Robustness in ASR: An Experimental Study of the Interrelationship between Discriminant Feature-Space Transformation, Speaker Normalization and Environment Compensation

Alireza Keyvani

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

January 2007

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Engineering.

2007/01/15

Abstract

This thesis addresses the general problem of maintaining robust automatic speech recognition (ASR) performance under diverse speaker populations, channel conditions, and acoustic environments. To this end, the thesis analyzes the interactions between environment compensation techniques, frequency warping based speaker normalization, and discriminant feature-space transformation (DFT). These interactions were quantified by performing experiments on the connected digit utterances comprising the Aurora 2 database, using continuous density hidden Markov models (HMM) representing individual digits.

Firstly, given that the performance of speaker normalization techniques degrades in the presence of noise, it is shown that reducing the effects of noise through environmental compensation, prior to speaker normalization, leads to substantial improvements in ASR performance. The speaker normalization techniques considered here were vocal tract length normalization (VTLN) and the augmented state-space acoustic decoder (MATE). Secondly, given that discriminant feature-space transformation (DFT) are known to increase class separation, it is shown that performing speaker normalization using VTLN in a discriminant feature-space leads to improvements in the performance of this technique. Classes, in our experiments, corresponded to HMM states. Thirdly, an effort was made to achieve higher class discrimination by normalizing the speech data used to estimate the discriminant feature-space transform. Normalization, in our experiments, corresponded to reducing the variability within each class through the use of environment compensation and speaker normalization. Significant ASR performance improvements were obtained when normalization was performed using environment compensation, while our results were inconclusive for the case where normalization consisted of speaker normalization. Finally, aimed at increasing its noise robustness, a simple modification of MATE is presented. This modification consisted of using, during recognition, knowledge of the distribution of warping factors selected by MATE during training.

Sommaire

Cette thèse adresse le problème général de maintenir l'exécution robuste de reconnaissance automatique de la parole (ASR) pour divers populations de locuteur, états de canal, et environnements acoustiques. À cet effet, la thèse analyse les interactions entre les techniques de compensation d'environnement, les techniques de la normalisation de locuteur basé sur la déformation de la fréquence, ainsi que la transformation discriminante de l'espace des attributs (DFT.) Ces interactions ont été mesurées en effectuant des expériences sur des émissions de paroles de chiffres relies compris dans la base de données d' Aurora 2, en utilisant des modèles de Markov cachées a densité continue (HMM). Premièrement, étant donné que l'exécution des techniques de normalisation de locuteur dégrade en présence du bruit, il est démontré que la réduction des effets du bruit par la compensation environnementale, avant la normalisation de locuteur, mène aux améliorations substantielles de l'exécution d'ASR. Les techniques de normalisation de locuteur considérées ici étaient la normalisation de la longueur de tractus aérien (VTLN) ainsi que le décodeur acoustique de l'espace d'état augmenté (MATE.) Deuxièmement, étant donné que la transformation discriminante de l'espace des attributs (DFT) est connue pour augmenter la séparation de classes, il est démontré que la normalisation de locuteur employant le VTLN dans un espace des attributs discriminant mène aux améliorations de l'exécution de cette technique. Dans nos expériences, les classes correspondaient aux états d'HMM. Troisièmement, dans le but de réaliser une discrimination de classe plus élevée, les données de la parole employées pour estimer la transformation discriminante de l'espace des attributs ont été normalisées. Dans nos expériences, la normalisation a été effectuée par une réduction de la variabilité dans chaque classe à l'aide de la compensation d'environnement et la normalisation du locuteur. Des améliorations significatives d'exécution d'ASR ont été obtenues quand la normalisation a été effectuée en utilisant la compensation d'environnement, alors que les résultats obtenus avec la normalisation du locuteur ont été peu concluants. En conclusion, une modification simple de MATE est présentée afin d'augmenter la robustesse contre le bruit. Cette modification, lors de la phase d'identification consistait à employer la connaissance de la distribution des facteurs de déformation choisis par MATE pendant la formation.

Acknowledgments

I would like to express my gratitude towards Dr. Richard Rose for his guidance, supervision, support and friendliness, which made the performing of this research all the more rewarding. I would also like thank my co-supervisor Dr. Peter Kabal for his constant kind support and motivation. Another thank you goes to Dr. Antonio Miguel whose support and collaboration with this work was outstanding. Finally, I would like to thank everybody at McGill who's everyday work keeps this institution up and running.

Contents

List of Figures

List of Tables

List of Acronyms

ASR	Automatic Speech Recognition
AFE	Advanced Front-End
DFT	Discriminant Feature-Space Transformation
DSR	Distribued Speech Recognition
ETSI	European Telecommunications Standards Institute
HDA	Heteroscedastic Discriminant Analysis
HMM	Hidden Markov Model
LDA	Linear Discriminant Analysis
MATE	augMented stAte-space acousTic dEcoder
MDA	Multiple Discriminant Analysis
MFCC	Mel-Frequency Cepstral Coefficients
MLLT	Maximum Likelihood Linear Transformation
PCA	Principle Component Analysis
SD	speaker dependent
SI	speaker independent
SWP	SNR-dependent Waveform Processing
SNR	Signal to Noise Ratio
VTLN	Vocal Tract Length Normalization
WER	Word Error Rate

Chapter 1

Introduction

The field of Automatic Speech Recognition (ASR) has undergone substantial advances during recent years. Today's ASR systems are able to perform with considerable accuracy. As a result, the technology is finding more and more applications: ASR is now used in systems such as automated call centers, cellular telephones, automobiles, and personal computers. Nevertheless, the performance of such systems is far from perfect.

There are various issues concerning the operation of modern ASR systems, which lead to reductions in their efficiency and accuracy. Among these issues is the existence of various forms of variability in speech. These include variations in acoustic environment, communications channel, and speaker characteristics. Hence, much effort has been devoted to finding ways of dealing with these issues. Furthermore, the notion of class discrimination plays an important part in determining the performance of ASR systems. In the following subsections, we will first describe these sources of variability as well as the notion of class discrimination, and then, present the main goals of this thesis as well as its outline.

1.1 Variations in Acoustic Environment and Communications Channel

A major factor that leads to degradations in the performance of ASR systems is the presence of noise in the environment. Such degradations in performance can be explained in terms of the mismatch between the conditions in which the systems are trained and the ones in which they are operated. For example, Lockwood and Boudy reported a 70% degradation in

the accuracy of a conventional word recognizer when it was trained in clean conditions and tested in a car travelling at 90km/h [1]. Also, as another example, Das *et al.* reported an error rate of 50% from the IBM Tanagora speech recognition system when the system was trained with isolated words under clean conditions and tested in a cafeteria environment, while the same system had yielded a mere 1% error rate under clean conditions [2].

The approaches designed to battle the mismatch between training and operating environment can be classified into three categories [1]. The first category is concerned with the feature analysis stage of ASR, the function of which is to obtain information from the speech waveform that is relevant to the pattern classification task, discarding any redundant information. The approaches in the first category are aimed at extracting features from the speech signal which are *insensitive* to noise. For example, Xu and Wei proposed a scheme which assumes the statistics of the noise signal to be stationary over successive speech frames. Therefore, by extracting features based on the difference of short-time power spectrum of the speech signal, the effects of additive noise are effectively cancelled [3]. Other examples include feature extraction algorithms based on the human auditory system, such as the one presented by Kim *et al.* [4].

The second category of methods is also concerned with the feature analysis stage, and consists of techniques which attempt to *remove* the effects of noise from the speech signal. These techniques often utilize *a priori* information about the properties of speech and noise to recover clean speech from noisy speech. For example, the model-based Wiener filter (MBW) method combines spectral subtraction, Wiener filtering, and minimum mean square error estimation, based on a Gaussian mixture model representing pre-trained knowledge of speech, to reduce the effects of noise [5]. Another example is the European Telecommunications Standards Institute advanced front-end (ETSI-AFE) [6] which combines blind equalization, a two-stage Wiener filter design and SNR-dependent waveform processing (SWP) [7] to perform this task. The ETSI-AFE is the method of choice for environment compensation in this thesis and is further described in Chapter 2.

The methods in the third category, instead of compensating the noisy speech signal, attempt to adapt the acoustic model, originally trained under clean conditions, to the noisy environment. Parallel model combination (PMC) [8] is one such approach which is applied to hidden Markov model (HMM)-based systems [9]. The technique models noise as separate HMMs which are combined with the clean condition acoustic HMMs to yield the compensated models.

While the three categories of methods mentioned above focus on the effects of the acoustic environment, other methods are required to deal with the characteristics of the communications channel as a major source of ASR performance degradation. For example, in the case of packet-based mobile and IP networks, codec distortion and packet loss are two major problems to contend with. When the ASR system is located on a server connected through a network to the terminal device where the speech input occurs, a distributed speech recognition (DSR) architecture can be used to avoid codec distortion. This is done by moving the feature extraction stage of ASR to the terminal device. Also, to battle packet loss, techniques have been used such as the maximum a-posteriori (MAP) estimation of lost packets using the statistics of the packet stream [10].

1.2 Variations in Speaker Characteristics

Another major source of performance degradation in ASR is the variability caused by physiological and dialectical differences among different speakers. Evidence of this can be found by comparing the performance of speaker dependent (SD) and speaker independent (SI) ASR systems. A speaker dependent system is trained using data from a single speaker, and is intended for use by that *same* speaker. On the other hand, a speaker independent system is trained using data from a large population of different speakers for use by speakers that are not necessarily in the training population. To illustrate the difference in performance, Huang and Lee performed a comparison of SI and SD systems on the DARPA Resource Management Task. They reported a word error rate of 4.3% using a SI system, while, using a SD system on the same task, the error rate was only 1.4% [11].

As the performance degradation stated above is typical of many SI systems where training and testing are performed using utterances from many differing speakers, it is clear that there is a need for techniques that make ASR systems more robust against differences among speaker characteristics. These techniques can be divided into two main categories: speaker adaptation and speaker normalization.

Speaker adaptation techniques require the existence of a model which has already been trained for one or many speaker. The goal of such techniques is to tune the parameters of this model to a new speaker. To this end, a set of speech samples from the new speaker are used as adaptation data. Depending on how the adaptation is performed, these techniques differ in the amount of data they need. For example, the speaker adaptation algorithm pro-

posed in [12] utilizes maximum posteriori (MAP) estimation to perform speaker adaptation of HMM-based speech recognizers. The approach requires a large amount of adaptation data since it only updates those models for which there are examples in the data [13]. Another technique for adapting HMM-based systems is the maximum likelihood linear regression (MLLR) approach [13] which requires a relatively small amount of adaptation data. This data is used to compute a number of linear transformations which are applied to the distribution means contained in the model.

Speaker normalization techniques, on the other hand, instead of compensating the model, perform transformations on the speech signal to compensate for speaker variabilities. Due to the significant variations in the vocal tract length of different speakers, the positions of the formants produced by different speakers can vary as much as 25% [14]. Therefore, a major category of recent speaker normalization techniques are focused on normalizing the effective vocal tract length across different speakers. Vocal tract length normalization (VTLN) [14] and augmented state-space acoustic decoder (MATE) [15] perform this by applying a linear warping to the frequency axis of the utterance, normalizing the position of spectral peaks or formants of speech. These methods are further described in Chapter 2.

1.3 Class Discrimination in Statistical Pattern Recognition

The function of the feature extraction stage of ASR is to reduce the high amount of information contained in a speech signal and produce features that are most relevant to the ASR task, discarding any redundancy. One of the most common feature analysis techniques in ASR involves concatenating a set of static and dynamic feature vectors for each given speech frame. The static features are often computed using cepstral analysis [16] which is based on the magnitude of the short-time spectral envelope of the signal. Dynamic features, on the other hand, are computed to capture a measure of the time evolution of the spectral content of the signal [17]. Such information has been shown to play an important role in human speech perception [18] and also to increase ASR performance [17]. For a given frame, these dynamic features are computed based on the derivatives of the trajectories of the spectral parameters of the current frame [17].

Although the above technique has been widely used in the literature, the fact remains that the simple concatenation of dynamic features to the static features is not necessarily

the best way to capture the time evolution of speech frames. Therefore, efforts have been made to perform this task in a more mathematically sound framework. The underlying assumption of such developments is the consideration of ASR as a type of classification problem. Although, as it is not clear what the best definition of classes should be, various choices such as words and phones have been considered in the literature [19]. Regardless of this choice, the fundamental problem in ASR is to discriminate between such classes, and hence, ASR can be viewed as a special type of classification problem in statistical pattern recognition.

Considering the classification problem inherent in ASR, recent techniques have approached the problem of capturing the time evolution of speech frames in a new light. In such approaches, instead of performing the simple concatenation of static and dynamic features discussed above, the extracted features for a given frame are concatenated with features from a number of surrounding frames. Then, a discriminant feature-space transformation (DFT) technique is used to reduce the dimensionality of the resulting vector while maximizing class discrimination. Linear discriminant analysis (LDA) [20] is a standard method for dimensionality reduction with the constraint of maximizing class discrimination [21, 22]. In the transformed space, the feature vectors, while containing information about the time evolution of the spectral content of the signal in the current frame, are optimal in the sense that they allow for maximum class discrimination.

Despite the popularity of LDA, the method has shown gains for small vocabulary tasks, while yielding mixed results for large vocabulary tasks [21, 22]. One reason for this shortcoming is that, while the underlying ASR models are often trained with the assumption that the data has diagonal covariance, LDA produces a projected space whose dimensions might be highly correlated. Therefore, maximum likelihood feature-space transformations (MLLT) have been used to diagonalize the resulting space [19]. Another shortcoming of LDA is the fact that it assumes the data assigned to each class to have the same covariance. Therefore, subsequent work was concerned with the generalization of LDA to heteroscedastic discriminant analysis (HDA) which does not require such an assumption [23]. These two discriminant feature-space transformation (DFT) method are further described in Chapter 2.

1.4 Thesis Statement

The purpose of this thesis is to perform an experimental study concerned with the notion of robustness against sources of variability in automatic speech recognition (ASR). We believe that by combining environment compensation, speaker normalization, and discriminant feature-space transformation (DFT), we can improve the robustness of ASR systems, and therefore improve their performance. The specific claims that were investigated in this respect are stated below. The research performed as part of this thesis was aimed at motivating and investigating the validity of these claims through experiments.

1.4.1 Claims

The first claim considered in this thesis states that the performance of speaker normalization should improve through the use of environment compensation. We will motivate this claim by considering the effects of noise on the processes comprising the speaker normalization techniques, and subsequently, examining the effects of environment compensation on these processes.

To state the second claim, we note that, in the sense outlined in Section 1.3, class discrimination has a direct impact upon the performance of speech recognition. Furthermore, it can be expected that class discrimination also affects the performance of speaker normalization algorithms. Therefore, we argue that when discriminative algorithms, such as LDA and HDA, are used to increase class discrimination, the performance of speaker normalization techniques, such as MATE and VTLN, should improve. This comprises the second claim of this thesis.

On the other hand, class discrimination is reduced as a result of the increase in within-class variance caused by variabilities in environment, channel and speaker characteristics. As a result, since environment compensation and speaker normalization techniques essentially reduce within-class variance due to such sources of variability, we claim that these techniques should improve the performance of algorithms aimed at increasing class discrimination, which should in turn lead to improvements in ASR performance.

This constitutes the final claim of this thesis and is evaluated based on the following scheme. First, we use environment compensation to remove variabilities in each class due to channel and environment effects. Then, we use speaker normalization techniques such as VTLN and MATE to reduce inter-speaker variabilities in each class. As a result, the

variance of the "normalized" data in each class should now have been minimized. Finally, we apply discriminant feature-space transformation techniques, such as LDA and HDA to project the normalized data into a space where classes are maximally separated. In this new space, we expect the performance of ASR to improve. Here, it is important to recognize a duality embedded in the above claims. On the one hand, we claim that increasing class discrimination leads to improved speaker normalization, while on the other hand, we claim that reducing speaker variability leads to better class discrimination.

1.5 Thesis Outline

This thesis starts by presenting the required technical background information in Chapter 2. In this chapter, we will first describe the Mel-frequency cepstral coefficients (MFCC) used as the underlying feature analysis algorithm in our experiments. Second, we will provide an overview of continuous density hidden Markov model (HMM)-based speech recognition. Third, we will describe the standard noise-robust feature analysis algorithm we used for environment compensation. Fourth, we will look at speaker normalization techniques, including two specific techniques used in our experiments, namely vocal tract length normalization (VTLN) and augmented state-space acoustic decoder (MATE.) Finally, the chapter will present an overview of discriminant feature-space transformation (DFT) techniques, including linear discriminant analysis (LDA) and heteroscedastic discriminant analysis (HDA.)

