Understanding and adapting to speaker variability using correlation-based principal component analysis

Zhihong Hu
B.E., Tsinghua University, 1990

A dissertation submitted to the faculty of the
Oregon Graduate Institute of Science and Technology
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Computer Science and Engineering

October 1999

The dissertation "Understanding and adapting to speaker variability using correlation-based principal component analysis" by Zhihong Hu has been examined and approved by the following Examination Committee:

Etienne Barnard
Associate Professor
Thesis Research Adviser

Ronald A. Cole
Professor and director
Center for Spoken Language Understanding

Pieter J. Vermeulen
Associate Professor

Richard Stern
Professor
Dept. of Electrical and Computer Engineering
Carnegie Mellon University

Peter A. Heeman
Assistant Professor

Dedication

To my dear husband Johan and my family.

Acknowledgments

This thesis is done with the help with many people. I need to express my thanks to all of them. First, I thank my dear husband Johan for all the love he put into the birth of the thesis, for always being there and all the encouragement, for perpetual support and many laughs over the years. I thank my parents for a lot of encouragement and support for me to study science and engineering and lots of loving support from my mom and dad from South Africa. I also want to thank my friend Kay for all these years of friendship and encouragement.

I thank Etienne for all these years of patient advise and Ron for giving me the opportunity to do research in speech and continuing support. I thank Pieter for always being there to give me often needed help and advise. I appreciate Rich for being part of my committee and many thoughtful suggestions and advise. I appreciate Peter for giving me many helpful suggestions for the thesis. Finally, I thank Charlene for her great effort of helping me arrange the final stage of putting the thesis together.

Contents

List of Tables

List of Figures

Abstract

Understanding and adapting to speaker variability using correlation-based principal component analysis

Zhihong Hu

Supervising Professor: Etienne Barnard

In this research, we study the relationship amongst different sounds in speech of a speaker and between speakers. We propose a new speaker normalization/adaptation model that incorporates correlations amongst phoneme classes, and explore the applications of the model. Using principal component analysis we construct a speaker space based on a speaker covariance matrix obtained from the training data. The speaker covariance matrix is constructed in such a manner as to explicitly describe the correlations between classes.

The hypothesis of this thesis is that by explicitly modeling these correlations it is possible to adapt the model or normalize the speaker's features with limited adaptation data. This proposed method also allows researchers to understand some of the properties of speaker variability and gives insight into some physical aspects of speaker differences.

This hypothesis is tested in segment classification tasks where other variant conditions (such as contextual variation) are minimized. Various possible applications of this method are also tested on an HMM speech recognizer.

Chapter 1

Introduction

Automatic speech recognition (ASR) is rapidly advancing to become part of our everyday lives. As a natural human-computer communication mode, speech recognition technology will greatly enhance information accessibility.

Although current state-of-the-art systems perform well under controlled conditions, such as the Wall Street Journal read speech corpus, spontaneous speech remains an important indicator that computer speech recognition is not yet capable of matching human speech perception. Research on the Switchboard corpus (spontaneous speech over the telephone) is an example. For this problem, only 40% word accuracy is achieved.

Why is it, then, that humans do so well? One important reason for this is that human perception is extremely good at adapting to variations. These might be variations in acoustic environments, speaker, accent, and domain (topic). For example, a person can not physically produce the same utterance in exactly the same acoustic form twice. The speed, tone and loudness may vary. People from different regions have different accents and even grammars. Additionally, the background noise, or the communication channel (microphone/telephone) could be different. These factors contribute to the variability in the speech signal.

It is difficult to design a speech-recognition system that can adjust to all kinds of variability in the signal. Having the ability to adapt to these variations - thus eliminating them from consideration - should greatly improve computer speech recognition. For example, various channel normalization techniques [17] improve channel robustness significantly; speaker adaptation techniques such as MLLR[33] and MAP[43] help to improve the speaker-independent system's performance.

The purpose of this thesis is to investigate methods to study speaker variances and compensate for them. Specifically, this thesis addresses methods that could be used to shed light on and understand speaker variabilities and furthermore use this knowledge to efficiently adapt a speaker-independent recognition system to approach the performance of a speaker-specific system. Hence this will improve the overall recognition performance.

1.1 Variabilities in speech signals

Variabilities in speech can be grouped into the following categories: *domain variabilities*, *speaker variabilities*, and *environmental variabilities.*

- **Domain variabilities** are the variabilities introduced by different topics of speech. Different domains of speech contain different vocabularies. These variabilities are modeled in language models in speech recognition systems.

- **Speaker variabilities** are the variabilities caused by the differences among speakers. These include the following:

  - **Physiological differences:** Different people have different physiological characteristics. For example, male speakers generally have a longer vocal tract compared to female speakers. A longer vocal tract in turn relates to lower resonant frequencies. Most of the speaker adaptation techniques currently used attempt to compensate for speaker differences and the effect they have on modeling the speech data.

  - **Linguistic differences:** People talk differently. The differences can arise from tone, accent, or talking mode. The realizations of the same word can be different among speakers. The realizations of the same word can even be different for the same speaker under different circumstances. Factors such as stress, duration, vowel coloration, and intonation may all be different.

- **Environmental variabilities** are introduced into the speech signal from the environment. These include the following:

- **Background noise** that gets mixed into the speech signal such as door bangs, telephone rings, music, etc. Generally, these kinds of noises are additive.
- **Channel variation** is caused by different channels, such as the telephone hand sets, microphones, transmission lines, etc. These effects can be described as convolutional noise.

Designing a speech recognition system to cope with all these variations is a difficult problem. Nearey's [35] experiments suggest that the speaker-dependent variation is the second biggest source of variation in vowel quality (the largest one is the vowel identity itself), and thus a worthy candidate for elimination. Other researches have concentrated on environmental variances such as RASTA, CMS, etc. In this research, we concentrate on compensating for variabilities, which are typically caused by physiological and habitual differences between speakers.

1.2 Speaker dependent and speaker independent acoustic modeling

One crucial area of any speech-recognition system is the ability to accurately model the set of acoustic classes chosen as the basic units of recognition. If the acoustic modeling is poor, the effects of additional knowledge gained from language modeling or lexicons are limited. Recognition is performed by matching the observed unknown speech to the model which is most likely to have generated that speech. Generally, a large amount of training data is required for a sufficient degree of modeling accuracy.

Speaker independent systems (SI) are designed to be used by all users. They are trained with data from many speakers. As a result the models describe the distribution of features for all speakers, rather than a specific speaker. These systems have to model both intra-speaker and inter-speaker variations, and typically contain on the order of millions of parameters to do so. To estimate these parameters accurately requires a great amount of training data. Given these complex systems, there is, however, no guarantee that the system will perform equally well for each possible test speaker.

