Sep. 2005.

[64] S. Shin, H. Y. Jung, T. Y. Kim, and B. H. Juang, "Discriminative linear-transform based adaptation using minimum verification error," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4318–4321, Mar. 2010.

[65] P. Ramesh, C. Lee, and B. Juang, "Context dependent anti subword modeling for utterance verification," in *Proc. Fifth International Conference on Spoken Language Processing*, pp. 3233–3236, 1998.

[66] Q. Fu and B. H. Juang, "Investigation on rescoring using minimum verification error (mve) detectors," in *Proc. Interspeech*, pp. 677–680, Sep. 2006.

[67] Q. Fu and B. H. Juang, "A study on rescoring using hmm-based detectors for continuous speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 570–575, Dec. 2007.

[68] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 347–354, Dec. 1997.

[69] S. Yaman and C. Lee, "A flexible classifier design framework based on multiobjective programming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 779–789, May 2008.

[70] K. Lee and H. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 1641–1648, Nov. 1989.

[71] S. Young, G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book version 3.4*. Cambridge University Engineering Department, 2006.

[72] L. Bahl, S. Balakrishnan-Aiyer, J. Bellgarda, M. Franz, P. Gopalakrishnan, D. Nahamoo, M. Novak, M. Padmanabhan, M. Picheny, and S. Roukos, "Performance of the ibm large vocabulary continuous speech recognition system on the arpa wall street journal task," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 41–44, May 1995.

[73] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," *The Journal of the Acoustical Society of America*, vol. 108, p. 3036, 2000.

[74] S. Furui, "Recent advances in spontaneous speech recognition and understanding," in *ISCA & IEEE workshop on spontaneous speech processing and recognition*, 2003.

[75] M. G. Rahim, C. H. Lee, and B. H. Juang, "Discriminative utterance verification for connected digits recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 266–277, May 1997.

[76] R. Sukkar and C. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 420–429, Nov. 1996.

[77] E. Lleida and R. Rose, "Utterance verification in continuous speech recognition: Decoding and training procedures," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 126–139, Mar. 2000.

[78] E. Lehmann and J. Romano, *Testing statistical hypotheses*. Springer Verlag, 2005.

[79] M. Koo, C. Lee, and B. Juang, "Speech recognition and utterance verification based on a generalized confidence score," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 821–832, Nov. 2001.

[80] B. H. Juang and S. Furui, "Automatic recognition and understanding of spoken language-a first step toward natural human-machine communication," *Proceedings of the IEEE*, vol. 88, pp. 1142–1165, Aug. 2000.

[81] C. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, pp. 109–111, Oct. 2004.

[82] Y. Zhao, S. Shin, E. Robledo-Arnuncio, and B. Juang, "A study on recognizing distorted speech over local distributed transducer networks," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4181–4184, Apr. 2009.

[83] J. Deller, J. Proakis, and J. Hansen, *Discrete-time processing of speech signals*. Macmillan Publishing Company New York, 1993.

[84] A. Biem, J. Ha, and J. Subrahmonia, "A bayesian model selection criterion for hmm topology optimization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 989–992, May 2002.

[85] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. The Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152, July 1992.

[86] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," in *In Advances in Neural Information Processing Systems 12*, pp. 470–476, MIT Press, 1999.

[87] D. Mansjur, *Statistical pattern recognition approaches for retrieval-based machine translation systems*. PhD thesis, Georgia Institute of Technology, 2011.

[88] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[89] N. Ueda and Z. Ghahramani, "Bayesian model search for mixture models based on optimizing variational bounds," *Neural Networks*, vol. 15, no. 10, pp. 1223–1242, 2002.

[90] C. Chesta, O. Siohan, and C. Lee, "Maximum a posteriori linear regression for hidden markov model adaptation," in *Proc. EuroSpeech*, pp. 211–214, 1999.

[91] O. Siohan, T. Myrvoll, and C. Lee, "Structural maximum a posteriori linear regression for fast hmm adaptation," *Computer Speech and Language*, vol. 16, no. 1, pp. 5–24, 2002.

[92] C. Breslin, K. Chin, M. Gales, K. Knill, and H. Xu, "Prior information for rapid speaker adaptation," in *Proc. Interspeech*, pp. 1644–1647, 2010.

[93] P. T. Akhil, S. P. Rath, S. Umesh, and D. R. Sanand, "A computationally efficient approach to warp factor estimation in vtln using em algorithm and sufficient statistics," in *Proc. Interspeech*, pp. 1713–1716, 2008.

[94] A. Gupta and T. Varga, *Elliptically contoured models in statistics*. Kluwer Academic Publishers, 1993.

[95] E. Lukacs, *Stochastic convergence*, vol. 39. Academic Press New York, 1975.

[96] H. J. Kushner and G. G. Yin, "Stochastic approximation algorithms and applications," 1997.

[97] S. Katagiri, B.-H. Juang, and C.-H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2345–2373, 1998.

[98] J. Bernardo, A. Smith, and M. Berliner, *Bayesian theory*. Wiley New York, 1994.

[99] K. Shinoda and C.-H. Lee, "Structural map speaker adaptation using hierarchical priors," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pp. 381–388, IEEE, 1997.

[100] K. Shinoda and C.-H. Lee, "A structural bayes approach to speaker adaptation," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 276–287, 2001.