Chapter 3 presents the specifics of the experimental setup used for the purpose of our experiments. The chapter includes details regarding the speech corpus as well as the baseline speech recognition system our experiments were based on. Furthermore, the chapter describes specifically how environmental compensation, speaker normalization and discriminant feature-space transformation were combined to assess the validity of our claims.

Chapter 4 focuses on presenting the actual experiments performed to motivate and assess the three claims of this thesis, as stated in Section 1.4. In evaluating the claims of this thesis, we noted certain shortcomings of the MATE speaker normalization technique in noise, and correspondingly devised a simple modification to this technique to increase its robustness. This new technique is also presented as part of Chapter 4.

Finally, Chapter 5 will summarize the work done for the purpose of this thesis. Major conclusions are highlighted, and potential areas for future work are stated.

Chapter 2

Background

The overall focus of this thesis is robustness against sources of variability in ASR systems. To this end, experiments are performed to investigate the interaction between techniques that reduce the effects of speaker and environment variability, as well as techniques that increase class separability in ASR. The purpose of this chapter is to briefly introduce the various techniques and algorithms employed for the purpose of our experiments. First, we will describe the Mel-frequency cepstral coefficients (MFCC) used as the underlying feature analysis algorithm in our experiments. Second, we will provide an overview of continuous density hidden Markov model (HMM)-based speech recognition, which our experiments are based on. Third, we will describe the standard noise-robust feature analysis algorithm we used for environment compensation. Fourth, we will look at speaker normalization techniques. This will include vocal tract length normalization (VTLN), which is aimed at removing global utterance-level variability, as well as the augmented state-space acoustic decoder (MATE), which attempts to remove localized frame-level variability. Finally, the chapter will present an overview of discriminant feature-space transformation (DFT) techniques. The two approaches considered, namely linear discriminant analysis (LDA) and heteroscedastic discriminant analysis (HDA), differ in their assumptions concerning the class-specific distribution of data.

2.1 Feature Analysis

The feature analysis component of an ASR system plays a crucial role in the overall performance of the system. As in any pattern classification problem, the goal of feature analysis

is to obtain information from the speech waveform that is relevant to the pattern classification task, and discard information which is redundant or does not contribute to class separability. In ASR, as the definition of what constitutes relevant information depends on the design of the classifier and on assumptions made about speech, many feature extraction techniques have been developed. The most widely used methods are based on smoothed estimates of the short-time stationary spectral magnitude of speech. These include linear predictive cepstral coefficients (LPCC), perceptual linear predictive coefficients (PLP) and Mel-frequency cepstral coefficients (MFCC) [24]. The MFCC was the feature analysis algorithm of choice for our experiments

2.1.1 Overview of Mel-Frequency Cepstral Coefficients

The Mel-frequency cepstral coefficients (MFCC) feature extraction technique is currently one of the most widely used in ASR systems. The procedure starts by breaking up the signal into short (e.g. 25ms) frames, windowing them and calculating the magnitude of the short-time Fourier transform of each. Then, each frame goes through a filterbank consisting of a set of triangular weighting functions in the spectral magnitude domain. The center and cut-off frequencies of the filters are uniformly spaced according to a non-linear scale. The non-linear frequency scale used here is an approximation to the Mel-frequency scale which is approximately linear for frequencies below 1kHz and logarithmic for frequencies above 1kHz [16]. This is motivated by the fact that the human auditory system becomes less frequency-selective as frequency increases above 1kHz [24].

The MFCC features correspond to the cepstrum of the log filterbank energies. To calculate them, the log energy is first computed from the filterbank outputs as

$$S_t[m] = \ln\left(\sum_{n=0}^{N-1} |X_t[n]|^2 H_m[n]\right) \qquad 0 \le m < M, \tag{2.1}$$

where $X_t[n]$ is the discrete Fourier transform of the tth input speech frame, $H_m[n]$ is the frequency response of mth filter in the filterbank, N is the window size of the transform and M is the total number of filters. Then, the discrete cosine transform (DCT) of the log energies is computed as

$$\vec{c}_t[m] = \sum_{n=0}^{M-1} S_t[n] \cos\left(\pi m \left(\frac{n-0.5}{M}\right)\right) \qquad 0 \le m < M. \tag{2.2}$$

Since the human auditory system is sensitive to time evolution of the spectral content of the signal, an effort is often made to include the extraction of this information as part of feature analysis. As such, in order to capture the changes in the coefficients over time, first and second difference coefficients are computed as

$$\Delta\vec{c}_t = \vec{c}_{t+2} - \vec{c}_{t-2} \tag{2.3}$$

$$\Delta\Delta\vec{c}_t = \Delta\vec{c}_{t+1} - \Delta\vec{c}_{t-1} \tag{2.4}$$

respectively. These dynamic coefficients are then concatenated with the static coefficients $\vec{c}_k$ according to

$$\vec{x}_t = [\vec{c}_t \quad \Delta\vec{c}_t \quad \Delta\Delta\vec{c}_t]^T, \tag{2.5}$$

making up the final output of feature analysis representing the tth speech frame.

2.2 Hidden Markov Model-Based Speech Recognition

Hidden Markov model (HMM) [9] is a statistical modelling tool which finds widespread use in speech recognition systems. It is defined as a discrete first order Markov chain where the output of each state is a continuous or discrete-valued random variable with a corresponding probability density function. Continuous density hidden Markov models (CDHMMs) were used in this thesis.

As part of the acoustic model of a given ASR system, an HMM is trained for each recognizable phonetic or word unit. Then, based on the specific ASR application, a network or a similar structure consisting of HMMs is created which determines the allowable sequence of the phonetic or word units. Such a network comprises the language model of the ASR system. In order to perform recognition, the Viterbi algorithm is used to find the path though this network which is most likely for the observed speech. To perform training of individual HMMs, the Baum-Welch or the segmental K-means algorithm is used. These algorithms are discussed in more detail below.

2.2.1 Algorithm Description

Consider a hidden Markov model defined by the triplet $\lambda = (\pi, A, B)$, consisting of states described by state indices $\{q_j\}_{j=1}^{S}$, where S is the number of states. In this definition,

$\pi = [\pi_1, \pi_2, \ldots, \pi_S]$ is the vector of initial state probabilities, $A = [a_{i,j}]$ is an $S \times S$ state transition matrix, where $a_{i,j}$ is the transition probability from state q_i to state q_j, and $B = \{b_i\}_{i=1}^{S}$ is the set of observation probability density functions associated with the HMM states. Also, note that $\sum_{j=1}^{S} a_{i,j} = 1$ for $i = 1, \ldots, S$ and $\sum_{j=1}^{S} \pi_j = 1$. The sequence of observation vectors $X = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T)$ generated by feature analysis is considered a manifestation of a hidden state sequence $\Gamma = (s_1, s_2, \ldots, s_T)$, where $s_t \in \{q_1, q_2, \ldots, q_S\}$ [25].

The Viterbi Algorithm

The Viterbi algorithm is a dynamic programming procedure that is used here to find the most likely HMM state sequence, given the model λ and the observation sequence X. The algorithm is implemented by finding the most likely path in a two dimensional trellis consisting of observation vectors along one dimension and HMM states along the other. This is performed according to an inductive approach given by the equation

$$\phi_j(t) = \max_i \{\phi_i(t-1)a_{i,j}\}\, b_j(\vec{x}_t), \tag{2.6}$$

where $\phi_j(t)$ is the probability of the most likely state sequence, which has generated the observation sequence $(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_t)$, terminating in HMM state q_j at time t, and $b_j(\vec{x}_t)$ is the probability of the observation x_t in state q_j. The best path for the entire observation sequence X is the one corresponding to the highest likelihood $\max_j \phi_j(T)$ [24].

The Baum-Welch and the Segmental K-Means Algorithms

Given a model λ, the probability of generating an observation sequence X is given by

$$f(X|\lambda) = \sum_{\Gamma} \pi_{s_1} b_{s_1}(\vec{x}_1) \prod_{t=2}^{T} a_{s_{t-1},s_t} b_{s_t}(\vec{x}_t), \tag{2.7}$$

where the summation is over all possible state sequences Γ. Now, the goal of maximum likelihood estimation is to maximize $f(X|\lambda)$ over all parameters λ for a given observation sequence X. This can be done using an efficient iterative approach referred to as the Baum-Welch algorithm, which guarantees a monotonic increase in the likelihood function [25].

One drawback of the using Equation 2.7 as the optimization criterion is that it requires considering all state transition paths in the likelihood calculation. In addition, given that the state dependent observation probabilities b_i vary over a very large dynamic range, evaluation of the likelihood over every possible path will inevitably run into numerical difficulties [25]. Therefore, the segmental K-means algorithm has been devised, which is based on a different likelihood function, given by

$$\max_{\Gamma} f(X, \Gamma|\lambda) = \max_{\Gamma} \pi_{s_1} b_{s_1}(\vec{x}_1) \prod_{t=2}^{T} a_{s_{t-1},s_t} b_{s_t}(\vec{x}_t), \tag{2.8}$$

which only focuses on the most likely state sequence. The approach consists of the iteration of two steps: the segmentation step and the optimization step [25].

In the segmentation step, given the current model and the observation sequence, the most likely state sequence is found. This can be done using the Viterbi algorithm. In the optimization step, a new set of model parameters are estimated to maximize the likelihood function given in Equation 2.8. These steps are then repeated until the likelihood converges within a certain threshold [25].

2.2.2 General Considerations

For the purpose of this thesis, left-ro-right HMMs are used, which means that

$$a_{i,j} = 0 \qquad if \qquad (j < i \quad \vee \quad j > i + 1). \tag{2.9}$$

In other words, a transition can only take place from a state to itself or to the immediately following state. This topology serves to capture the existence of quasi-stationary segments in the non-stationary speech signal [24]. Furthermore, observation densities are mixtures of multivariate Gaussians defined as [24]

$$b_j(\vec{x}_t) = \sum_{k=1}^{G} w_{jk} N(\vec{x}_t, \vec{\mu}_{jk}, \vec{\Sigma}_{jk}) \tag{2.10}$$

where $N(\vec{x}_t, \vec{\mu}_{jk}, \vec{\Sigma}_{jk})$ denotes a single multivariate Gaussian density function for state q_j with mean $\vec{\mu}_{jk}$ and covariance matrix $\vec{\Sigma}_{jk}$, G denotes the number of mixtures, and w_{jk} is the weight for the kth mixture component in state q_j satisfying

$$\sum_{k=1}^{G} w_{jk} = 1. \tag{2.11}$$

In order to greatly reduce the computational complexity and to reduce the amount of data required in training, the covariance matrices $\vec{\Sigma}_{jk}$ are assumed to be diagonal. This is referred to as the diagonal modelling assumption.

2.3 Environment Compensation

As stated previously, speech variabilities caused by acoustic environment and communications channels present a major source of ASR performance degradation. In Section 1.1 we presented an overview of the major techniques used for environment compensation. In particular, the second category of the techniques we mentioned included procedures aimed at reducing the effects of noise in the feature-space [5, 6]. In our experiments, we used one such technique, namely the WI008 European Telecommunications Standards Institute advanced front-end (ETSI-AFE) [6].

The following techniques are employed by the ETSI-AFE to increase noise-robustness [15]:

1. Noise reduction: Wiener filters are often used in speech applications to perform noise reduction. In this implementation, a *two-stage* Wiener filter design is used, where the output of the first stage is used as input for the second stage. Since the inputs to the second stage have higher SNR than the first, we gain more flexibility in filter design by making different decisions based on the SNR at each stage. This non-linear behavior is difficult to obtain in a single-stage design [26].

2. SNR-dependent waveform processing (SWP): This technique takes advantage of the quasi periodicity observed in voiced speech resulting from glottal excitation [7]. Since, by contrast, the noise energy is relatively constant in an interval corresponding to the fundamental pitch, the SNR is variable. Therefore, SWP increases the effective SNR by emphasizing the amplitude of the high SNR portions of the waveform while de-emphasizing the low SNR portions [26].

3. Blind equalization: This technique relies on an adaptive filter to reduce the convolutional distortion resulting from the mismatch between the acoustic environments

where training and testing are performed. The Least Mean Square (LMS) algorithm is used to minimize the mean square error between the current cepstrum and the cepstrum of a flat spectrum [26]. It is worth mentioning that this same result can also be achieved through cepstral mean normalization (CMN).

In this front end, using the Wiener filter scheme described above, noise reduction is first performed on the speech signal. Next, SNR-dependent waveform processing is applied to the de-noised signal. Cepstrum coefficients are then extracted from the resulting signal in the same manner as for the standard MFCCs (see Section 2.1.1.) Finally, blind equalization is performed on the resulting cepstral features [26]. This procedure has been shown to produce substantial improvements over the standard MFCC procedure. For example, in [26], experiments performed on a telephone-based connected digit database show a 52% improvement averaged over various noise conditions.

2.4 Speaker Normalization

Physiological and dialectical differences among different speakers are a major source of performance degradation in ASR systems. Evidence of this can be found by comparing the performance of speaker dependent (SD) systems, which are trained and evaluated on data from a single speaker, and speaker independent (SI) ASR systems, which are trained and evaluated on data from a large population of different speakers. For example, on the DARPA Resource Management Task, Huang and Lee reported a word error rate of 4.3% using a SI system, while, using a SD system on the same task, the error rate was only 1.4% [11]. As this is typical of many SI systems where training and testing are performed using utterances from many differing speakers, it is clear that there is a need for techniques that make ASR systems more robust against differences among speaker characteristics [27]. As mentioned in Section 1.2, these techniques are referred to as speaker normalization and speaker adaptation procedures. In order to experiment with the effects of robustness against speaker variabilities, this thesis focuses on speaker normalization techniques.

The effectiveness of many speaker normalization techniques can be attributed to the fact that they attempt to model some physiological variability in the human speech production apparatus [14]. To this end, the length of the vocal tract has been identified as one such source of variability. While the actual shape of the vocal tract is a crucial source of

phonetic information, the length does not carry any such information. However, when identical sounds are uttered by different speakers, inter-speaker variations in vocal tract length result in discrepancies among the positions of the formants produced by different speakers. In fact, it can be shown that the length of the vocal tract is inversely proportional to the position of the formant peaks [14]. Hence, given that the vocal tract length in human population can range from 13cm for females to 18cm for males, the formant frequencies can differ by up to 25%. [14]

Since ASR features are based on the spectral envelope of short-time segments of speech, discrepancies among the positions of formants produced by different speakers used for training cause an increase in the variance within each of the phonetic classes. This in turn causes overlap in the acoustic models trained for each phonetic class, as well as the occurrence of speakers who are *statistical outliers*. This phenomenon occurs when a given test speaker is not well represented by the acoustic models that have been trained from the population of speakers in the training data.

Various speaker normalization techniques have been developed which are based on performing a linear scaling of the frequency axis of speech utterances. The underlying theory maintains that if we could normalize the vocal tract of all speakers to have the same length, formant position variations due to vocal tract length differences would be reduced. As a result, the variance within each phonetic class is decreased, reducing the problems associated with class overlap and statistical outliers. Two such techniques were employed for the purpose of our experiments: vocal tract length normalization (VTLN) and augmented state-space acoustic decoder (MATE). The following will describe each of these techniques in more detail.

2.4.1 Vocal Tract Length Normalization (VTLN)

Vocal tract length normalization [14] is a speaker normalization technique which compensates for long-term average mismatch between an utterance from a test speaker and the acoustic model used for ASR. This is done by applying a warping function to the frequency axis of the test utterance. In the implementation described in [14], the warping function takes the form of a linear warping factor α, which is used to scale the frequency axis of the utterance.

To describe the procedure for estimating α, we define the following notation. Let

$S_t(\omega)$ represent the frequency domain representation of the tth speech frame. Let $\vec{x}_t$ be the corresponding MFCC feature vector as described in Section 2.1.1, and let the entire utterance be represented by $X = \{\vec{x}_t\}_{t=1}^{T}$. Finally, if $S_t^{\alpha}(\omega) = S_t(\alpha\omega)$ is the frequency domain representation of the tth speech frame of the warped utterance, then $\vec{x}_t^{\alpha}$ is the corresponding warped feature vector and $X^{\alpha} = \{\vec{x}_t^{\alpha}\}_{t=1}^{T}$.