On the other hand, a speaker-dependent (SD) system is designed to be used by a single speaker. Since the system is trained on only one speaker's data it does not have to model the inter-speaker variation. The resulting recognizer can be very precise for this particular speaker but tends to perform considerably worse when tested on other speakers. Assuming that sufficient data is available to train both systems, speaker-dependent systems typically perform better than equivalent speaker independent systems, as mentioned in the results on Resource Management by Huang [20].

1.3 Speaker adaptation and speaker normalization

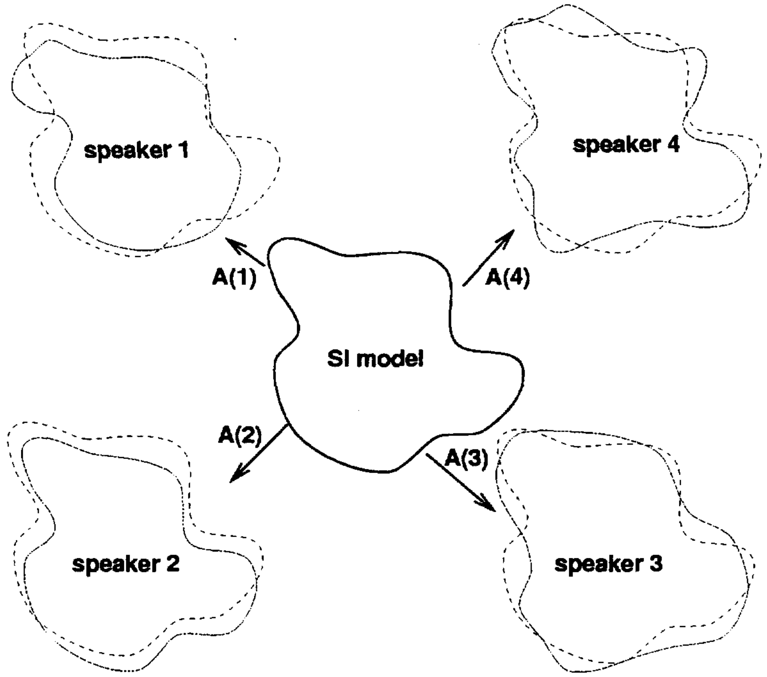


Figure 1.1: Speaker adaptation methods derive a mapping between the SI reference space and the speaker specific space (depicted by the solid-line images). The speaker adapted models (depicted in the dashed lines), generated by applying the speaker-specific mappings ($A(1)$, $A(2)$, etc.) to the SI model, match more accurately for the new speakers than the SI model.

In many applications, it is impossible for each user to provide a sufficient amount of training data to train a speaker-specific recognizer. Speaker-dependent performance is still, however, desired. To address this problem, and as an improvement of the speaker independent system, speaker adaptation technology has been developed. Speaker adaptation generally uses the information (adaptation data) available from the new speaker to adjust the speaker independent recognizer to be specific to the new speaker. This improves the recognition performance for this new speaker and hence improves the system's overall performance.

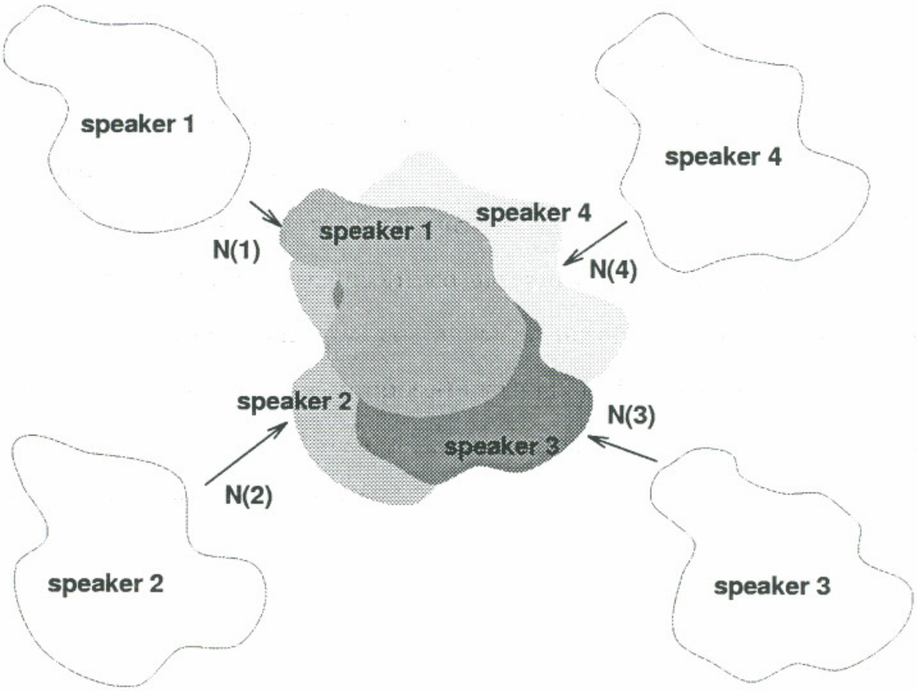


Figure 1.2: Speaker normalization processes construct a "neutral" speaker space that any speaker can be projected into and the inter-speaker differences are minimized.

Figure 1.1 illustrates the general principles of speaker adaptation in a two dimensional space. Speaker normalization attacks the problem of speaker variances from the other end of the process. Speaker normalization algorithms attempt to convert the characteristics of speech from each speaker to those of a reference or "neutral" speaker. The normalization process is applied to both the training and test speakers. As depicted in Figure 1.2, the objective of speaker normalization is to construct a "neutral" speaker space into which speech from any speaker can be projected (via $N(i)$) so that inter-speaker variations will be minimized. This results in less variance in the feature space and thus more precise modeling.

1.4 Adaptation modes

The process of adapting a speech recognition system can be performed in several different modes. These modes differ in (1) how and when the adaptation takes place, and (2) how the adaptation data are used.

The amount of available adaptation data is generally the most important factor in choosing between these modes.

- The adaptation is called *supervised adaptation* when the content of the adaptation data (typically, the orthographic transcription) is presented to the adaptation process.
- Otherwise, the adaptation process is called *unsupervised* adaptation.