[101] O. S. T. Myrvoll and C.-H. Lee, "Structural maximum a posteriori linear regression for fast hmm adaptation," *ICSA ITRW ASR*, 2000.

[102] W. Chou, O. Siohan, T. A. Myrvoll, and C.-H. Lee, "Extended maximum a posterior linear regression (emaplr) model adaptation for speech recognition.," in *INTERSPEECH*, pp. 616–619, 2000.

[103] R. Chengalvarayan, "Speaker adaptation using discriminative linear regression on time-varying mean parameters in trended hmm," *IEEE Signal Processing Letters*, vol. 5, pp. 63–65, Mar. 1998.

[104] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[105] A. Herschtal and B. Raskutti, "Optimising area under the roc curve using gradient descent," in *Proceedings of the twenty-first international conference on Machine learning*, p. 49, ACM, 2004.

[106] S. H. Shin, H. Y. Jung, and B. H. Juang, "An adaptive utterance verification framework using minimum verification error training," *ETRI Journal*, vol. 33, pp. 423–433, June 2011.

[107] D. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Nashua, NH: Athena Scientific, 1996.

[108] R. Fletcher, *Practical methods of optimization*. New York: Wiley, 2000.

[109] Y. Sawaragi, H. Nakayama, and T. Tanino, *Theory of multiobjective optimization*, vol. 176. Academic Press New York, 1985.

[110] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[111] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[112] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[113] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1, pp. 91–108, 1995.

[114] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Proc. of Eurospeech*, pp. 1895–1898, 1997.

[115] A. Martin and M. Przybocki, "The nist 1999 speaker recognition evaluation: an overview," *Digital signal processing*, vol. 10, no. 1, pp. 1–18, 2000.

[116] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[117] W. Conover and R. L. Iman, "Rank transformations as a bridge between parametric and nonparametric statistics," *The American Statistician*, vol. 35, no. 3, pp. 124–129, 1981.

[118] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, ACM, 2006.

[119] K. Ataman, W. N. Street, and Y. Zhang, "Learning to rank by maximizing auc with linear programming," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, pp. 123–129, IEEE, 2006.

[120] C. Cortes and M. Mohri, "Auc optimization vs. error rate minimization," *Advances in neural information processing systems*, vol. 16, no. 16, pp. 313–320, 2004.

[121] L. Yan, R. Dodier, M. C. Mozer, and R. Wolniewicz, "Optimizing classifier performance via the wilcoxon-mann-whitney statistics," in *Proceedings of the 20th international conference on machine learning*, pp. 848–855, Citeseer, 2003.

[122] T. Joachims, "Making large scale svm learning practical," 1999.

[123] K. Miettinen, *Nonlinear multiobjective optimization*. New York: Springer, 1999.

[124] K. Deb, "Multi-objective optimization," *Multi-objective optimization using evolutionary algorithms*, pp. 13–46, 2001.

[125] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach," *Evolutionary Computation, IEEE Transactions on*, vol. 3, no. 4, pp. 257–271, 1999.

[126] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *Evolutionary Computation, IEEE Transactions on*, vol. 6, no. 2, pp. 182–197, 2002.

[127] J. Lévesque, A. Durand, C. Gagné, and R. Sabourin, "Multi-objective evolutionary optimization for generating ensembles of classifiers in the roc space," in *Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference*, pp. 879–886, ACM, 2012.

[128] J. W. Lee and S. H. Kim, "Using analytic network process and goal programming for interdependent information system project selection," *Computers & Operations Research*, vol. 27, no. 4, pp. 367–382, 2000.

[129] M. Tamiz, D. Jones, and C. Romero, "Goal programming for decision making: An overview of the current state-of-the-art," *European Journal of operational research*, vol. 111, no. 3, pp. 569–581, 1998.

[130] C. Marrocco, M. Molinara, and F. Tortorella, "On linear combinations of dichotomizers for maximizing the area under the roc curve," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 41, no. 3, pp. 610–620, 2011.

[131] C. Chatelain, S. Adam, Y. Lecourtier, L. Heutte, and T. Paquet, "A multi-model selection framework for unknown and/or evolutive misclassification cost problems," *Pattern Recognition*, vol. 43, no. 3, pp. 815–823, 2010.

[132] M. D. Del Castillo and J. I. Serrano, "A multistrategy approach for digital text categorization from imbalanced documents," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 70–79, 2004.

[133] S. Gao, C. Lee, and J. Lim, "An ensemble classifier learning approach to roc optimization," in *Proc. 18th International Conference on Pattern Recognition (ICPR)*, vol. 2, pp. 679–682, Aug. 2006.

[134] Q. Tao and R. Veldhuis, "Threshold-optimized decision-level fusion and its application to biometrics," *Pattern Recognition*, vol. 42, no. 5, pp. 823–836, 2009.

[135] M. J. Gacto, R. Alcalá, and F. Herrera, "Adaptation and application of multi-objective evolutionary algorithms for rule reduction and parameter tuning of fuzzy rule-based systems," *Soft Computing*, vol. 13, no. 5, pp. 419–436, 2009.

[136] M.-H. Siu, B. Mak, and W.-H. Au, "Minimization of utterance verification error rate as a constrained optimization problem," *Signal Processing Letters, IEEE*, vol. 13, no. 12, pp. 760–763, 2006.

[137] J. Nocedal and S. Wright, *Numerical optimization*. New York: Springer verlag, 1999.