As illustrated in Figure 2.1, the procedure starts by performing a first recognition pass to obtain a preliminary word transcription W_{pre} for the unwarped utterance X using a reference model λ. This model is obtained by performing Baum-Welch training [24] on the unwarped training data. The optimum warping factor $\hat{\alpha}$ is then obtained according to

$$\hat{\alpha} = \arg\max_{\alpha} \Pr(X^{\alpha}|\lambda, W). \tag{2.12}$$

In other words, the optimum warping factor $\hat{\alpha}$ is the one that, given the reference model λ and preliminary transcription W_{pre}, maximizes the likelihood of the warped utterance X^{α}. Since it is difficult to obtain a closed-form solution for the above equation, VTLN evaluates the above likelihood for an ensemble of N warping factors over a range corresponding to 12% compression and 12% expansion of the frequency axis, and chooses the warping factor yielding the maximum likelihood. Having obtained the optimum warping factor, a second recognition pass is then performed to obtain the final transcription W_{final}.

An iterative approach can be used to further train the reference model using warped utterances. To this end, for each utterance in the training set, a warping factor is selected using the procedure outlined above, the utterance is warped accordingly and the current model is retrained using the warped utterance. In [14], convergence properties of iterating this process are studied. Its is shown that, on a telephone-based connected digit database, while the average likelihood of the training data increases with each iteration, the word error rate for the test data does not increase beyond the first iteration. As depicted in Figure 2.1, while the reference model λ is used in the first recognition pass, the retrained model λ' is used in the warping factor estimation process, as well as the second recognition pass.

It is worth noting that, in the implementation depicted in Figure 2.1, the actual warping of the frequency axis is performed as part of the feature analysis. This is achieved by varying the spacing and width of the component filters of the filterbank front-end [14]. This procedure is encapsulated in the box marked *Feature Analysis (Freq. Warp)* in Figure

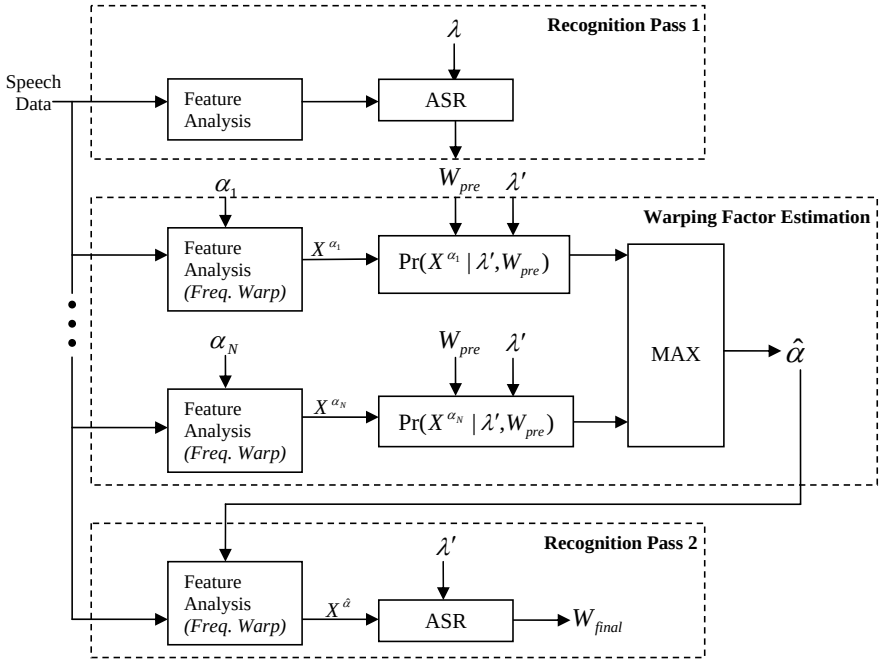


Figure 2.1 ASR using speaker normalization

2.1.

The performance of this procedure was evaluated on a telephone-based connected digit recognition task [14]. It was found that performing VTLN reduced the Word Error Rate (WER) from 3.4% for the baseline system to 2.7% using VTLN. This corresponds to a WER reduction of approximately 20%.

2.4.2 Augmented State-Space Acoustic Decoder (MATE)

Whereas VTLN tries to remove *average* variability throughout a *whole* utterance, MATE attempts to capture and remove frame-level variability. This is achieved by using a modified search algorithm that estimates an optimum linear warping factor for each analysis frame [15]. This search algorithm is implemented using a modified Viterbi decoder in an augmented state-space. The original Viterbi algorithm was described in Section 2.2.1.

MATE augments the search-space of the Viterbi algorithm such that the state-space is expanded to include the discrete ensemble of warping functions $\{\alpha_n\}_{n=1}^{N}$. The modified trellis consists of warped observation vectors $\{x_t^{\alpha_n}\}_{t=1,n=1}^{T,N}$ and states $\{q_j^n\}_{j=1,n=1}^{S,N}$. Each state q_j^n in the augmented state-space corresponds to the HMM state index j and the

frequency warping index n. Hence, given a reference model λ, the optimum sequence of states is identified using a modified version of Equation 2.6 according to

$$\phi_{j,n}(t) = \max_{i,m} \left\{ \phi_{j,m}(t-1) a_{i,j}^{m,n} \right\} b_j(x_t^{\alpha_n}), \tag{2.13}$$

where $\phi_{j,n}(t)$ is the likelihood of the optimum path terminating at state q_j^n at time t and $a_{i,j}^{m,n}$ is the transition probability from the combined HMM/warping state q_i^m to q_j^n [15]. Note that in Equation 2.13, the observation probability density function $b_j(x_t^{\alpha_n})$ is not dependent on the warping factor α_n. This is because in the current implementation of the MATE decoder, the observation densities are tied so that the parameters for $b_j(x_t^{\alpha_n})$ are shared for all warping factors α_n.

By setting a subset of the transition probabilities to zero, we can limit the search-space of the algorithm in order to reduce its complexity. For example, by imposing the condition

$$a_{i,j}^{m,n} = 0, \quad if \quad |m-n| > 1, \tag{2.14}$$

we can ensure that the degree of frequency warping applied to adjacent frames does not differ significantly. This is a physiologically motivated constraint, which is applied in all of the MATE implementations described in this thesis. Note that the above constraint is applied in addition to the left-to-right state transition constraint given in Equation 2.9.

Similar to VTLN, the MATE decoder can be used to further train the reference model λ. To this end, using frame-specific warping factors selected by the MATE decoder according the Equation 2.13, the frequency axis of each training utterance is warped. Following the frequency warping of each training utterance, the reference model is retrained using the warped utterance.

It is important to note here that the strength of MATE lies in its ability to perform warping and recognition in a single pass over the utterance. By contrast, VTLN in general requires that the warping factor estimation be performed in a first recognition pass, and recognition on the warped utterance be performed in the second pass. However, any such technique that requires multiple passes over an utterance results in response latencies that may be unacceptable for many human-machine applications.

The performance of this procedure was evaluated on a telephone-based connected digit recognition task [15]. It was found that performing MATE reduced the WER from 0.90%

for the baseline system to 0.78% using MATE. This corresponds to a WER reduction of approximately 13%. However, using VTLN in the same setup yielded a WER of 0.85% which corresponds to a 6% improvement over the baseline. Therefore, for this particular ASR task, the improvement obtained using MATE was greater than VTLN by a factor of two.

2.5 Discriminant Feature-Space Transformation

In statistical pattern recognition, one approach for tackling the issues of high dimensionality is to use a linear transformation to lower the dimensionality of the data. Classically, these approaches are divided into two categories: principal component analysis (PCA) and techniques based on Fisher's linear discriminant. While PCA aims at finding the projection that *best represents* the data in a least-squares sense, the techniques based on Fisher's linear discriminant are aimed at finding the projection that *maximizes the separability* of the data [20]. As a result, these latter techniques have been widely used in the feature analysis stage of ASR systems.

As discussed in Section 1.3, a common procedure used in the feature analysis component of ASR systems includes concatenating static features and dynamic features. Static features are extracted using cepstral analysis [16], while dynamic features are computed based on derivatives of the trajectories of the cepstral parameters of the current frame. We also discussed how it is not clear if capturing the time evolution of speech frames in this manner is actually optimal for ASR.

In recent years, techniques such as linear discriminant analysis (LDA) and heteroscedastic discriminant analysis (HDA) have been used as alternatives to the standard procedure discussed above [19, 22]. To this end, for each frame, a super vector consisting of feature vectors associated with the current as well as several preceding and succeeding frames is first created. Either HDA or LDA is then used to lower the dimensionality of this high-dimensional vector while maximizing the separability between predefined classes of feature vectors. These classes may be defined as phones, HMM states, or some other arbitrarily specified notion of class [22]. Consequently, not only have we used a proper mathematical basis for extracting information about the time evolution of speech frames, but also we can expect higher performance in a new space where classes, however defined, have been maximally separated.

Having motivated the use of discriminant analysis as a preprocessing step for ASR, we will now proceed with more details regarding two of the most widely used techniques: LDA and HDA. A comparison between the two techniques will follow.

2.5.1 Algorithm Description

Consider the set $X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}$ of vectors in $\Re^P$, each of which belongs to one-and-only-one class $c \in \{c_1, c_2, ..., c_m\}$, where m is the number of classes. We refer to this set X along with the corresponding class labels as our training data. We are interested in a classification problem where, using our training data, we would like to be able to determine the class label of *any* given vector $\vec{x} \in \Re^P$. The goal of discriminant analysis is to estimate the parameters of an $M \times P$ matrix A, with $M \leq P$ to transform vectors from a P dimensional space into an M dimensional space where class discrimination is maximized. The transformation is performed according to

$$\vec{y}_i = A\vec{x}_i, \tag{2.15}$$

where x is an arbitrary vector in the source space and y is the transformed version of x.

Now assume that each class contains N_j elements and is characterized by its mean vector $\vec{\mu}_j$ and covariance matrix Σ_j with $j = 1, ..., m$. We define the following [19]:

- within-class scatter:

$$S_W = \frac{1}{N} \sum_{j=1}^{m} N_j \Sigma_j \tag{2.16}$$

- between-class scatter:

$$S_B = \frac{1}{N} \sum_{j=1}^{m} N_j \vec{\mu}_j \vec{\mu}_j^T - \bar{\mu}\bar{\mu}^T \tag{2.17}$$

where $N = \sum_{j=1}^{m} N_j$ and $\bar{\mu} = \sum_{i=1}^{n} \vec{x}_i$. We can see that the within-class scatter is a measure of the average variance of the data *within* each class, while the between-class scatter represents the average distance *between* the means of the data in each class and the global mean.

LDA

Given a transformation matrix A, LDA defines the following measure of class separability [20]:

$$J_L(A) = \frac{\left|AS_BA^T\right|}{|AS_WA^T|} \tag{2.18}$$

where we have, in the transformed space, normalized a measure of the average distance between the centroids of each class, by a measure of the average within-class variance. Therefore, the highest separability is attained where this ratio is maximized.

Fortunately, finding the transformation A_{LDA} that maximizes $J_L(A)$ has a closed-form solution. The columns of the matrix are given by the generalized eigenvectors corresponding to the largest eigenvalues in the equation [20]:

$$S_BA^i_{LDA} = \lambda_i S_W A^i_{LDA} \tag{2.19}$$

where the A^i_{LDA} are the columns of the matrix A_{LDA}. It can be shown that the value of the LDA measure of separability is proportional to the sum of the magnitude of the largest M eigenvalues of $S_W^{-1}S_B$ [28].

HDA

Given a transformation matrix A, taking into account the individual contribution of each class, HDA uses a different objective function to maximize separability [19]:

$$\prod_{j=1}^{m}\left(\frac{\left|AS_BA^T\right|}{|A\Sigma_jA^T|}\right)^{N_j} = \frac{\left|AS_BA^T\right|^N}{\prod_{j=1}^m |A\Sigma_jA^T|^{N_j}} \tag{2.20}$$

Taking the logarithm and rearranging terms we get [19]:

$$J_H(A) = \sum_{j=1}^{m} -N_j \log\left|A\Sigma_jA^T\right| + N\log\left|AS_BA^T\right| \tag{2.21}$$

Now, the objective is to maximize the above function. Unfortunately, it is not possible to find a closed-form solution for matrix A_{HDA} that maximizes $J_H(A)$, and a numerical approach is used instead.

Discussion

Brown was one of the first to apply LDA as a preprocessing step in ASR [29]. In his work, an augmented feature vector was used to take into account context information from neighboring frames (as discussed earlier in this section.) In the following years, performance gains where reported from utilizing LDA in small-vocabulary tasks, while mixed results where reported for large-vocabulary tasks [22, 23].

The major drawback of LDA is that it assumes that all classes have identical variance. In the case of the vectors generated by the feature analysis component of an ASR system, this assumption is generally not true. Therefore, HDA was proposed which does not make such an assumption [21]. The example depicted in Figure 2.2 illustrates this idea qualitatively. Part (a) of the figure shows two-dimensional data belonging to two classes whose covariance matrices are significantly different. We are interested in transforming the two-dimensional data into scalar data. Therefore, the transformation is equivalent to projecting the data points onto a straight line. The direction of the projections computed using both LDA and HDA are also shown in part (a) of the figure. Part (b) and (c) show the distribution of transformed data generated by LDA and HDA respectively. Since the area which is below the curve for both classes corresponds to a measure of classification error, we can see that HDA performs a better job in separating the data.

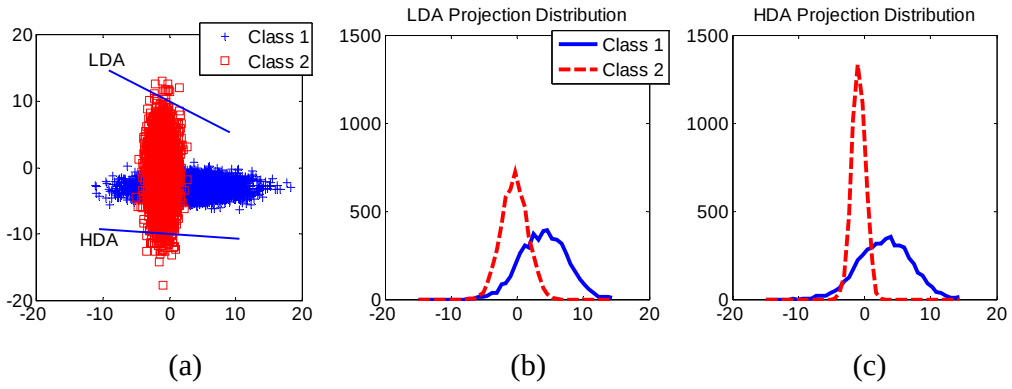


Figure 2.2 LDA/HDA comparison when covariance of classes are significantly different

Another drawback of LDA and HDA is the fact that they may produce a space where projected vectors have full covariance. Indeed, this is in contrast to the diagonal modelling

assumptions used in most ASR systems (see Section 2.2.1.) Therefore, a maximum likelihood linear transform (MLLT) can be calculated which finds an $M \times M$ transform $\hat{\psi}$, which minimizes the loss in likelihood between full and diagonal covariance models [23, 30]:

$$A_{MLLT} = \arg \max_{\psi \in \Re^{M \times M}} \sum_{j=1}^{m} -\frac{N_j}{2} \left(\log |\text{diag}(\psi \hat{\Sigma}_\text{j} \psi^\text{T})| - \log |\psi \hat{\Sigma}_\text{j} \psi^\text{T}| \right) \tag{2.22}$$

where $\hat{\Sigma}_j = A \Sigma_j A^T$, with A denoting A_{LDA} or A_{HDA} depending on whether LDA or HDA is being performed. The resulting matrix is used to perform an additional transformation following HDA. We should note here that the reason why we can use MLLT directly after HDA is the fact that the value of the HDA objective function is invariant to linear transformations in the original space [19].

2.6 Summary

The purpose of this chapter was to present the technical background necessary for the experiments contained in this thesis. In Section 2.1, we described the function of the feature analysis stage of ASR system, and presented an overview of the Mel-frequency cepstral coefficients (MFCC) feature extraction technique, which was used in our experiments. In Section 2.2, we presented an overview of the continuous density hidden Markov model (CDHMM)-based speech recognition, which our experiments were based on. Our presentation included a discussion of the various algorithms used for training and recognition in such systems. In Section 2.3, we described the MFCC-based European Telecommunications Standards Institute advanced front-end (ETSI-AFE), which was used for environment compensation in our experiments. In Section 2.4, we presented a discussion of speaker normalization techniques, describing the vocal tract length normalization (VTLN) technique and the augmented state-space decoder (MATE) in detail. Finally, in Section 2.5, we presented a description of the discriminant feature-space transformation (DFT) techniques, which are aimed at increasing the class discrimination in statistical pattern classification. In particular, we considered the linear discriminant analysis (LDA) and the heteroscedastic discriminant analysis (HDA), as well as a maximum likelihood linear transform (MLLT) aimed at diagonalizing the covariance of the feature-space.