The adaptation data can be used in different ways :

- When all the adaptation data is presented to the adaptation process before the final adapted system is produced, it is called *static adaptation.*
- When the adapted system is produced after only part of the adaptation data is presented, and is refined while more adaptation data is observed, it is called *dynamic adaptation.*

Different adaptation modes have different effects on accuracy, efficiency, and usability on different tasks. It is obvious that when a large amount of adaptation data is available and correctly labeled, the appropriate adaptation mapping can be determined accurately, while limited adaptation data would hurt the estimation accuracy and unsupervised adaptation data may lead to errors in the estimation. In some tasks, a quick speaker enrollment is required, hence it may be desirable to initially start with a speaker independent system and use dynamic unsupervised adaptation. In tasks where high recognition accuracy is desired, more time may be available to collect adaptation data, and a static supervised mode can be used.

1.5 Organization of the thesis

Our goal in this research is to develop speaker adaptation and normalization techniques for efficient and fast adaptation of general-purpose speaker-independent recognizers and study the nature of the variation amongst speakers. The focus of this work is to use principal component analysis to determine the linear components which are the most important contributors to speaker differences. This work incorporates the correlations among the different classes in the adaptation process to enable adaptation with limited data. Different feature representations are studied. Various applications of this method are also explored and discussed.

Chapter 2 describes previous work on various speaker normalization and adaptation techniques. Chapter 3 describes the proposed method in mathematical detail, and Chapter 4 studies various mechanisms related to the proposed method. Chapter 5 provides experimental results to support the proposed method. In Chapter 6, various applications of this method are described and results are presented. A Summary and conclusions of this research are presented in Chapter 7.

Chapter 2

Background

In speech recognition, there usually is a mismatch between the data observed during training and the data observed during testing. This mismatch will result in a performance degradation compared to the matched condition (training and testing on a known set of speakers in a known environment). The task of speaker adaptation is to adapt a speaker independent model to the data of an unseen test speaker.

The mismatch between training and testing conditions can be represented in signal space, feature space, or model space as shown in Figure 2.1. In this figure, R represents the reference speech signal and T represents the target speaker's signal. **Fr** and **Ft** represent the feature transformations of the input signal for the reference and target speakers respectively. Similarly Mr and Mt represent the models for the reference and target speaker.

Various speaker adaptation methods that attempt to solve the problem of mismatch due to speaker variation have been developed. These methods operate either on the feature space (transformation **Tf**) or the model space (transformation **Tm**). Since the feature-extraction process can reduce the dimensionality of the speech signal and has proved to be efficient in representing the information of the original speech signal in speech recognition process, speaker adaptation has rarely been applied directly to the speech signal.

In the following sections, various adaptation techniques are presented. Section 2.1 presents techniques of compensating for the speaker variances in feature space. Section 2.2 describes model-based approaches, which are most related to the work in this thesis. Section 2.3, summarizes the chapter and briefly presents the context within which this research fits.

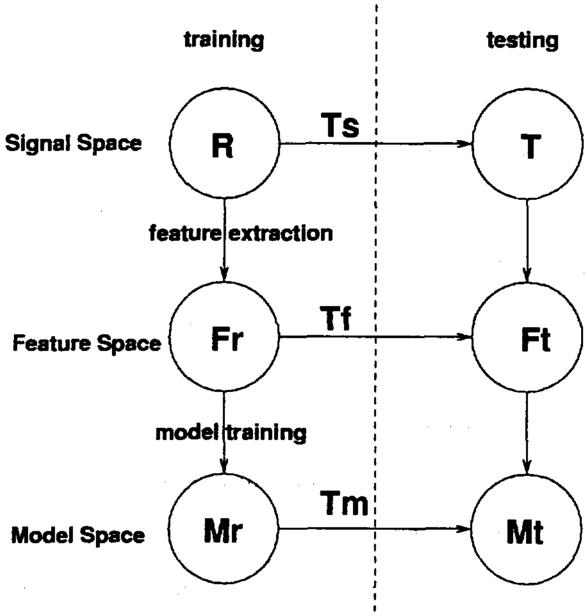


Figure 2.1: Different representations of mismatch conditions and their corresponding transformations.

2.1 Feature transformation techniques

The most prominent feature transformation approaches are vocal tract normalization (VTN) and spectral transformations. Vocal tract normalization normalizes the speaker's features during the feature extraction process. In contrast spectral transformations attempt to learn a mapping based on the features extracted.

2.1.1 Vocal tract normalization (VTN)

Vocal tract normalization (VTN) is an approach that normalizes the speaker variability caused by the physical differences in vocal tract length. The idea of vocal tract normalization emerged as early as 1977, when Wakita [41] presented a method of vowel formant frequency normalization.

Modeling the vocal tract directly from the speech signal has proven to be difficult. The features used in most speech recognition systems are therefore based on the perceptual model of the ear rather than an explicit vocal tract model. These features are typically cepstral transformations of the warped frequency spectrum, for example MFCC

(Mel scale) or PLP (Bark scale). To perform VTN, an optimal frequency warp scale is determined for each training speaker. Normally this requires the estimation of a single scalar normalization parameter.

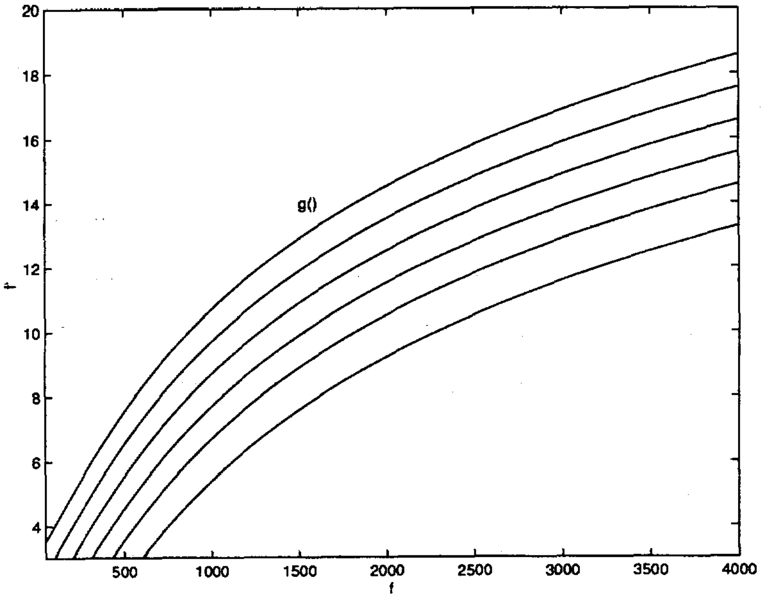


Figure 2.2: Representation of a typical frequency scale warping function.

In frequency scale warping, a parametric warping of the frequency axis f can be generally written as:

$$f' = g(k_s, f) \tag{2.1}$$

where k_s is a scalar which compensates for the vocal tract length of speaker s. Figure 2.2 depicts a typical warping function g. g can be linear or non-linear. The effect of the parameter k_s is to shift the warping function, either to the left or right. Shifting of the warped frequency scale in effect moves the formant values of the speaker towards a "neutral" position.