Chapter 3

Experimental Setup

This thesis is an experimental study concerning the relationship between the techniques of environment compensation, speaker normalization and discriminant feature-space transformation. Our experiments were performed in the context of a connected digit recognition task and a continuous density hidden Markov model (CDHMM)-based ASR system. The purpose of this chapter is to describe the speech corpus that is used to define this task, and to describe the configuration of the baseline ASR system, as well as to describe how the techniques of Chapter 2 where incorporated in this system.

The chapter is organized as follows. In the first two sections we will describe the speech corpus and the underlying baseline ASR systems used in our experiments. We claimed in Section 1.4.1 that the techniques of speaker normalization and discriminant feature-space transformation (DFT) have the potential to complement one another. Hence, the last section of this chapter presents the specifics of how these techniques were combined to investigate this conjecture.

3.1 Speech Corpus

The speech corpus used in our experiments was a subset of the European Telecommunications Standards Institute (ETSI) Aurora 2 database. The database has been created by adding simulated noise samples to connected digit utterances from the TIDigits database. Eight different noise environments are simulated in Aurora 2: subway, speech babble, automobile, exhibition hall, restaurant, street, airport, and train station. Each utterance was created by adding noise from one of these categories to a corresponding clean utterance

from TIDigits. The addition of noise was done under seven signal-to-noise ratio (SNR) assumptions, namely -5dB, 0dB, +5dB, +10dB, +15dB and +20dB [31].

Aurora 2 provides two standard training sets, each containing a total of 8440 utterances from 55 males and 55 females. The clean training set consists of utterances recorded in a quiet acoustic environment. A model trained with this set yields the highest performance when testing with clean speech. However, when testing with noisy speech, the mismatch between the model and the noisy speech causes a decrease in the recognition performance. In order to reduce this mismatch, the multi-condition training set can be used. This set consists of 20 subsets where each subset corresponds to utterances distorted by one of the first four noise types mentioned above (i.e. subway, speech babble, automobile, exhibition hall) under one of the first five noise SNR assumptions mention above (i.e. clean conditions, +20dB, +15dB, +10dB and +5dB.) When trained from this set, the models capture information about the noise as well as the speech. Therefore, when testing in noisy conditions, there is less mismatch between the model and the noisy test utterances. Therefore, a higher performance can be obtained.

The subset "A" of the Aurora 2 database was used for our experiments. It consists of seven test sets corresponding to the seven SNR conditions mentioned above. Within each test set there are 1001 utterances for each of the same four noise types used for multi-condition training (i.e. subway, speech babble, automobile, exhibition hall), for a total of 4004 utterances (13159 words) per test set. These utterances were recorded from 52 males and 52 females. For the purpose of our experiments, out of the seven available test sets, we used the four test sets corresponding to the clean, +20dB, +15dB and +10dB SNR levels.

One of the issues concerned with the use of this speech corpus is the fact there are limitations on how well speech in a noisy environment can be simulated by adding noise samples to clean speech. For example, referred to as the Lombard effect, speakers usually attempt to speak more effectively and therefore differently (in terms of loudness, speed, emphasis, etc.), in noisy environments [32]. As all the utterances in Aurora 2 were recorded in a quiet environment, such effects are absent from this corpus. This poses some limitations on how realistic the obtained test results are.

Despite the mentioned limitations, the existence of various types and levels of noise have made the Aurora 2 database suitable for experiments concerned with noise-robustness in ASR. In addition, the standardized nature of this database makes it an effective tool for performing comparisons among different algorithms and techniques. For example, the

performance of various techniques incorporated in the front-end of ASR systems have been evaluated and compared on this database (see for example [7, 33].) Furthermore, these same properties make the database also suitable for evaluating distributed speech recognition (DSR) systems (see for example [26].)

3.2 ASR Platform

The ASR system used for the purpose of this thesis was based on continuous density hidden Markov models (CDHMM) with mixtures of Gaussians for the state observation probability density functions (see Section 2.2 for details.) There were a total of 11 word models, corresponding to digits "one" to "nine", as well as "oh" and "zero". These models contained 16 states with 3 Gaussians per state. The model representing silence at the beginning and end of an utterance consisted of 3 states, while a 1-state model was used to represent inter-word silence. The silence models contained 6 Gaussian mixtures per state.

As discussed in Section 2.1.1, the feature analysis stage of the baseline ASR platform was based on Mel-frequency cepstral coefficients (MFCCs). Feature analysis was performed with a window size of 25ms and a 10ms update interval. A vector of thirteen cepstral coefficients was extracted for each frame. For the baseline system, this vector was then augmented with first and second difference coefficients as discussed in Section 2.1.1 for a total of 39 features.

In order to perform environment compensation, the ETSI advanced front-end (ETSI-AFE) was used in place of the baseline front-end discussed above. As discussed in Section 2.3, this front-end is also based on MFCCs. In our system, the front-end was configured with the same window size and update interval as above, and the same procedure was used to extract 39-component feature vectors consisting of 13 cepstral features augmented with first and second difference cepstrums.

The baseline models were generated by performing 20 iterations of the maximum likelihood segmental K-means algorithm, as discussed in Section 2.2.1. The models used for performing recognition with VTLN were generated by performing an additional iteration of training on the baseline models. Prior to this iteration, the frequency axis of each utterance in the training set was warped using a maximum likelihood warping factor, as discussed in Section 2.4.1. Likewise, an additional training iteration was performed on the baseline models to generate the models used for performing recognition based on the MATE de-

coder. Prior to this iteration, the frequency axis of each training utterance was warped using frame-specific warping factors selected by the MATE decoder, as discussed in Section 2.4.2.

The VTLN procedure was configured with an ensemble of 11 warping factors equally spaced along a range from a minimum of 12 percent compression and a maximum of 12 percent expansion of the frequency axis. MATE on the other hand was configured with an ensemble of 5 equally-spaced warping factors covering the same range. This configuration matches the one used in [15] and [28].

3.3 Combining Discriminant Feature-Space Transformation with Speaker Normalization

The techniques of discriminant feature-space transformation (DFT) and speaker normalization were described individually in Sections 2.5 and 2.4 respectively. Regarding these techniques, we stated two claims in Section 1.4.1 which were inversely related. On the one hand, increasing class discrimination through DFT should improve the performance of speaker normalization techniques. On the other hand, reducing within-class variance of the data through speaker normalization should improve the degree of class discrimination gained though DFT. In this section we will describe how the two techniques were combined for the purpose of evaluating these claims experimentally.

In our system, heteroscedastic discriminant analysis (HDA) was used to compute a transform aimed at maximizing class separability. Then, a maximum likelihood linear transformation (MLLT) was computed to minimize the loss in likelihood between full and diagonal covariance models. In Section 3.3.1 below, we will describe how these transformations were estimated from the training data. Applying these transforms in a system incorporating speaker normalization will allow us to evaluate the effects of DFT on speaker normalization. In Section 3.3.2, we will describe how the transformations were applied in such a system.

Conversely, in order to evaluate the effects of speaker normalization on DFT, we need to estimate the HDA transformation from *speaker normalized* training data. To this end, speaker normalization is performed on the training data prior to estimating the transform parameters. Section 3.3.3 describes this procedure. The procedure used for applying the transforms is the same as when the transforms are estimated from unnormalized data,

which is given in Section 3.3.2.

3.3.1 Estimating the Transforms

Figure 3.1 shows an overall view of the transform estimation process. First, feature analysis generates 13-dimensional observation vectors $\vec{x}_t$ consisting of cepstral coefficients as discussed in Section 2.1.1. We refer to the feature-space defined by these vectors as the *Original Space*. Then, HDA is performed to compute a matrix, A_{HDA}, to transform these observation vectors into 39-dimensional vectors, $\vec{x}_t^{HDA}$. These vectors are in the *HDA Space*, where classes are, according to the criterion given in Equation 2.21, maximally separated. Next, a MLLT, A_{MLLT}, is computed which transforms the $\vec{x}_t^{HDA}$ into another set of 39-dimensional vectors, $\vec{x}_t^{HDA/MLLT}$. These are in the *HDA/MLLT Space*, where in addition to maximal class separation, the resulting vectors are distributed according to a diagonal Gaussian distribution. Below, we will provide more details on this procedure.

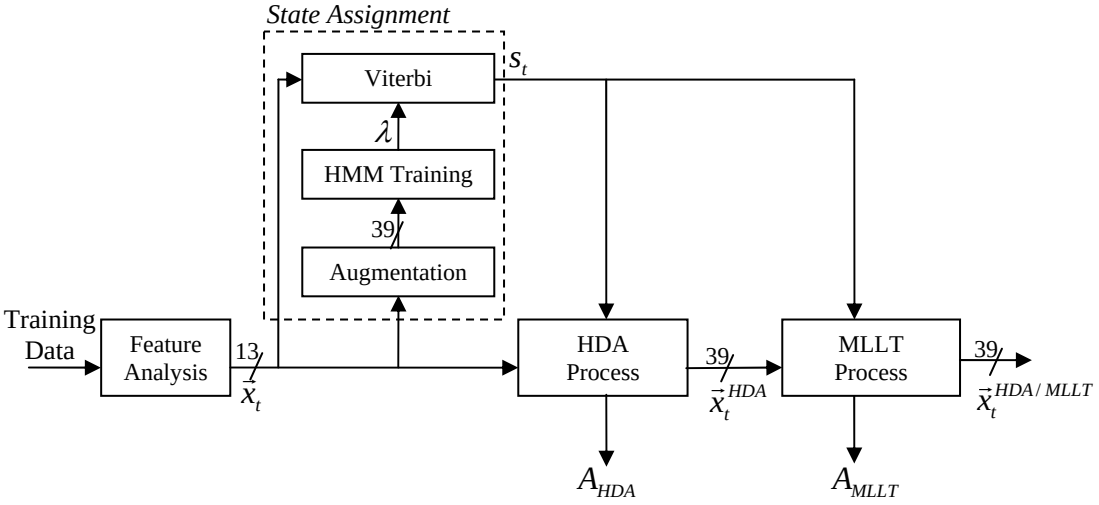


Figure 3.1 Overview of the DFT process

Defining Classes

In order to perform HDA, it is necessary to assign the observation vectors $\vec{x}_t$ to a number of classes to be maximally separated. In our experiments, we have adopted the use of individual states of HMMs as classes [28]. That is, for a sequence of observation vectors

$X = (\vec{x}_0, \vec{x}_1, \ldots, \vec{x}_T)$, given the model λ, the most likely state sequence $s = (s_0, s_1, \ldots, s_T)$ determines the state assignment, and therefore, the class assignment of each observation vector in this sequence. The Viterbi algorithm, as described in Section 2.2, is used to find this most likely state sequence.

In Figure 3.1, the procedures necessary for generating the state assignment of the observation data are marked as *State Assignment*. The model λ is trained from observation vectors $\vec{x}_t$ augmented by first and second difference coefficients (as discussed in Section 2.1.1.) Given this model, the Viterbi algorithm is then used to generate the state indices s_t corresponding to observation vectors $\vec{x}_t$.

Performing Heteroscedastic Discriminant Analysis (HDA)

To further explain the DFT estimation process, Figure 3.2 illustrates an expanded view of the *HDA Process* box of Figure 3.1. In this process, each observation vector is first concatenated with four preceding and four succeeding vectors to form a 117-dimensional "super vector". This is further illustrated in Figure 3.3. These "super vectors" are then classified according to the state assignments generated by the Viterbi algorithm as discussed above. Given that, as stated in Section 3.2, there are eleven 16-state models, one 3-state model and one 1-state model, there are a total of 180 HMM states, and therefore 180 classes.

Having classified the 117-dimensional vectors computed from the training data into 180 classes, a 117 by 39 matrix is estimated which maximizes the HDA measure of class separability as given in Equation 2.21. This matrix is then used to transform the data back to a 39-dimensional space.

Note that the target dimension 39 was selected for convenience, since implementation issues, such as dealing with the wide dynamic range of computed likelihoods, had already been addressed in the baseline system for a 39-dimensional space. In general, the dimensionality of the discriminant transformation is determined by the rank of the target feature-space. For LDA, this rank may be determined by observing the magnitude of the eigenvalues given in Equation 2.19. It is more typical to determine the rank empirically on a development test set. This is a potential topic for future work.

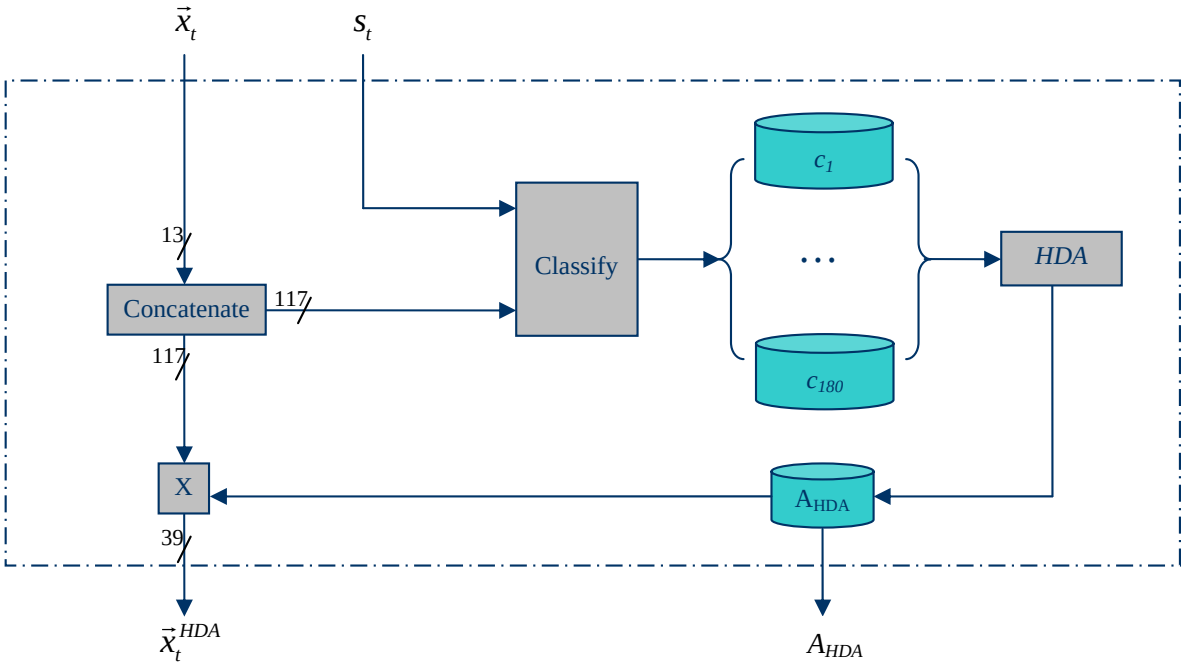


Figure 3.2 The HDA process: matrix computation and feature-space transformation

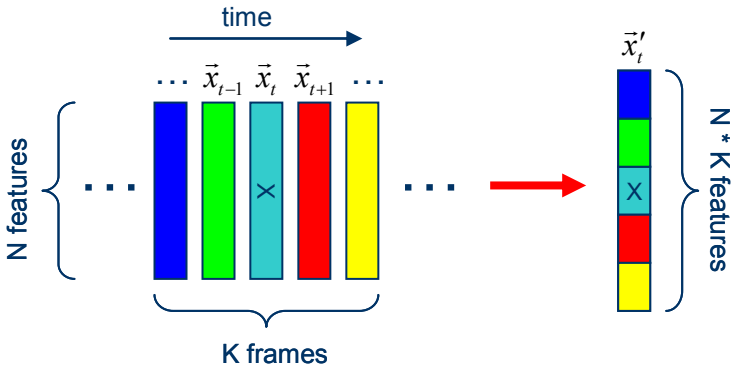


Figure 3.3 The feature vector concatenation process

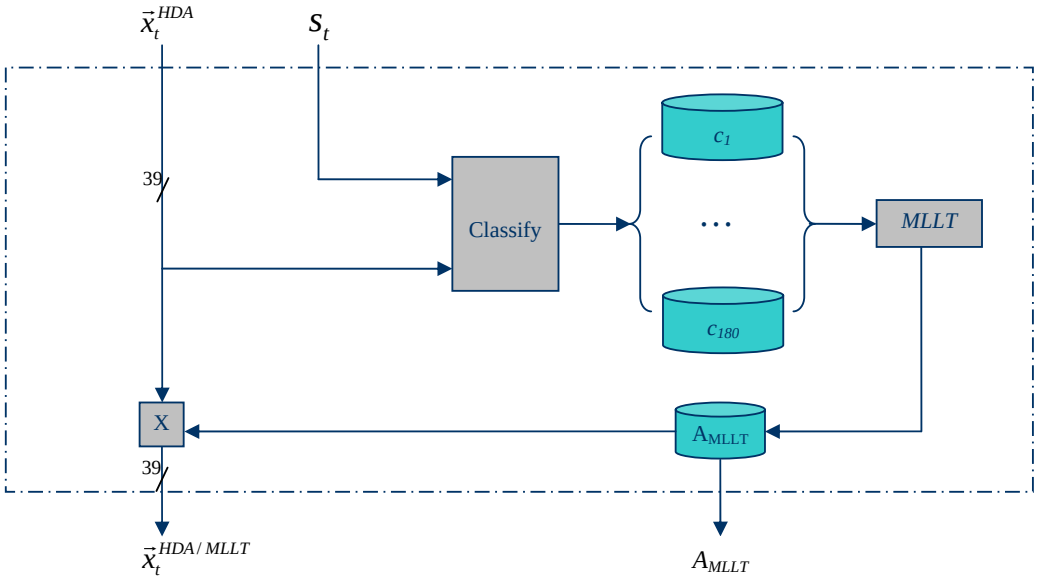


Figure 3.4 The MLLT process: matrix computation and feature-space transformation

Performing Maximum Likelihood Linear Transformation (MLLT)

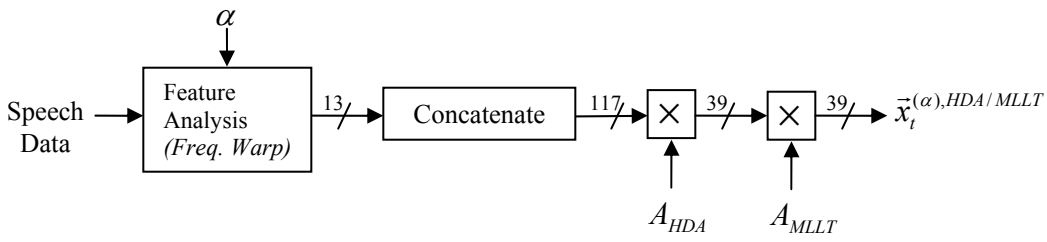
The MLLT computation stage is very much similar to the HDA stage. This is shown in Figure 3.4 as an expanded view of the *MLLT Process* box of Figure 3.1. Given the 39-dimensional vectors transformed using HDA, a 39 by 39 maximum likelihood linear transform (MLLT) is computed according to Equation 2.22. This matrix is then used to transform the data into a new space where the likelihood of the data with respect to the class-specific diagonal Gaussian covariance models is maximized.

Incorporating Environment Compensation

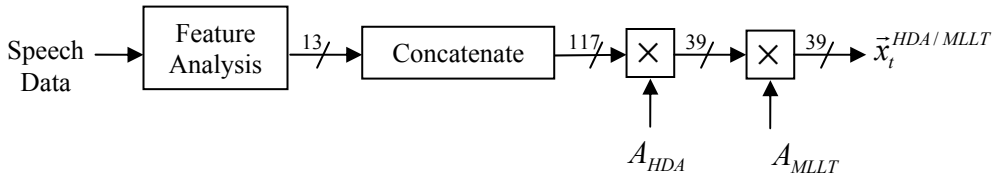
In Section 1.4.1 we argued that reducing the within-class variance using environment compensation should improve the performance of DFT. In order to evaluate this claim, we need to be able to perform DFT both with and without environment compensation. Therefore, in Figure 3.1, when environment compensation is used, the *Feature Analysis* box represents the noise robust ETSI advanced front-end (ETSI-AFE) (described in Section 2.3); otherwise, it refers to the baseline MFCC-based front-end (described in Section 2.1.1)

3.3.2 Applying the Transforms

In order to perform speaker normalization in conjunction with DFT, we simply perform warping inside the feature analysis component prior to applying the transformation. Section 2.4.1 describes how the actual frequency warping is performed as part of feature analysis. Figure 3.5(a) shows how a given speech waveform can be warped using a given warping factor, and subsequently transformed. On the other hand, when speaker normalization is not used, the HDA matrix and the MLLT are simply applied to the observation vectors generated by the feature analysis component. This process is depicted in Figure 3.5(b).



(a) With frequency warping



(b) Without frequency warping

Figure 3.5 Applying the discriminant feature-space transforms

Going back to Figure 2.1 in Chapter 2, we can see that VTLN consists of three stages: recognition pass one, warping factor estimation and recognition pass two. Figure 3.6 illustrates how each of these was modified to include matrix transformations. Here, the *DFT* box corresponds to the process outlined in Figure 3.5(b), while the *DFT (Freq. Warp)* box corresponds to Figure 3.5(a). Also, trained in the transformed space, $\lambda^{HDA/MLLT}$ and $\lambda^{HDA/MLLT'}$ correspond to the baseline model λ and retrained model λ' in Figure 2.1.

Much like the warping factor estimation process in VTLN shown in Figure 3.6, our implementation of MATE consists of performing frequency warping and transformations for an ensemble of warping factors and storing the results. At each stage of the modified

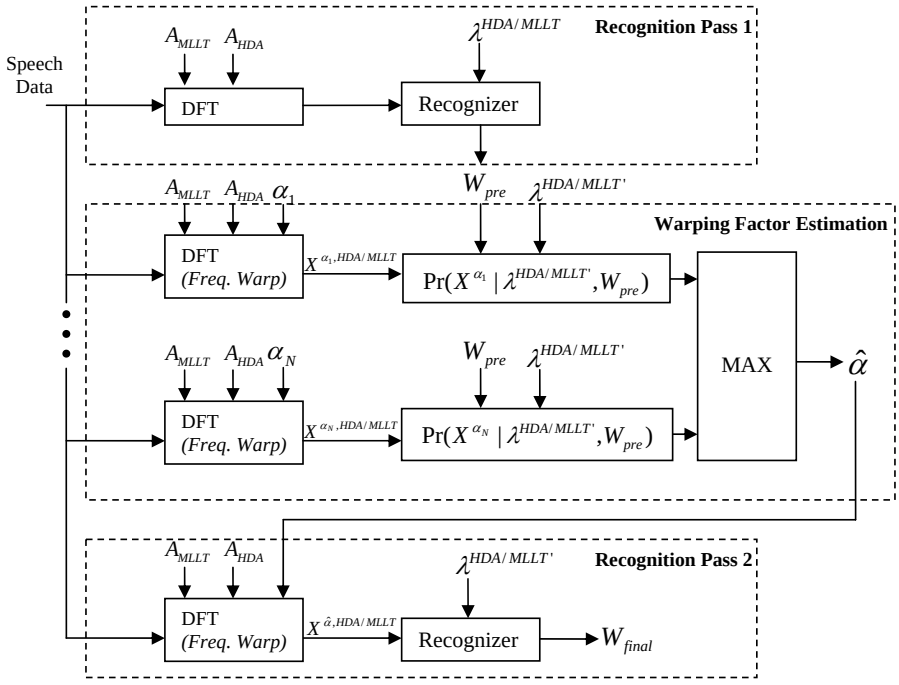


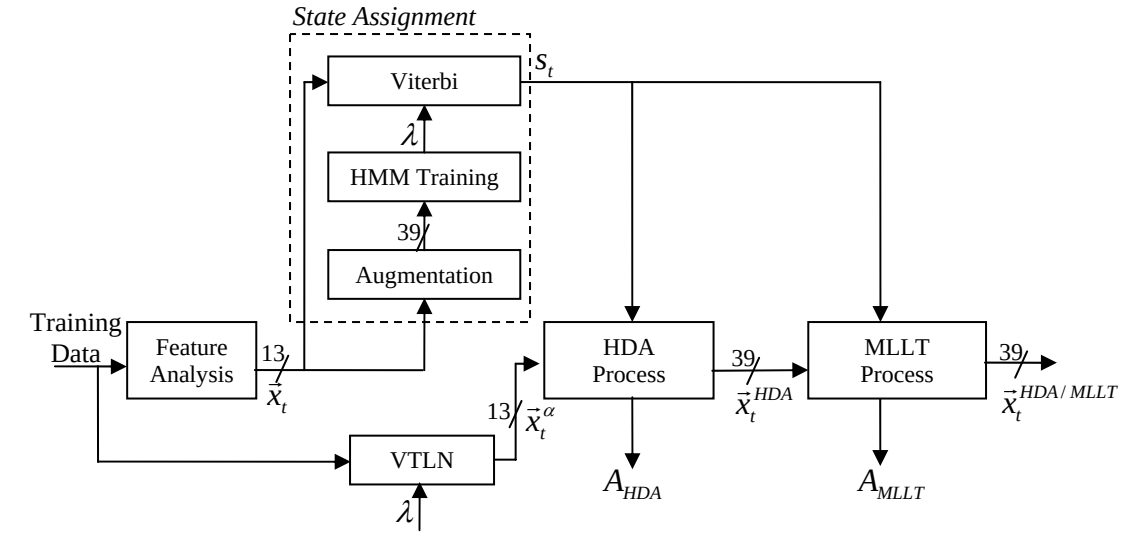
Figure 3.6 Combining speaker normalization with DFT

Viterbi algorithm, the MATE decoder then obtains the frames with the required degree of warping from this stored ensemble of warped and transformed vectors.

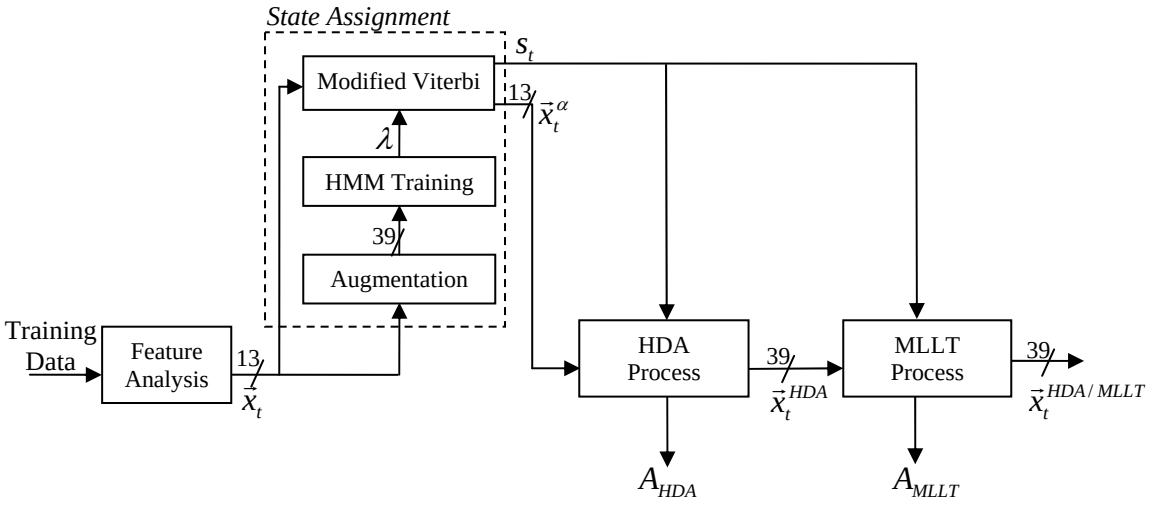
3.3.3 Estimating the Transforms from Speaker Normalized Data

Figure 3.7 shows an overall view of the process of estimating the HDA matrix from speaker-normalized data. As stated in Section 1.4.1, we expect speaker normalization to reduce the variability due to differences in speaker characteristics within each of the classes used to estimate the HDA matrix. This will in turn allow HDA to yield higher class discrimination. As depicted in Figure 3.7(a), when using VTLN to normalize the data, each training utterance is warped to maximize its likelihood given the model λ. In the case of MATE, as depicted in Figure 3.7(b), the modified Viterbi decoder described in Section 2.4.2 is used to select maximum likelihood warping factors for individual frames, as well as determining the corresponding state assignments s_t.

Using the warped utterances, concatenation and classification are performed in the exact same manner as described in Section 3.3.1. Given the warped classes, the procedure used for computing the HDA matrix as well as the MLLT is also the same as previously



(a) Using VTLN



(b) Using MATE

Figure 3.7 Estimating DFT parameters from speaker-normalized data

described. The only difference is that classes now contain warped data, as well as the fact that, in the case of performing speaker normalization using MATE, the state assignment of the observation data is determined by the modified Viterbi algorithm.

3.4 Summary

This chapter described the experimental setup used for investigating the claims of this thesis, as stated in Section 1.4.1. In Section 3.1, we described the ETSI Aurora 2 connected digit speech corpus used in our experiments, while in Section 3.2, we described our baseline continuous density hidden Markov model (CDHMM)-based ASR system. Finally, in Section 3.3, we described how the speaker normalization techniques of VTLN and MATE where employed in a discriminant feature-space. To this end, we showed how heteroscedastic discriminant analysis (HDA) was used to estimate a transform aimed at increasing class discrimination, where classes corresponded to HMM states. We also showed how a maximum likelihood linear transform (MLLT) was estimated to diagonalize the covariance of the feature vectors generated by HDA. Subsequently, we showed how these transforms could be applied in conjunction with the speaker normalization techniques of VTLN and MATE. Our description also included how the parameters of the discriminant feature-space transform could be estimated from data which was normalized using environment compensation as well as the VTLN and MATE speaker normalization techniques.

Chapter 4

Experimental Analysis

This chapter presents the experimental study that was performed to investigate the interaction between techniques that reduce the effects of speaker and environment variability and techniques that increase class discrimination in ASR. In Section 1.4.1, we stated the specific claims that were to be investigated as part of this thesis. In this chapter, we will explain how each claim was motivated, present the corresponding experiments, and analyze the obtained results.

We will start this chapter by discussing some general considerations regarding the evaluation of our experimental results, including a note on statistical significance. Then, we will look at the effects of environment compensation on the performance of speaker normalization techniques, namely vocal tract length normalization (VTLN) and the augmented state-space acoustic decoder (MATE). Next, we will look at how discriminant feature-space transformation (DFT) techniques can effect the performance of speaker normalization techniques. More specifically, we will look at the performance of VTLN and MATE in a discriminant feature-space generated by heteroscedastic discriminant analysis (HDA). Then, we will consider how the techniques of environment compensation and speaker normalization can improve DFT by estimating the DFT parameters from data that has been normalized using these techniques.

In evaluating the claims of this thesis, our experimental results revealed that the MATE speaker normalization technique does not perform well in noisy conditions. Hence, we devised a modification to the original decoding algorithm used by MATE. This modification was based on utilizing, during recognition, knowledge about the distribution of warping

factors selected by MATE during training. The final section of this chapter defines and presents the results obtain by this new algorithm.

4.1 Evaluation Metrics

In Section 3.1, two sets of training data were described: clean and multi-condition. For the models trained from each training set, four test sets were evaluated corresponding to the clean, 20dB, 15dB and 10dB signal-to-noise ratio (SNR) conditions. For each test set, recognition was performed in three speaker normalization modes: baseline (no speaker normalization), VTLN, and MATE. Section 3.2 describes how the models used in each of these modes were trained, while Section 3.1 contains more information regarding the training and testing data.

The measure of recognition performance used in our experiments was the word error rate (WER), define as

$$\%WER = \frac{n_I + n_S + n_D}{n_T} \times 100, \tag{4.1}$$

where n_T is the total number of words in the reference transcriptions, and n_I, n_S, and n_D are the total number of insertions, substitutions, and deletions respectively. For a given test utterance, the number of insertions, substitutions and deletions are obtained using a dynamic programming algorithm to produce the best alignment between the recognized transcription and the correct transcription.

Our analysis of experimental results involved comparing the WER results obtained from different system configurations. In order to determine how well the obtained WER improvements could be generalized, we had to quantify the statistical significance of these improvements. Suppose we are interested in showing that method B has a performance that is significantly different from method A. Also, suppose performing n independent trials, where the outcome of each trial is either an error or a success, method A yields an error rate of p_A while method B yields p_B. Given that the number of trials, n, is large enough, we can assume the number of errors to be a random variable with a Gaussian distribution, P, with mean $\mu = np_A$ and variance $\sigma^2 = np_A(1 - p_A)$. To state that method B is in fact significantly different than method A within a level of confidence c, the area under the curve P bounded by the interval $(2np_A - np_B, np_B)$ should be larger than or equal to c.

This is referred to as the Normal test of significance [24].