The formant values in speech correspond to resonance frequencies of the vocal tract. Eide and Gish [10] use a parametric normalization method to compute k_s. This value k_s is estimated with formant values for a test speaker as the median of the speaker's third formant over a subset of frames satisfying certain criteria. This method, however, involves a large amount of computation in estimating the appropriate warp parameters.

In similar work done by others [5, 23, 31, 32], the vocal tract length variation is crudely compensated for through a step-wise search of the parameter k_s. The parameter k_s applied to each speaker's speech is estimated using the speaker's utterance(s) within a model-based framework. In this approach the effect of the scalar (k_s) is evaluated either within a maximum likelihood (ML) framework or using a Viterbi approximation.

The advantage of VTN is that it only needs to estimate a single adaptation parameter per speaker and therefore requires a small amount of adaptation data. However, estimating this parameter normally involves a large amount of computation or difficult estimation of accurate formant values. Since the estimation of the adaptation parameter relies on the accuracy of estimation of formant values, this is also not desirable in real time implementations. Furthermore, since vocal tract normalization only models the variance caused by the differences in vocal-tract length, it does not have the ability to model other types of variances between speakers.

2.1.2 Spectral transformation approaches

The spectral mapping approach to speaker adaptation has been widely investigated in recent years. This approach applies a linear transformation to either the input feature vector, or to the whole model set. The main difference in the two implementations are in the manner in which the transformations are estimated.

Spectral transformations were first used with spectral-template based recognizers [14, 15, 16]. These approaches considered the reference templates as representing speech from a reference speaker and computed spectral transformations to minimize the distance between the new speaker and the reference speaker. The disadvantage of these techniques is that they are text dependent.

More success has been achieved by utilizing phone-specific transformations [21, 44], but these methods can be viewed as model transformations instead (see next section). A more detailed review of this topic can be found in [2, 33].

2.2 Model transformation techniques

The aim of the model transformation techniques is to derive a transformation that adjusts the speaker-independent model parameters in order to represent the new speaker more accurately. The transformation is chosen from a set with specified parametric form.

A generic linear model-based speaker adaptation transformation can be represented as follows:

$$u' = Au + b \tag{2.2}$$

where u' is the adapted model mean, u is the speaker independent or "prototype" model mean, A is the adaptation transformation matrix and b is the offset. Figure 2.3 depicts this process in a two dimensional feature space.

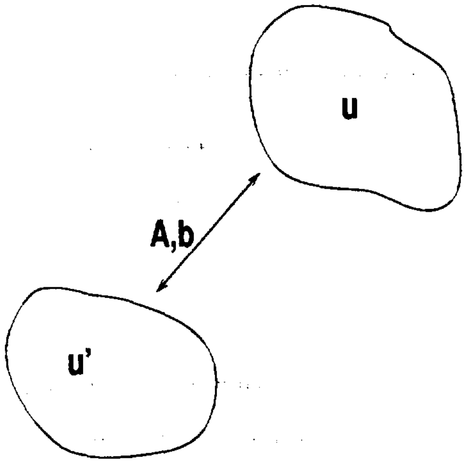


Figure 2.3: Linear transformation (A, b) adapts the "neutral" mean (u) to the speaker specific mean (u').

The transformation matrix A and offset b are used to adapt the speaker-independent model towards the new speaker. The speaker- independent model mean in effect represents the "neutral" speaker's model. Using the inverse transform we can adjust the speaker towards this neutral speaker.

The adaptation parameters represented by the transformation (A,b) may be estimated using different methods. Most modern adaptation can be categorized by the estimation used, which is either maximum likelihood (ML), or maximum *a-posteriori* (MAP) training (Bayesian approach).

2.2.1 Maximum likelihood linear regression (MLLR)

The aim of MLLR is to estimate an appropriate transformation for the mean vectors of each mixture component so that the original system is adapted to the new speaker. The adaptation parameters (A, b) are estimated to maximize the likelihood of the model on the adaptation data.

In the original formulation of MLLR [33, 39], linear transformations are associated with each component distribution within the Hidden Markov Model (HMM) framework and estimated using a maximum likelihood approach similar to the standard HMM parameter estimation. The transformations capture the general characteristics between the speaker independent parameters and the new speaker.

Regression classes for MLLR are studied in [12, 33]. Regression classes are used to group classes that are similar enough that they could share transformations. When there are insufficient data to estimate a transformation for a particular component distribution, the transformation is derived from the data within the same regression class.

Leggetter [33] reports an average 10% reduction in word error rate on the Wall Street Journal task using a 65000 word dictionary.

2.2.2 Speaker adaptive training (SAT)

The aim of speaker adaptive training [2] (SAT) is to estimate acoustic models that are invariant to long-term variations, specifically speaker-induced variations. SAT minimizes the effects of these variations in the training data and at the same time estimates the acoustic model parameters in a unified maximum likelihood framework.

SAT can be realized in both feature space and model space. In a feature space approach, each speaker's features can be normalized towards the "neutral" speaker using an optimum transformation. This transformation can be computed within the MLLR framework. The pooled "neutral" features from all speakers are then used to estimate the model. This will result in a model which has smaller variance, since all features are transformed to the "neutral" speaker's feature space. The smaller variance results in a more precise model.

Anastasakos [2] reported that in a model space approach, speaker normalization is incorporated within a maximum likelihood speaker-independent training paradigm. During training, the speaker dependent transformations (A, b) are estimated for each speaker in the training data. The "normalized" model is updated using the inverse transform.

In most experimental tests, the SAT adaptive training technique results in reducing the average word-error by approximately 10% [2] compared to baseline systems, that incorporate MLLR adaptation.

2.2.3 Prediction adaptation

Although MLLR is a very powerful adaptation technique it does not model correlations among phonemes explicitly. With limited data available for adaptation (i.e. not all classes are necessarily represented), not all model means are adapted optimally.

Since all sounds (phonetic classes) from a speaker are produced by the same vocal tract there are certain covariant relations between different phonetic classes. An example of this relationship between different sounds can be found by studying the formant frequencies of vowels. A plot of the mean first-formant frequency ($F1$) and the difference between the first and second formant frequencies ($F2 - F1$) for vowels exhibits a pattern called the vowel-triangle [27], where the point vowels (/iy/ as in "beet", /aa/ as in "father" and /uw/ as in "boot") have extreme $F1$-$F2$ values and most other vowels have formant values close to one of the sides of the triangle. For any given speaker, if the positions of vowel /uw/ and /iy/ are known, the positions of the vowels /aa/ may be estimated relatively well. The work in this thesis explores these relations to more effectively normalize the speaker parameters.