To use the Normal test in our experiments, we regarded the recognition of individual words as independent trials. As stated in Section 3.1, each test set consisted of 13159 words, and therefore $n = 13159$ independent trials. In the following presentation of our experimental results, all comparisons between WER results are considered taking into account the minimum relative WER reduction required at a 95% confidence level.

4.2 Improving Speaker Normalization Using Environment Compensation

Table 4.1 displays the results obtained from a system trained from MFCC features without environment compensation. The three rows of data presented correspond to the speaker normalization mode used. The numbers in square brackets next to the baseline results indicate the minimum relative WER reduction required for the improvement to be significant at a 95% confidence level. The numbers in the parentheses next to the results obtained from VTLN and MATE indicate the actual relative WER reduction.

Table 4.1 Recognition results in MFCC space (% WER)

	Clean Training				Multi-condition Training			
	Clean	20dB	15dB	10dB	Clean	20dB	15dB	10dB
Baseline	0.90[18][1]	3.83[9]	11.41[5]	28.54[3]	1.49[14]	2.09[12]	2.47[11]	4.56[8]
VTLN	0.89(2)[2]	3.47(10)	10.90(5)	27.94(2)	1.27(15)	1.88(10)	2.39(3)	4.49(1)
MATE	0.78(13)	3.66(5)	12.08(-6)	29.77(-4)	1.09(27)	1.78(15)	2.27(8)	4.57(0)

[1]Minimum relative reduction in WER required at a 95% confidence level

[2]Actual relative reduction in WER compared to baseline

The table shows that, in the case of clean training, MATE is not able to yield significant improvements over the baseline system (it even degrades performance at low SNR conditions), while VTLN only yields significant improvements in 15dB and 10dB SNR test conditions. Although these improvements are more significant in the case of multi-condition training, a common trend under both cases is that as the test SNR conditions decrease the degree of improvements obtained by speaker normalization is reduced.

The above results indicate that the presence of noise interferes with the performance of speaker normalization techniques. Therefore, in Section 4.2.3, we will present the results

obtained when environment compensation is used in conjunction with speaker normalization to counter these effects. But first, we will perform an experimental study regarding how the presence of noise effects the processes by which VTLN and MATE perform speaker normalization. More specifically, the effects of noise on the warping factor estimation process of both techniques (as described in Section 2.4) is considered in Section 4.2.1, while the effects of noise on the first recognition pass of VTLN (as described in Section 2.4.1.) is considered in Section 4.2.2.

4.2.1 The Effect on Warping Factor Estimation

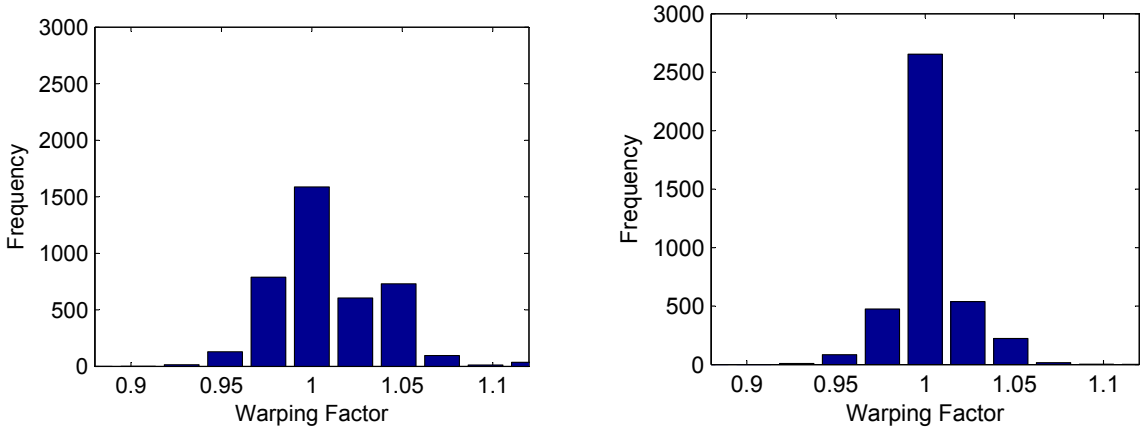
As stated in Section 2.4, both VTLN and MATE use a maximum likelihood framework to select warping factors. This involves using an HMM-based model to calculate the probability of observation vectors generated by feature analysis. Therefore, if these observation vectors are corrupted by noise, the resulting probabilities are also affected.

In the case of VTLN, in order to choose the optimal warping factor for a given utterance, the likelihood of the utterance for an ensemble of warping factors is computed and the one yielding the highest likelihood is selected. Consequently, if these likelihoods are obtained from observation probabilities computed from corrupt data, it is possible for a suboptimal warping factor to yield the highest likelihood, and therefore be selected.

In the case of MATE, the modified Viterbi algorithm propagates paths into a three dimensional trellis, which has an ensemble of warping factors along one axis, and selects the path with the highest likelihood. As in the case of VTLN, if the likelihoods are obtained from observation probabilities computed from corrupt data, it is possible for a suboptimal path containing suboptimal warping factors to be selected.

To further illustrate the effects of noise on the warping factor estimation process, we extracted the warping factors estimated by VTLN during recognition on the clean and 10dB SNR test sets. The models used in this experiment were trained from the clean training set. Figure 4.1 illustrates the distribution of these warping factors for both the clean and noisy cases.

Noting the comparison made in Figure 4.1, it is clear that the distribution of warping factors selected in clean conditions has a much wider spread than the distribution of warping factors selected in noisy conditions. It is apparent that noise causes VTLN to favor performing no or little warping of the utterance. Since VTLN selects the warping factor



(a) Clean test set (b) 10dB SNR test set

Figure 4.1 Distribution of warping factors selected by VTLN during recognition (clean training)

yielding the highest likelihood, $P(X^{\alpha}|W,\lambda)$, we have plotted these likelihoods in Figure 4.2 for a sample utterance at 12 discrete values of α ranging from 0.88 to 1.12 under a range of different SNR levels. The models used for generating these likelihoods were trained from the clean data.

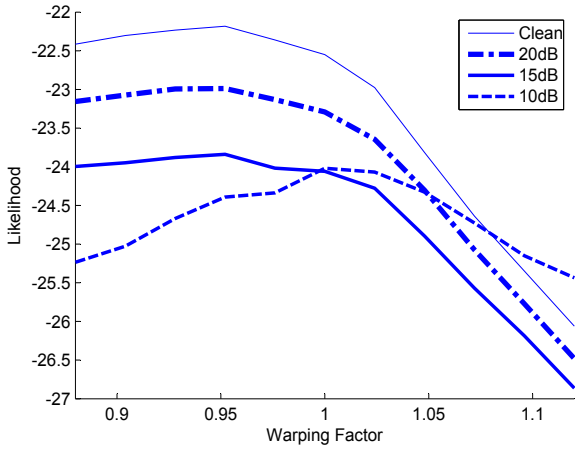


Figure 4.2 Warping factor likelihoods for various noise levels (clean training)

It can be seen that, as the level of noise increases, the likelihood of the warped utterance decreases for all warping factors. However, this decrease in likelihood is more pronounced

in the case of warping factors which correspond to applying a higher degree of compression/expansion to the frequency axis of the utterance. Therefore, in accordance with what we observed from Figure 4.2, as the level of noise increases, VTLN chooses no or little warping for this utterance. For instance, in the above figure, the peaks of the graphs, corresponding to the maximum likelihood warping factors for the clean, 20dB and 15dB cases, are close to 0.95, while the peak for the 10dB case is close to unity. The reasons behind this phenomenon can be traced back to the disproportionate mismatch between the warped data in noise with respect to the acoustic model.

4.2.2 The Effect on the First Recognition Pass of VTLN

As shown in Figure 2.1, VTLN performs a first recognition pass to obtain a preliminary word transcription for each unwarped utterance. This transcription, W_{pre} is subsequently used to calculate the likelihood, $P(X^{\alpha_i}|\lambda, W_{pre})$, for each member of the ensemble of warping factors, selecting the warping factor yielding the highest likelihood. In noisy conditions, it is only natural to expect a degradation in the accuracy of this preliminary transcription. As a result, it is likely that, based on this error-prone preliminary transcription, a non-optimal warping factor is selected. This will in turn affect the performance of ASR.

In order to illustrate this issue, we trained a model using MFCC features without environment compensation from the multi-condition training set. Then, using VTLN, we performed recognition on the clean, as well as the 15dB SNR test sets. The first column of Table 4.2 shows the Word Error Rate (WER) of the first recognition pass of VTLN. We can see that the WER is much higher in the case of noisy data.

Table 4.2 Recognition results on clean vs. noisy speech using VTLN (% WER)

Condition	Pass 1	Pass 2	Pass 2 (actual transcriptions)
Clean	1.11	0.99	0.97
15dB SNR	1.54	1.35	1.26

The second column of Table 4.2 illustrates the WER of the second recognition pass of VTLN. In order to gauge how inaccuracies in the results of the first recognition pass can effect the performance of the second pass, we performed an experiment where, instead

of using the results of the first pass, we supplied the actual correct transcriptions to the second pass. The results are given in the third column of Table 4.2. Comparing the two columns we can see that the performance of VTLN in 15dB SNR conditions would improve by 6.7% if the first recognition pass yielded completely correct results.

4.2.3 Applying Environment Compensation

Table 4.1 displays the results obtained from a system trained from the environment compensated features generated by the ETSI advanced front-end (described in Section 2.3.) We will refer to these as the environment compensated MFCC features. The three rows of data presented correspond to the speaker normalization mode used. The numbers in square brackets next to the baseline results indicate the minimum relative WER reduction required for the improvement to be significant at a 95% confidence level. The numbers in the parentheses next to the results obtained from VTLN and MATE indicate the actual relative WER reduction.

Table 4.3 Recognition results in environment compensated MFCC space (% WER)

	Clean Training				Multi-condition Training			
	Clean	20dB	15dB	10dB	Clean	20dB	15dB	10dB
Baseline	0.96[17][1]	1.95[12]	3.18[9]	7.04[6]	1.11[16]	1.54[14]	2.11[12]	3.69[9]
VTLN	0.87(10)[2]	1.76(10)	2.84(11)	6.51(8)	1.00(10)	1.35(12)	1.90(10)	3.44(7)
MATE	0.81(15)	1.47(25)	2.63(17)	6.06(14)	0.92(17)	1.28(16)	1.82(13)	3.55(4)

[1]Minimum relative reduction in WER required at a 95% confidence level

[2]Actual relative reduction in WER compared to baseline

Comparing the results in Table 4.3 with Table 4.1 we can see that the improvements obtained by MATE and VTLN are more significant when environment compensation is used. Table 4.4 illustrates this further by comparing the results from Table 4.1 and Table 4.3 for the case with clean training and 10dB SNR testing. For example, we can see in this table that MATE yields a 14% WER reduction when using environment compensation, while it actually increased the WER by 4% when environment compensation was not used.

Indeed, the above developments lead us to concluded that environment compensation does in fact improve the performance of speaker normalization techniques. To substantiate this result further, we will now examine how environment compensation alleviates the

Table 4.4 Speaker normalization improvements due to environment compensation (% WER)

	10dB SNR	10dB SNR (environment compensation)
Baseline	28.54[3][1]	7.04[6]
VTLN	27.94(2)[2]	6.51(8)
MATE	29.77(-4)	6.06(14)

[1]Minimum relative reduction in WER required at a 95% confidence level

[2]Actual relative reduction in WER compared to baseline

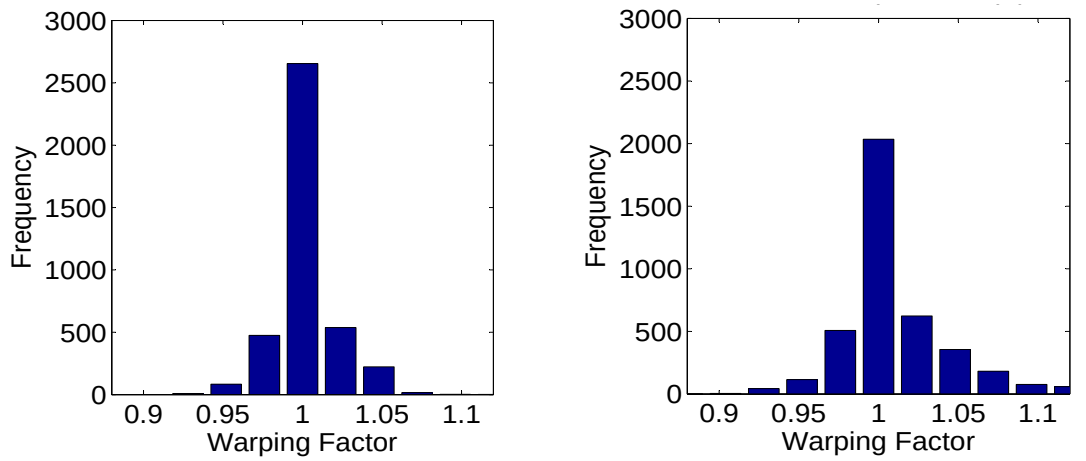
effects of noise on the processes mentioned above, namely the warping factor estimation process (Section 4.2.1) and the first recognition pass of VTLN (Section 4.2.2.)

The Effect on Warping Factor Estimation

In accordance with our analysis from Section 4.2.1, where we examined the distribution of the warping factors selected by VTLN, we now consider the same distribution in the presence of environmental compensation. Figure 4.1 compared the distribution of warping factors resulting from clean test conditions against that of 10dB SNR conditions. Now, in Figure 4.3, we examine, for the 10dB SNR conditions, how this distribution changes when environmental compensation is used. The figure suggests that when environmental compensation is used, the spread of the warping factors selected by VTLN widens. This means that the tendency of VTLN to choose no or little warping in the presence of noise, as discussed in Section 4.2.1, has been diminished. This serves as an indication of how environmental compensation can improve the performance of frequency warping-based speaker normalization techniques such as VTLN and MATE.

The Effect on the First Recognition Pass of VTLN

The above result is also confirmed by considering the performance of the first recognition pass of VTLN (as we did in Section 4.2.2.) Our experiments showed that, under multi-condition training and 15dB test conditions, when using environmental compensation, the WER obtained from the first recognition pass of VTLN was 2.1%. This shows an improvement of 15% compared with the 1.54% WER obtained when environmental compensation is not used (see Table 4.2).



(a) 10dB SNR test set

(b) 10dB SNR test set (using environment compensation)

Figure 4.3 Distribution of warping factors selected by VTLN during recognition when environment compensation is used

4.3 Combining Discriminant Feature-Space Transformation and Speaker Normalization

Section 3.3 introduced the idea of performing speaker normalization in a discriminant feature-space. The underlying theory maintained that when classes are maximally separated, speaker normalization can yield a higher performance. The details of our proposed setup for combining the two techniques were also presented. In turn, we dedicate this section to presenting the results obtained from experiments performed based on this setup.

In order to gauge the performance of our proposed scheme, as described in Section 3.3.1, we estimated the HDA transform and the MLLT based on environment compensated MFCC features. Then, in the resulting environment compensated HDA/MLLT space, we trained models based on both the clean and the multi-condition training sets. We then performed recognition using both VTLN and MATE in this space, as described in Section 3.3.2. The corresponding WER results are given in Tables 4.5 and 4.6 for VTLN and MATE respectively.

In Table 4.5, we compare the WER results obtained by performing VTLN recognition in the environment compensated HDA/MLLT space as described above, against the WER results obtained in the environment compensated MFCC space, as described in Section

4.2.3. The idea is to see how the performance of VTLN improves when it is used in a discriminant feature-space. Hence, the first row of Table 4.5 lists the results obtained in the environment compensated MFCC space (which were also given in Table 4.3), with the numbers in brackets indicating the minimum relative WER reduction required for any improvements to be significant at a 95% confidence level, while the second row contains the results obtained in the environment compensated HDA/MLLT space, with the numbers in parentheses indicating the actual WER reductions. In the case of MATE, the same comparison is performed in Table 4.6.

Table 4.5 Comparing recognition results using VTLN in environment compensated MFCC space vs. environment compensated HDA/MLLT space (% WER)

	Clean Training				Multi-condition Training			
	Clean	20dB	15dB	10dB	Clean	20dB	15dB	10dB
MFCC	0.87[18][1]	1.76[13]	2.84[10]	6.51[6]	1.00[17]	1.35[15]	1.90[12]	3.44[9]
HDA/MLLT	0.55(36)[2]	1.33(24)	2.36(17)	5.66(13)	0.90(9)	1.14(16)	1.63(14)	3.25(6)

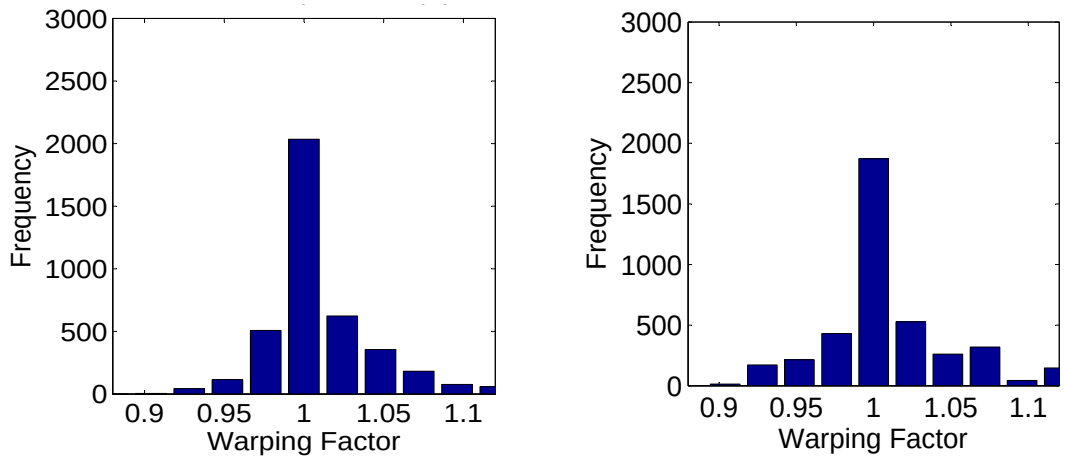
[1]Minimum relative reduction in WER required at a 95% confidence level

[2]Actual relative reduction in WER compared to environment compensated MFCC space

Analyzing VTLN

Our results for VTLN indicate significant improvements for both the clean and the multi-condition training cases. Although these improvement are less significant in the case of multi-condition training, the results lead us to conclude that increasing class discrimination through HDA improves the performance of speaker normalization using VTLN. This can also be examined by considering the distribution of warping factors selected by VTLN in the environment compensated HDA/MLLT space.

In Figure 4.4(a) and Figure 4.4(b) we have plotted the distribution of warping factors for a 10dB SNR test set in the environment compensated MFCC space and the environment compensated HDA/MLLT space respectively. In Section 4.2.1, using a similar plot we pointed out that in the presence of noise VTLN tends to perform no or little warping. Subsequently, in Section 4.2.3 we showed how the use of environment compensation can mitigates this problem. In turn, we can see in Figure 4.4 that this tendency is further diminished in the environment compensated HDA/MLLT space.



(a) 10dB SNR test set in environment compensated MFCC space

(b) 10dB SNR test set in environment compensated HDA/MLLT space

Figure 4.4 Distribution of warping factors selected by VTLN during recognition comparing MFCC and HDA/MLLT space

The Effects of Noise

In Table 4.6, we can see that results obtained for MATE are generally less promising than VTLN. Nevertheless, a common trend seen for both VTLN and MATE is that smaller improvements are obtained for lower SNR conditions. This observation is in line with the fact that the presence of noise increases the within-class variance, which in turn reduces the class discrimination achieved by HDA, degrading the overall performance of ASR. The effects of noise on class discrimination will be considered further in Section 4.4.1.

Analyzing MATE

Given that Table 4.6 shows how MATE generally did not produce significant improvements in the environment compensated HDA/MLLT space for low SNR conditions, we analyzed the partitioning of the resulting WER into insertions, deletions and substations. Through this analysis we realized that, compared to when MATE is performed in MFCC space, the resulting insertion rate increases significantly, while the deletion and substitution rates are reduced. This is illustrated in Table 4.7, which compares the resulting insertion rates for the environment compensated MFCC and HDA/MLLT cases.

Although, as shown in Table 4.6, performing MATE on clean test data seems to improve

Table 4.6 Comparing recognition results using MATE in environment compensated MFCC space vs. environment compensated HDA/MLLT space

	Clean Training				Multi-condition Training			
	Clean	20dB	15dB	10dB	Clean	20dB	15dB	10dB
MFCC	0.81[19][1]	1.47[14]	2.63[10]	6.06[7]	0.92[18]	1.28[15]	1.82[13]	3.55[9]
HDA/MLLT	0.55(33)[2]	1.30(11)	2.61(1)	5.95(2)	0.82(11)	1.30(-1)	1.79(2)	3.44(3)

[1]Minimum relative reduction in WER required at a 95% confidence level

[2]Actual relative reduction in WER compared to environment compensated MFCC space

Table 4.7 Comparing insertion rates obtained from recognition using MATE in environment compensated MFCC space vs. environment compensated HDA/MLLT space

	Clean Training				Multi-condition Training			
	Clean	20dB	15dB	10dB	Clean	20dB	15dB	10dB
MFCC	0.02[121][1]	0.11[51]	0.17[41]	0.53[23]	0.08[60]	0.14[46]	0.20[38]	0.52[24]
HDA/MLLT	0.05(-133)[2]	0.18(-60)	0.40(-130)	1.16(-117)	0.08(-10)	0.24(-68)	0.34(-73)	0.67(-29)

[1]Minimum relative reduction in insertions required at a 95% confidence level

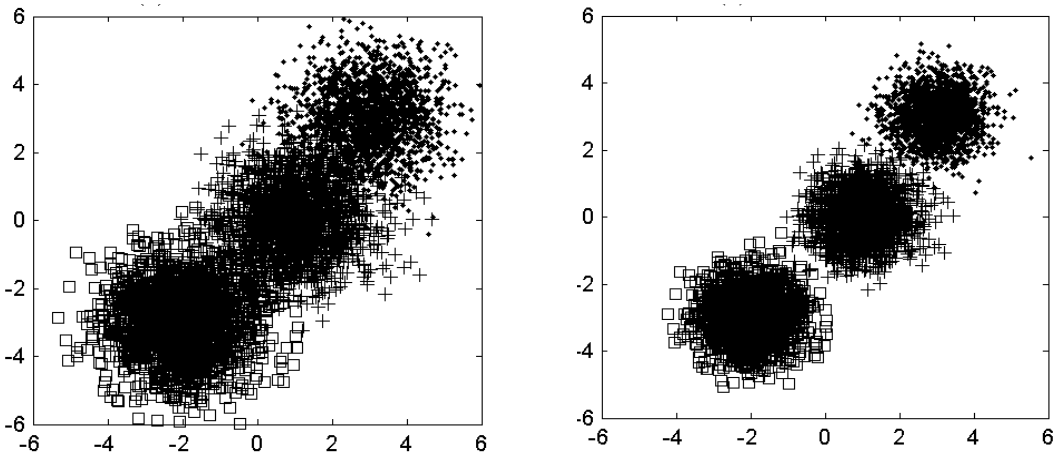
[2]Actual relative reduction in insertions compared to baseline

in the HDA/MLLT space, we are not able to verify the claim that increasing class discrimination through HDA improves the performance of speaker normalization using MATE. This is because, as a result of an increase in the resulting insertion rate, the overall performance of MATE does not seem to improve in the presence of noise in the HDA/MLLT space.

4.4 Estimating Discriminant Feature-Space Transformations from Normalized Data

Section 1.4.1 introduced the idea of normalizing speech data prior to performing discriminant feature-space transformation (DFT). Indeed, performing environment compensation and speaker normalization on speech data should reduce the variance in each class due to variabilities in environment, channel, and speaker characteristics. This should enable, heteroscedastic discriminant analysis (HDA) to achieve a higher degree of class discrimination as determined by the separability criterion given in Equation 2.21. Figure 4.5 illustrates visually how a reduction in class variance can lead to better class discrimination. The figure

shows a hypothetical situation where we have three classes of 2-dimensional Gaussian data.



(a) Gaussian data with unit variance

(b) Gaussian data with variance 0.7

Figure 4.5 A visual demonstration of the effects of a reduction in within-class variance on separability of data

Our analysis of the effects of normalization on the performance of DFT is divided into two parts, based on how the actual normalization is performed. In Section 4.4.1, we will look at environment compensation as the normalizing agent, while in Section 4.4.2, we will look at speaker normalization.

4.4.1 Environment Compensation

In order to investigate the effects of normalization using environment compensation on DFT, Table 4.8 compares the baseline ASR results for two different system configurations. To obtain the results in the first row, as described in Section 3.3.1, we estimated the HDA transform and the MLLT based on MFCC features. Then, in the resulting HDA/MLLT space, we trained models based on both the clean and the multi-condition training sets, and performed recognition using the baseline system. To obtain the results in the second row, the same procedure was used, with the exception that, in this case, *environment compensated* MFCC features were used to estimate the HDA matrix. In this manner, performing environment compensation on the MFCC features used to estimate the HDA matrix, constitutes normalization of the data prior to DFT computation.

Table 4.8 Baseline system in HDA/MLLT space: effects of estimating DFT parameters from data normalized using environment compensation (% WER)

	Clean Training				Multi-condition Training			
Normalization	Clean	20dB	15dB	10dB	Clean	20dB	15dB	10dB
Off	0.68[21]	3.11[10]	10.99[5]	30.44[3]	1.13[16]	1.95[12]	2.68[10]	4.67[8]
On	0.62(9)	1.55(50)	2.65(76)	5.90(81)	0.96(15)	1.25(36)	1.83(32)	3.48(26)

[1]Minimum relative reduction in WER required at a 95% confidence level

[2]Actual relative reduction in WER compared to unnormalized case

One interesting aspect of Table 4.8 is that, for the clean training case, we obtain larger WER reductions as the SNR decreases. This can be explained by noting the mismatch that arises when we transform the noisy test data using an HDA matrix which was estimated from the clean training data. Since normalizing the data using environment compensation reduces this mismatch by making noisy test data "resemble" clean data more closely, the lower the SNR the more mismatch is being removed, and therefore, compared to the unnormalized case, the more improvement we obtain. The same trend is not seen in the multi-condition case, since the HDA matrix is estimated from a number of different conditions, and therefore "normal" does not constitute clean.

The results in Table 4.8 lead us to conclude that normalizing the speech data using environment compensation prior to estimating the DFT parameters does in fact improve ASR performance. As stated in Section 4.4, we believe that these improvements are obtained because normalization causes an increase in class separability by reducing the variance of the data in each class. Having stated in Section 2.5.1 that the LDA measure of separability is proportional to the sum of the magnitude of the largest eigenvalues of the LDA matrix, we can examine this increase in class separability by considering these eigenvalue magnitudes.

To this end, according to the LDA measure of class separability given in Equation 2.18, we computed LDA matrices for the multi-condition training data, once with environment compensation and once without. In Figure 4.6 we have plotted the magnitude of the 30 largest eigenvalues for both cases. Comparing the two plots, it is evident that normalizing the data using environment compensation results in eigenvalues considerable larger in magnitude, and therefore higher class separation.

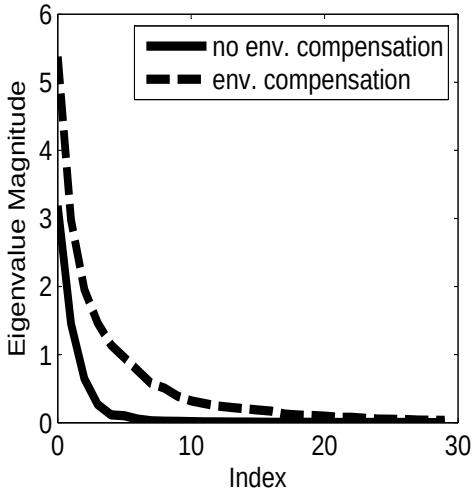


Figure 4.6 The effects of environment compensation on eigenvalues of the LDA matrix

4.4.2 Speaker Normalization

In Section 3.3.3 we presented the procedures used for estimating the HDA matrix from speaker normalized data, where speaker normalization was either performed using VTLN or MATE. Table 4.9 contains the corresponding recognition results obtained when speaker normalization is performed using VTLN. To obtain the results in the first row (which were also presented in Table 4.5), we estimated the HDA transform and the MLLT based on environment compensated MFCC features. Then, in the resulting environment compensated HDA/MLLT space, we trained models based on both the clean and the multi-condition training sets, and performed recognition using VTLN. To obtain the results in the second row, the same procedure was used, except that we normalized the environment compensated MFCC features which were used for estimating the HDA matrix. The normalization was performed using VTLN, according to the procedure given in Section 3.3.3. The results obtained when recognition and normalization are performed using MATE are presented in Table 4.10.

The results given Tables 4.9 and 4.10 indicate that we were not able to obtain significant improvements from removing speaker variabilities prior to estimating the HDA transform. As stated in Section 4.4, any improvements should be a result of an increase in class separability. Therefore, we can analyze our results further by obtaining a measure of the

Table 4.9 VTLN in HDA/MLLT space: effects of estimating DFT parameters from speaker normalized data (% WER)

	Clean Training				Multi-condition Training			
Normalization	Clean	20dB	15dB	10dB	Clean	20dB	15dB	10dB
Off	0.55[23][1]	1.33[15]	2.36[11]	5.66[7]	0.90[18]	1.14[16]	1.63[13]	3.25[9]
On	0.61(-10)[2]	1.39(-5)	2.58(-9)	5.66(0)	0.78(13)	1.19(-5)	1.73(-6)	3.07(6)

[1]Minimum relative reduction in WER required at a 95% confidence level

[2]Actual relative reduction in WER compared to unnormalized case

Table 4.10 MATE in HDA/MLLT space: effects of estimating DFT parameters from speaker normalized data (% WER)

	Clean Training				Multi-condition Training			
Normalization	Clean	20dB	15dB	10dB	Clean	20dB	15dB	10dB
Off	0.55[23][1]	1.30[15]	2.61[10]	5.95[7]	0.82[19]	1.30[15]	1.79[13]	3.44[9]
On	0.59(-7)[2]	1.28(2)	2.42(7)	5.65(5)	0.91(-11)	1.21(7)	1.90(-6)	3.53(-3)

[1]Minimum relative reduction in WER required at a 95% confidence level

[2]Actual relative reduction in WER compared to unnormalized case

change in class separability caused by removing speaker variabilities.

Having stated in Section 2.5.1 that the LDA measure of separability is proportional to the sum of the magnitude of the largest eigenvalues of the LDA matrix, we can determine a measure of class separability by examining these eigenvalues. To this end, according to the LDA measure of class separability given in Equation 2.18, we computed LDA matrices for the multi-condition training data under four different settings. In the first case, we computed the LDA matrix from unnormalized MFCC features, while in the second case we performed environment compensation on the data before estimating the LDA matrix. In the third and fourth case, in addition to performing environment compensation prior to estimating the LDA matrix, we removed speaker variability through VTLN and MATE respectively. Table 4.11 lists the mean of the magnitude of the 30 largest eigenvalues of the LDA matrix determined in each of these cases.

Table 4.11 shows that, when environment compensation is performed, the mean of the magnitude of the eigenvalues has a threefold increase from 0.199 to 0.622. This translates to a considerable increase in class separability, and hence, as shown in Section 4.4.1, a significant improvement in ASR performance. However, when speaker normalization is

Table 4.11 Comparing the mean of the magnitude of the largest 30 LDA eigenvalues for different normalization modes

Normalization Mode	Eigenvalue Magnitude Mean
–	0.199
Environment compensation	0.622
Environment compensation and VTLN	0.633
Environment compensation and MATE	0.649

performed in addition to environment compensation, the additional increase in the mean of the magnitude of the eigenvalues, as shown in Table 4.11, is only about 2% in the case of VTLN, and 4% in the case of MATE. Therefore, the additional separability gained through performing speaker normalization using VTLN and MATE is relatively quite small. Hence, correspondingly, we can only expect relatively small improvements in ASR performance. Therefore, we believe that, the size of the test set used for generating the results in Table 4.9 and 4.10 was not large enough to resolve such small improvements in ASR performance, and experiments with larger test sets are required.

4.5 Utilizing Gender-Specific Warping Factor Priors in MATE

As stated at the beginning of this chapter, our experiments with the MATE speaker normalization technique revealed that this technique does not perform well in noisy conditions. For instance, when considering the recognition results obtained in the environment compensated MFCC space in Table 4.3, we can see that, as the testing SNR decreases, MATE yields smaller improvements over the baseline system. Given our analysis of the effects of noise on the distribution of warping factors selected by VTLN in Section 4.2.1, we can expect noisy environments to have similar effects on the warping factor selection process of MATE, leading to the limitations on its performance. Therefore, in order to increase its robustness against such effects, we devised a simple modification to the original MATE design.