Cox [6, 8, 7] was also concerned with this correlation. He presented an approach that uses the training data to build linear models between sounds. When only a subset of the phonemes are present in the adaptation data, all unseen phonemes (models) are adapted using this pre-computed linear relation.

The linear relations are modeled as:

$$X_n = \beta_0 + \beta_1 X_m \qquad\qquad n \neq m, \tag{2.3}$$

where X_m and X_n are mean vectors from phonetic classes m and n. The linear model parameters (β_0 and β_1) are estimated for all pairs of classes. This method assumes the relation between classes is identical for all speakers. This may not necessarily be true. For example, the differences of realizations of two vowels from speakers having different accents can be very different. In contrast, our research allows more flexibility in the relationships between classes for different speakers. Ahadi and Woodland [1] extended this method to a 1000 word vocabulary continuous speech recognition using context dependent mixture Gaussian models containing orders of magnitude more parameters. Their research shows that regression-based model prediction consistently outperforms MAP(Maximum *a posteriori*) estimation with the same amount of adaptation data.

2.2.4 Maximum *a posteriori* estimation(MAP)

In maximum *a posteriori* estimation (MAP) parameter estimation, the parameter set is chosen to maximize the *a posterior* probability

$$p(\lambda|\chi) = \frac{p(\chi|\lambda)p(\lambda)}{p(\chi)} \tag{2.4}$$

where χ represents the adaptation data with a probability distribution function given by $p(\chi)$, and λ represents the parameter set defining the distribution.

Brown first suggested using Bayesian estimation for adaptation in a connected-digit recognizer [4]. This work was further extended by Lee [30] who incorporated Bayesian adaptation within a segmental K-means framework. This was further extended by Gauvain and Lee [13] to adapt multiple mixture distributions.

Within the HMM framework, where the data is modeled as a random variable with Normal distribution, the model parameters can be adapted using maximum a-posterior training (MAP). Given the initial model mean u_0 with covariance Σ_0, the updated model parameters can be computed as [11]:

$$u' = u^N = \frac{\Sigma}{N}(\Sigma_0 + \frac{\Sigma}{N})^{-1}u_0 + \Sigma_0(\Sigma_0 + \frac{\Sigma}{N})^{-1}\frac{1}{N}\sum_{i=1}^{N} X_i \tag{2.5}$$

Here Σ represents the covariance of the adaptation data, and N the total number of examples presented. With N large this corresponds to computing the model mean directly from the speaker- specific adaptation data, whereas speaker-independent estimates predominate for small N.

$$\lim_{N \to \infty} u^N = \frac{1}{N}\Sigma_{i=1}^{N} X_i \tag{2.6}$$

In general, however, MAP corresponds to a linear interpolation between the speaker independent model mean and the speaker dependent model mean. In this method data from one class can not be used to update the model of another. To solve this problem, an extension of MAP (called EMAP) was introduced by Lasry and Stern [29] and further investigated by Rozzi [38]. The aim of the EMAP algorithm is to use the correlations between classes to estimate the transformation for unseen classes.

EMAP differs from MAP in that EMAP updates a concatenated mean of all classes simultaneously. This concatenated mean vector can be represented as

$$\mu^r = [\mu_1^r, \mu_2^r, ..., \mu_C^r] \tag{2.7}$$

where C represents the total number of classes in the recognition system. Each mean vector has D dimensions. Therefore the concatenated mean vector for speaker r, μ^r is a CD dimensional vector that is assumed to be normally distributed around *a priori* mean μ_0 with covariance matrix Σ_0. Each example of class $j \in 1..C$ across all speakers is assumed to be a random vector of a normal distribution with a mean μ_j and covariance Σ_j.

In this system, correlations amongst classes are modeled by Σ_0, and the matrices Σ_j model the within-class covariance. We define Σ as a CD by CD block diagonal matrix with Σ_j as the jth block. Σ is block diagonal due to the assumption of independent observations., i.e. individual realizations of phonemes are not correlated.

Similar to Equation 2.5 we can now compute the adapted mean vector as follows:

$$\mu' = \Sigma(\Sigma + L\Sigma_0)^{-1}\mu_0 + \Sigma_0(\Sigma + L\Sigma_0)^{-1}La, \tag{2.8}$$

where L is a diagonal matrix with the number of observations per class along the diagonal. The vector a denotes the concatenated mean vector computed from the adaptation data only.

Through the mean cross covariance matrix Σ_0, EMAP estimation is able to update all classes given any example from any class. We will later see in Chapter 3 that this mean cross covariance matrix can be interpreted as a covariance matrix for a speaker space. Because of the dimensionality it is very difficult to estimate Σ_0 accurately. To estimate this matrix accurately would require full coverage of each class (phoneme) spoken by every speaker in the training data with large number of speakers. Furthermore the matrix inversion $(\Sigma + L\Sigma_0)^{-1}$ is also expensive. Rozzi [38] extended the basic EMAP algorithm so that the computation is more efficient. These extensions, however, do not solve the problem of accurately estimating Σ_0 from a reasonable amount of data.

A combination of MLLR and MAP adaptation was recently proposed by Digalikis [9]. In this approach the transformations are used as a preprocessing step to transform the speaker independent models so that they better match the new speaker's characteristics. This provides improved prior information for MAP adaptation.

The reader is referred to Appendix A for a detailed derivation of MAP and EMAP.

2.3 This thesis in perspective

Each of the techniques presented above either adapts the speaker features or the speaker independent model to ultimately derive a speaker specific system. Methods such as MLLR adaptation have proven to be successful. However when very little adaptation data are available not all classes can be adapted optimally. Cox [7] built linear relations between classes. This allowed the adaptation of both the presented and unseen classes. This model, however, assumes an invariant relationship among classes (for all speakers), which is not valid in practice. The extended MAP algorithm provides a framework in which the correlation among classes can be used to adapt unseen models. This method however has severe data and computational constraints. Additionally, all these methods lack the ability to provide the environment to study the physical aspects of speaker variation so that the speaker variability and the mechanism of adaptation/normalization can be understood and analyzed more thoroughly.

One of the goals of this thesis is to create correlation models amongst classes and use the relationship among different sounds to improve adaptation with limited data. Another goal of this work is to study and understand the variability of the speaker differences.

We propose a new method of speaker normalization which incorporates correlations amongst classes. Using principal component analysis we construct a speaker space based on a speaker covariance matrix obtained from the training data. The speaker covariance matrix is constructed in such a manner as to explicitly describe the correlations between classes. By explicitly modeling these correlations it is possible to adapt the recognition model or normalize the speaker's features, from a very limited amount of speaker specific adaptation data.