In Section 2.4.2, we described how an additional iteration of training is performed using utterances warped using the MATE decoder. Essentially, this is done in an effort to utilize, during recognition, knowledge gained through the use of frequency warping on the training data. While HMM retraining attempts to capture this knowledge by modifying all of the

parameters of the acoustic model, obtaining prior information about the distribution of state-dependent warping factors selected by MATE is another means of capturing this knowledge. This approach was taken to incorporate gender-specific prior knowledge into the MATE decoder.

By observing the warping factors α_n that are decoded for each state q_j on the training utterances, it is possible to obtain an empirical estimate of the conditional probability $p(\alpha_n|q_j)$ for each of S HMM states. Given this conditional probability, it is possible to replace the observation probabilities $b_j(\vec{x}_t) = p(\vec{x}_t|q_j)$ in Equation 2.13, which represents the original MATE decoding algorithm, with the joint probability

$$p(\vec{x}_t, \alpha_n|q_j) = p(\vec{x}|q_j)p(\alpha_n|q_j), \tag{4.2}$$

under the assumption that $\vec{x}_t$ and α_n are independent given q_j. The new decoding algorithm is given by

$$\phi_{j,n}(t) = \max_{i,m} \left\{ \phi_{j,m}(t-1) a_{i,j}^{m,n} \right\} b_j(x_t^{\alpha_n}) g_j(\alpha_n), \tag{4.3}$$

where we modelled the probabilities $g_j(\alpha_n) = p(\alpha_n|q_j)$ as simple univariate Gaussians, with the means μ_j and variances σ_j^2 estimated from the training data. In other words, in selecting a given warping factor α_n for state q_j during recognition, we used prior knowledge about the distribution of warping factors selected for state q_j during training.

Given that female speakers usually have a shorter vocal tract length and therefore, their utterances contain higher formant frequencies, frequency warping tends to compress the frequency axis for these speakers [27]. On the other hand, frequency expansion is usually occurs for male speakers. As a result, in our current implementation of this approach we trained separate warping factor distributions for male and female speakers, and assumed knowledge of the gender of speakers during recognition. In a more realistic implementation, where the gender information is not available during recognition, decoding can be performed once for each gender assumption, and the transcription corresponding to the gender-specific warping factor distribution with higher likelihood would be selected.

Another aspect of this algorithm is related to the fact that MATE does not perform warping of the frames corresponding to silence states. Hence, as it is not possible to obtain estimates of the $g_i()$ for silence states, an empirically derived threshold was used in its place. In this manner, silence states are penalized to compensate for the absence of the

additional warping factor probability term in Equation 4.3.

4.5.1 Evaluation

Table 4.12 compares the results obtained by our proposed scheme against the results obtained by the original MATE. The models used in these experiments were trained using the environment compensated MFCC features. The first row of the table shows recognition result obtained using the original MATE algorithm, while the second row of the table shows the results obtained using the modified MATE.

Note that, the results reported here for the original MATE are slightly different from the ones reported previously in Table 4.3. As stated in Section 3.2, in our experiments, MATE was configured with an ensemble of 5 warping factors covering a range from 12% compression to 12% expansion of the frequency axis. However, in the modified MATE, in order to be able to have a higher resolution in estimating the distribution of warping factors, we configured the modified MATE with an ensemble of 12 warping factors covering the same range. This difference in corresponding configurations is the cause for the differences in WER results reported in Table 4.3 and Table 4.12.

Table 4.12 Comparing the recognition performance of modified MATE with the original MATE in environment compensated MFCC space (% WER)

	Clean Training				Multi-condition Training			
	Clean	20dB	15dB	10dB	Clean	20dB	15dB	10dB
MATE	0.78[19][1]	1.50[14]	2.56[11]	6.15[7]	0.97[17]	1.17[16]	1.79[13]	3.40[9]
Modified MATE	0.67(14)[2]	1.31(13)	2.33(9)	5.49(11)	0.97(1)	1.20(-3)	1.72(4)	3.31(3)

[1]Minimum relative reduction in WER required at a 95% confidence level

[2]Actual relative reduction in WER compared to MATE

The comparison in Table 4.12 shows that, for the clean training case, consistent improvements are obtained for all test cases. However, no significant improvements are obtained for the multi-condition training case. One possible explanation for this can be stated by noting that, as stated in Section 3.1, the multi-condition training set consists of utterances with a range of SNR conditions. Therefore, when using this training set, as with the acoustic model, the warping factor distributions are trained from utterances with a variety of different SNR conditions. Hence, it is conceivable that each one of these noise conditions results

in a different warping factor distribution, and therefore, using one Gaussian to model all of them would be inadequate.

4.6 Summary

In this chapter, we presented a number of experiments aimed at investigating the claims of this thesis as stated in Section 1.4.1. In Section 4.2, we showed that the use of environment compensation improved the performance of the VTLN and MATE speaker normalization techniques. As part of our analysis, we also examined the effects of noise on the warping factor estimation process and the first recognition pass of VTLN, subsequently showing how these effects were ameliorated through the use of environment compensation.

In Section 4.3, we considered the performance of speaker normalization techniques in a discriminant feature-space. We concluded from our results that the use of VTLN in a discriminant feature-space improves its performance, while the same cannot be said about MATE. Our analysis showed that the shortcomings of MATE were due to an increase in the insertion rate obtained by this technique during recognition.

Section 4.4 was concerned with the effects of estimating the discriminant feature-space transform (DFT) parameters from data which had been normalized using environment compensation and speaker normalization. In our analysis, using the magnitude of the eigenvalues of the LDA matrix as a measure of class discrimination, we determined that normalizing the data through environment compensation resulted in a much larger increase in class discrimination compared to when normalizing the data through speaker normalization. Correspondingly, quite significant WER reductions were obtained when normalizing with environment compensation, while a larger test set was required to resolve any improvements gained through the use of speaker normalization.

Finally, in Section 4.5 we presented a simple modification to the original MATE aimed at increasing its noise robustness. In the modified MATE, we employed, during recognition, gender-specific knowledge of the distribution of the warping factors selected during training. Evaluating this technique, when training from clean data, we obtained significant improvements compared to the original MATE, while no significant improvement were obtained for the case of multi-condition training.

Chapter 5

Summary and Conclusions

This thesis was concerned with the notion of robustness against sources of variability in automatic speech recognition (ASR) systems. To this end, we performed an experimental study concerning the interrelationship between three separate techniques: environment compensation techniques, which remove speech variabilities due to environment and channel characteristics, speaker normalization techniques, which remove variabilities due to speaker characteristics, and discriminant feature space-transformation techniques (DFT), which are aimed at increasing the class discrimination of the speech data.

This chapter provides a brief summary of our work, and is divided into three sections. First, we will provide a brief description of the speech corpus and the baseline ASR system, as well as the specific techniques employed in our experiments. Second, we will restate the claims that were investigated in this thesis, and will summarize the results and conclusions corresponding to each claim. Finally, we will explore some of the potential avenues for further research related to our work.

5.1 Experimental Context

Our experiments were based on a continuous density hidden Markov model (CDHMM)-based ASR system. In our system, Mel-frequency cepstral coefficients (MFCC) were extracted in the feature analysis stage, and environment compensation was performed according to the European telecommunications standard institute advanced front-end (ETSI-AFE).

The two speaker normalization algorithms used in our thesis were vocal tract length

normalization (VTLN) and the augmented state-space acoustic decoder (MATE). These techniques were based on normalizing the effective vocal tract length of different speakers by applying a linear warping to the frequency axis of speech utterances. More specifically, VTLN used a maximum likelihood framework to choose a single warping factor for each utterance, while MATE used a special maximum likelihood decoding algorithm to choose different warping factors for individual speech frames.

In order to increase class discrimination, the extracted features for a given frame were concatenated with features from a number of surrounding frames, and heteroscedastic discriminant analysis (HDA) was performed to reduce the dimensionality of the resulting vector while maximizing class discrimination. This was followed by a maximum likelihood linear transformation (MLLT) to diagonalize the covariance of the resulting feature-space.

Our experiments were performed using ETSI Aurora 2 speech corpus, which consists of connected digit utterances under a variety of noise types and noise levels. This standardized corpus provided us with a clean training set, a multi-condition training set and a number of test sets corresponding to a range of signal-to-noise ratio (SNR) assumptions.

5.2 Claims

Our first claim was concerned with improving the performance of speaker normalization techniques through the use of environment compensation. In Section 4.2, evaluating the performance of VTLN and MATE under a range of SNR assumptions, we showed that the performance of these techniques is degraded under noisy conditions. Subsequently, incorporating environment compensation in our system, we reevaluated VTLN and MATE, obtaining significant improvements in their performance. For example, under 10dB SNR testing and clean training conditions, without environment compensation, VTLN obtained a 2% reduction in WER compared to the baseline system, whereas, when environment compensation was used, this WER reduction grew to more than 8%. Furthermore, as part of our analysis, we examined the adverse effects of noise on the distribution of warping factors selected by VTLN, as well as the performance of the first recognition pass of this technique. Subsequently, we demonstrated how these effects were ameliorated through the use of environment compensation.

Our second claim stated that the performance of speaker normalization techniques should improve when applied in discriminant feature-space transformation (DFT). To in-

vestigate this, we performed HDA on environment compensated MFCC features to increase the class discrimination, where classes corresponded to HMM states, followed by MLLT as stated above. We then performed ASR training and recognition using VTLN and MATE in the resulting environment compensated HDA/MLLT space. These results indicated that VTLN performs significantly better in this space compared to the environment compensated MFCC space. For example, under 10dB SNR testing and clean training conditions, the WER obtained by VTLN was 13% lower in the environment compensated HDA/MLLT space. In the case of MATE, due to an increase in the insertion rate, no significant WER reductions were obtained.

Our final claim suggested that using environment compensation and speaker normalization to normalize the data prior to estimating the DFT parameters should increase the final class discrimination achieved by DFT, reducing the overall WER. Using the magnitude of the eigenvalues of the LDA matrix as a measure of class separability, we showed that normalizing the data through speaker normalization increased the class discrimination significantly. Correspondingly, our baseline recognition results showed significant WER reduction when the HDA matrix was estimated from environment compensated features. For example, under 10dB SNR testing and clean training, an 81% WER reduction was obtained. In the case of speaker normalization, considering the LDA matrix eigenvalues showed a very relatively small increase in class discrimination. Likewise, the baseline recognition results obtained when the HDA matrix was estimated from environment compensated and speaker normalized features did not show significant improvements. Considering the small increase in class discrimination, we believe that a larger test set is required ro resolve the corresponding small improvements in WER.

In addition, having realized during our experiments that MATE did not perform well in noisy conditions, we presented a modified algorithm to increase its robustness. The modification was based on using gender-specific distribution of the warping factors selected by MATE during training as prior knowledge when selecting warping factors during recognition. In order to evaluate this technique, we performed testing and training in the environment compensated MFCC space. The corresponding results indicated significant WER reductions when using the clean training data, while no significant improvements were obtained for the case of multi-condition training. For example, under 10db SNR testing and clean training, the WER obtained by the modified MATE was 11% lower than that of the original MATE.

5.3 Future Work

As stated in Section 3.1, the fact that the Aurora 2 corpus which was used in our experiments contained simulated noise posed some limitations on the reliability of our obtained results. Therefore, one obvious avenue for future work is repeating our experiments on a speech corpus which has actually been recorded in a noisy environment. Furthermore, we obtained very small improvements in class discrimination by estimating the DFT from speaker normalized data, and therefore, realized that a bigger test set was required to actually show the corresponding small reductions in WER. Hence, this issue can also be investigated by using a different speech corpus. In addition, we mentioned in Section 3.3.1 that the target dimension of the discriminant feature-space transformations (DFT) we estimated for our experiments was selected based on convenience, whereas the best dimension is usually selected empirically based on the rank of the feature-space. Therefore, in future experimenters the target dimension of the DFT can be selected based on this empirically determined feature-space rank.

Another major possibility for future work is improving the performance of the MATE speaker normalization algorithm. As stated in Section 4.5, our modified version of MATE did not produce significant improvements under multi-condition training. Subsequently, we explained this by noting the fact that different distributions of warping factors are expected from each of the different SNR conditions comprising the multi-condition training set. Therefore, using a simple Gaussian distribution did not seem to be adequate for representing the combined distribution of warping factors for each noise condition. Hence, the use of other more complex distributions, such as a mixture of Gaussians, can be investigated.

Furthermore, we indicated in Section 4.5, that the current implementation of our modification to MATE assumed knowledge of the speaker genders during recognition. One possible area for future investigation is to remove this assumption by performing recognition for both gender assumptions and selecting the transcription corresponding to the one yielding the higher likelihood.

Bibliography

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

[2] S. Das, R. Bakis, A. Nadas, D. Nahamoo, and M. Picheny, "Influence of background noise and microphone on the performance of the IBM TANAGORA speech recognition system," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process*, vol. 2, pp. 71 – 74, 1993.

[3] J. Xu and G. Wei, "Noise-robust speech recognition based on difference of power spectrum," *IEEE Electronic Letters*, vol. 36, pp. 1247 – 1248, July 2000.

[4] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 55 – 69, January 1999.

[5] T. Arakawa, M. Tsujikawa, and R. Isotani, "Model-based Wiener filter for noise robust speech recognition," vol. 1, pp. 537 – 540, May 2006.

[6] ETSI Technical Committee Speech processing, Transmission and Quality aspects (STQ), "ETSI ES 202 050 v0.1.0 distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm," tech. rep., European Telecommunications Standards Institute, 2002.

[7] D. Macho and Y. M. Cheng, "SNR-dependent waveform processing for improving the roboustness of ASR front-end," in *ICASSP*, vol. 1, pp. 305 – 308, 2001.

[8] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 352 – 359, September 1996.

[9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257 – 286, 1989.

[10] B. Milner and A. James, "Robust speech recognition over mobile and IP networks in burst-like packet loss," *IEEE Transactions on Speech and Audio Processing*, vol. 14, pp. 223 – 231, January 2006.

[11] X. D. Huang and K. F. Lee, "On speaker-independent, speaker-dependent, and speaker adaptive speech recognition," in *ICASSP*, vol. 2, pp. 877 – 880, 1991.

[12] J. Gauvain and C.-H. 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, January 1994.

[13] C. Leggetter and C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," *Eurospeech*, pp. 1155 – 1158, January 1995.

[14] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on speech and audio processing*, vol. 6, p. 49, 1998.

[15] A. Miguel, E. Lleida, R. Rose, L. Buera, and A. Ortega, "Augmented state space acoustic decoding for modeling local variability in speech," in *ICSLP*, vol. 1, pp. 3009 – 3012, 2005.

[16] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Trans. ASSP*, vol. 28, pp. 357 – 366, 1980.

[17] J. Hernando, "Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition," pp. 1267 – 1270, April 1997.

[18] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 52 – 59, February 1986.

[19] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," tech. rep., IBM T. J. Watson Research Center, 2000.

[20] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley Interscience, second ed., 2000.

[21] E. G. Schukat-Talamazzini, J. Hornegger, and H. Niemann, "Optimal linear feature transformations for semi-continuous hidden Markov models," in *ICASSP*, vol. 1, pp. 369 – 372, 1995.

[22] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *ICASSP*, vol. 1, pp. 13 – 16, 1992.

[23] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 25, pp. 283 – 297, 1998.

[24] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR, 2001.

[25] B. Juang and L. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 1639 – 1641, Sept 1990.

[26] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on Aurora databases," in *ICSLP*, vol. 1, pp. 17 – 20, 2002.

[27] L. Lee, "A frequency warping approach to speaker normalization," M. Eng. thesis, Massachusetts Institute of Technology, 1996.

[28] R. Rose, A. Keyvani, and A. Miguel, "On the interaction between speaker normalization, environment compensation, and discriminant feature space transformations," in *ICASSP*, vol. 1, pp. 985 – 988, 2006.

[29] P. F. Brown, *The acoustic modelling problem in automatic speech recognition*. PhD thesis, Carnegie Mellon University, 1987.

[30] R. A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *ICASSP*, vol. 2, pp. 661 – 664, 1998.

[31] D. Pearce and H.-G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ICSLP*, p. 18 20, 2000.

[32] R. S. K. Y. B. Tian, M. Sun, "A unified compensation approach for speech recognition in severely adverse environment," pp. 256 – 261, 2003.

[33] C. Chen, J. Bilmes, and K. Kirchhoff, "Low-resource noise robust feature post-processing on Aurora 2.0," in *ICSLP*, vol. 4, pp. 2445 – 2448, 2002.