Chapter 3

Correlation-based speaker normalization using principal component analysis

In this chapter, we introduce a new correlation-based speaker normalization and adaptation model using principal component analysis (CBPCA). This new model may be regarded as an extension of EMAP which requires less adaptation data and computation, while preserving most of the information captured in the EMAP model. This new model also allows us to explore the physical aspects of the speaker differences.

3.1 Mathematical method

Assume a set of R speakers ($\mathcal{R} = 1, 2, ..., R$), each producing at least one example of a set of C phonemes ($\mathcal{V} = 1, 2, ..., C$). Reducing the differences among speakers for these phonemes would reduce the variance among speakers and consequently increase the classification accuracy.

To construct a space that describes the speaker variation as well as taking advantage of the correlations among classes, we design a speaker space in which these aspects can be studied. We represent a speaker by a vector that is constructed by concatenation of the mean feature vector of each of the phonemes taken from the speaker:

$$\mu^r = [\mu_1^r \mu_2^r ... \mu_i^r ... \mu_C^r]$$

Here μ_i^r represents the mean feature vector for class $i \in \mathcal{V}$ for speaker $r \in \mathcal{R}$. The collection of such vectors (μ^r) from different speakers form our speaker space.

If we have C classes, and each feature vector for a class has D dimensions, then the speaker r is represented as a $1 \times CD$ dimensional vector μ^r. The covariance matrix of vectors for all speakers (Σ_0) therefore will represent the variance among speakers using the information from all the classes. The ordered set of eigenvectors (ϕ_i) of Σ_0 represent the directions in speaker space (in decreasing order) that contribute most to the variances between speakers.

The following example shows a simple instance of this idea. Let's assume there are two classes, $\mathcal{V} = \{/\text{ae}/, /\text{iy}/\}$, and the feature chosen is the average second formant value (F_2). The speaker space in our definition above will be a two dimensional vector space where the axes represent the second formant value for the classes /ae/ and /iy/ respectively. Figure 3.1 depicts the speaker space and it's eigen-directions for this example. The data used for this figure were obtained using the Hillenbrand data set (which is described in more detail in Section 4.1).

In Figure 3.1, each "+" represents a single speaker. The solid lines indicate the eigen-directions of the constructed speaker space. The eigen-direction shown as the longer solid line corresponds to the largest eigenvalue of the covariance matrix. In this example, the largest eigenvalue constituted 93.8% of the sum of all eigenvalues. This means that about 94% of the differences between speakers described in this speaker space can be described by the largest eigenvector of the speaker space covariance matrix. In practice, the few eigenvectors with the biggest associated eigenvalues would describe most of the variance. These directions are referred to as the principal components of the speaker space covariance matrix.

In this speaker space (which has CD dimensions), each speaker can be expressed as the sum of the group mean (μ_0) and the deviation of this speaker from the group mean (e^r):

$$\hat{\mu}^r = \mu_0 + \mathrm{e}^r \tag{3.1}$$

where $\hat{\mu}^r$ is an estimate of the true speaker mean μ^r.

Alternatively this model can be viewed as the sum of a speaker independent part (μ_0) and speaker dependent part (e^r). By thinking of μ_0 as the speaker independent model,

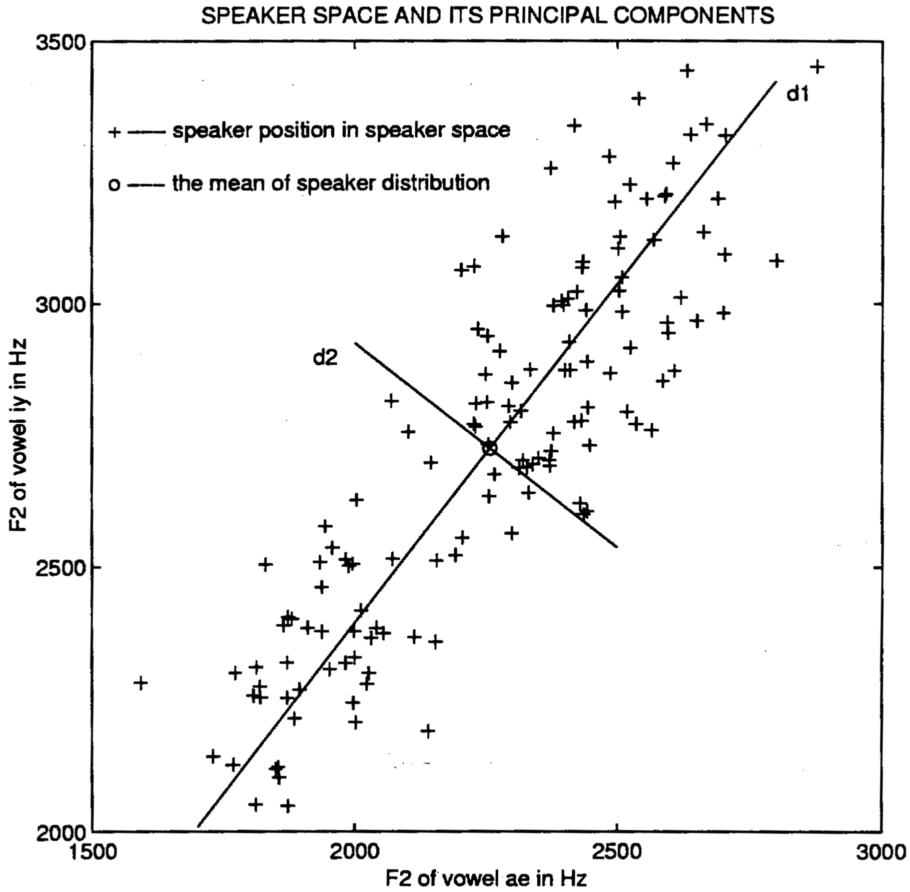


Figure 3.1: Principal components for speaker space constructed from the second formant values for the classes /ae/ and /iy/ taken from 139 speakers from the Hillenbrand data set. Each + in the figure represents one speaker. The vectors denoted as d_1 and d_2 are the eigen-directions.

the above formulation (Equation 3.1) describes a speaker adaptation process, where $\hat{\mu}^r$ represents the adapted speaker specific model.

By rewriting Equation 3.1 as

$$\hat{\mu}_0 = \mu^r - e^r \tag{3.2}$$

the model describes a speaker normalization process. The speaker specific vector e^r is used to move the speaker's data to a neutral position in the feature space denoted as $\hat{\mu}_0$.

The deviation vector e^r can be expressed in terms of the projections onto the orthogonal principal components:

$$e^r = \Phi\alpha^r = \sum_{i=1}^{CD} \phi_i \cdot \alpha_i^r \tag{3.3}$$

Here Φ is the eigen-matrix of Σ_0 and ϕ_i represents the ith ordered eigenvector. α_i^r is the projection of the speaker difference from the group mean $(e^r - \mu_0)$ onto these eigenvectors. Thus the adaptation formula (Equation 3.1) can be written as:

$$\hat{\mu}^r = \mu_0 + \sum_{i=1}^{CD} \phi_i \cdot \alpha_i^r \tag{3.4}$$

When Σ_0 has full rank, the projections onto the eigenvectors are computed as follows:

$$\alpha^r = \Phi^T(\mu^r - \mu_0) \tag{3.5}$$

If we substitute Equation 3.5 into Equation 3.4 we have:

$$\begin{aligned} \hat{\mu}^r &= \mu_0 + \Phi\Phi^T(\mu^r - \mu_0) && (3.6) \\ &= \mu^r && (3.7) \end{aligned}$$

In this scenario, the estimation is perfect.

The underlying assumption of the model we proposed here is that features which represent classes for the same speaker are correlated and that all speaker variances can be described by Φ.

3.2 Practical implementations of the principal component method

Theoretically, if examples for all classes used to construct the speaker space are available for a particular speaker, and all the eigen-components of the covariance matrix are used, perfect normalization could be achieved (i.e. the speaker vector can be mapped unto the group mean).

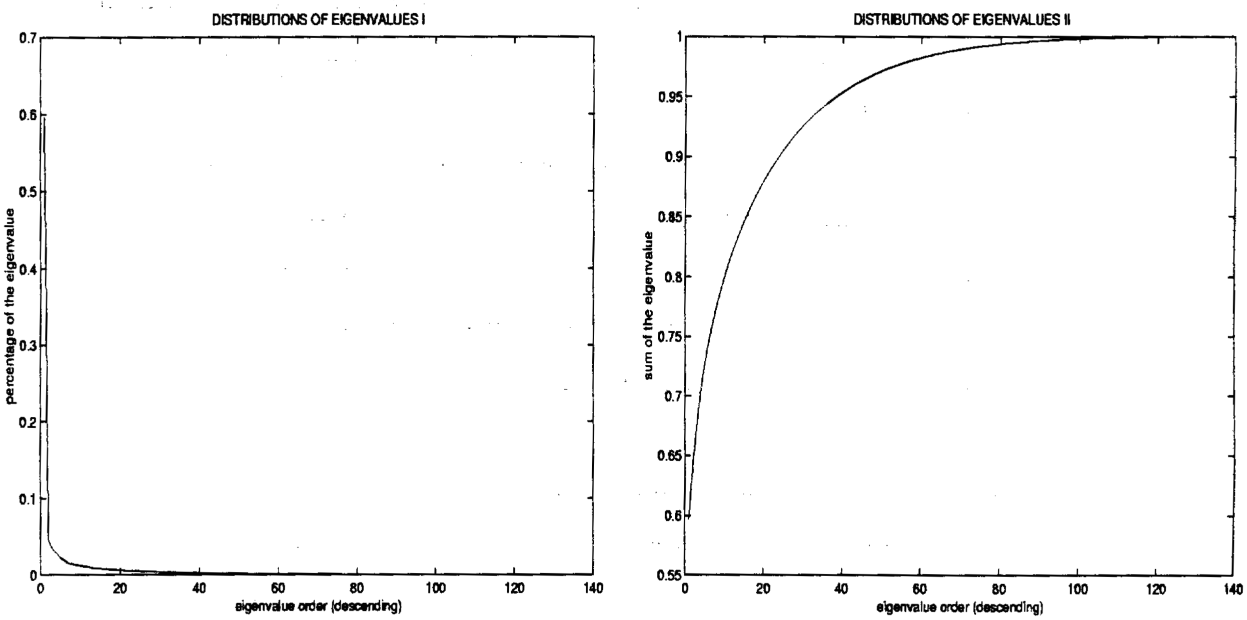
Since the principal components typically constitute the majority of the total variance of the speaker space, not all eigen-components are needed to sufficiently describe the variance among the speakers. Principal component analysis (PCA) is an optimal compression for variance. Components corresponding to the largest eigenvalues compose the subspace that contains most of the variance.

Using PCA, Equation 3.4 can therefore be approximated as:

$$\hat{\mu}^r = \mu_0 + \sum_{i=1}^{n} \phi_i \cdot \alpha_i^r \qquad\qquad n \ll CD \tag{3.8}$$

The number of eigenvectors (n) needed is determined by the contributions of the dominant eigenvectors to the variance between speakers. Selecting only the first few principal components has the further advantage that less data are needed for model adaptation. The projection coefficients α_i^r might not be perfect for the directions not included. However, since the principal components are calculated from the covariance matrix, they preserve the correlations among classes. Therefore α_i^r is still a good estimation when not all the data are available. When $n \ll CD$, the speaker space is a subspace of the original eigenspace. Figure 3.2 shows a typical eigenvalue distribution for a speaker space where the top few principal components describe most of the variance. In this particular example, we used the Hillenbrand data set [18] which consists of detailed hand-labeled formant trajectories for 12 vowels from 139 speakers.

During recognition a test speaker provides speech for adaptation (for supervised adaptation, the content of the speech is known; for unsupervised adaptation, the recognizer output is used to estimate what was spoken). Models are adapted using this part of the data and then used to recognize the test speaker's data. With limited adaptation data,


DISTRIBUTIONS OF EIGENVALUES I

DISTRIBUTIONS OF EIGENVALUES II

(a)

(b)

Figure 3.2: The distribution of the eigenvalues. Figure (a) shows the percentage of each eigenvalue (in descending order) of the sum of all the eigenvalues. Figure (b) shows the percentage of the sum of the top n eigenvalues out of sum of all the eigenvalues.

not all the classes used in constructing the speaker space are available for a particular speaker. In this condition however, we can estimate the projections (α_i^r) using the limited available data as follows:

$$\hat{\mu}^{r\prime} = \mu_0' + \sum_{i=1}^{n} \phi_i' \cdot \alpha_i^r \qquad\qquad n \ll CD, \tag{3.9}$$

where $\hat{\mu}^{r\prime}$, μ_0' and ϕ_i' represent a subspace of the constructed speaker space. In this scenario the projection coefficients α_i^r can be viewed as being class independent.

The normalization process consists of the following steps:

- First, we construct the speaker space and compute the principal components (ϕ_i) as the first n eigenvectors of the speaker space covariance matrix Σ_0 using data from many training speakers. It defines the subspace where we can model the speaker specific variances.

- Second, we use the new speaker's adaptation data to compute the projection (α^r) of the speaker's deviation from the group mean onto the principal components by solving the set of linear equations defined by Equation 3.9. α^r together with Φ describe the speaker specific between class relations.

- Finally, for the test data (i.e. previously unseen data), we use the projection coefficients to estimate the deviation from the group mean (e^r), from which we obtain the normalized model (vector in speaker space) $\hat{\mu}_0$ used for recognition. In an ideal scenario $\hat{\mu}_0$ will be equal or sufficiently close to the speaker space group mean μ_0.

In order to determine the realistic impact of these compromises, we conducted an experiment on the Hillenbrand data set using only the first three principal components. Figure 3.3 depicts the normalization results. In this experiment, we construct the speaker space using all the vowels and use all the vowels except the data for the classes /iy/ and /ae/ to estimate the projections coefficients α^r. This figure shows that although during normalization the data for /iy/ and /ae/ were not used in the estimation of α^r and only the first three principal components were used, significant normalization is achieved. In the figure, "+"s represent the original positions in speaker space and "o"s represent the

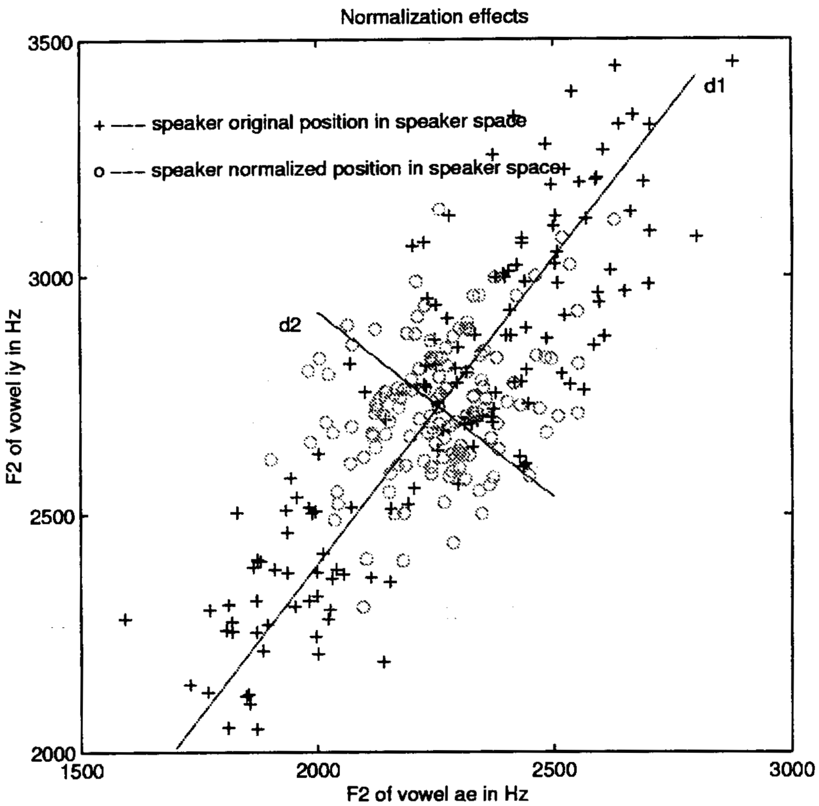


Figure 3.3: Normalization using three principal components and with partial data taken from the Hillenbrand data. "+"s represent the original positions in speaker space and "*o*"s represent the speaker normalized positions. The normalized speaker vectors have a smaller variance than the original.

speaker normalized positions. The variance of the normalized data is much smaller. This is strong evidence that our proposed method might work. In Chapter 4 we investigate the possible physical meaning of this approach and evaluate this method on various vowel classification tasks.

3.3 Discussion

The speaker space defined in this section is a "concentrated" subspace of Σ_0 used in the EMAP algorithm described in Chapter 2. The underlying assumption of this method is that most speaker variances can be described within the "concentrated" subspace of Σ_0. The advantage of this method over EMAP is that it reduces the complexity of the estimation problem for Σ_0. The top n eigenvectors of Σ_0 $(n \ll CD)$ can be estimated accurately using much less training data using singular value decomposition [26]. This method is similar to the computation of eigen-faces in image recognition [24]. At the same time, the proposed model retains most of the important information represented in Σ_0. Our work was first published in [19]. Similar method was proposed independently by Kuhn [25].

Comparing to regression-based prediction adaptation, our research allows more flexibility in the relationships between classes for different speakers. While other adaptation/normalization methods (such as MLLR) may lead to good performance, an important advantage of the proposed method is that it provides an environment in which various physical aspects of the speaker differences can be studied, and can thus offer insight into the underlying mechanism of the adaptation/normalization process. Since the importance of each eigen-component is indicated by the value of the corresponding eigenvalue, one can analyze which directions in speaker space are the most significant contributors to between-speaker variability. The results of these studies are presented in Chapter 4.

Chapter 4

Principles and properties of CBPCA

The experiments and discussion presented in this chapter are aimed at developing an understanding of the principles behind correlation based adaptation. In Section 4.2 we study the mechanisms on which the proposed adaptation model is built. Section 4.3 explores the underlying physical meaning of the speaker space covariance matrix. In particular, we concentrate on gender-related variations and show how these are modeled by the principal components of the speaker space covariance matrix. In Section 4.4 we compare CBPCA with EMAP and show how these two methods compare in the use of the underlying speaker covariance matrix (Σ_0).

4.1 Data sets

All experiments presented in this chapter and the following chapter are based on the following data sets.

1. **Hillenbrand:** The Hillenbrand data set [18] is an extension of the Peterson & Barney data [36], collected at the University of Michigan. In this data set, the first four formant ($F1$-$F4$) contours are measured for 12 vowels in the /h-V-d/ context [1]:

   `ae ah aw eh ei er ih iy oa oo uh uw`

   The database consists of speech from 45 men, 48 women, and 46 children. This data set has extra dynamic information compared to the Peterson and Barney data.

[1] Each vowel has phoneme "h" as the left context and phoneme "d" as the right context"

2. **TIMIT:** The TIMIT data set [28] is a phonetically hand-labeled data set which contains phonetically-balanced sentences. The entire corpus consists of 10 sentences recorded from each of 630 speakers of American English. In our training set (the training set defined by NIST), there are 462 speakers which consists of 326 men and 136 women; in our development set, there are 50 speakers consisting of 32 men and 18 women. For the TIMIT experiment we chose the following 10 vowels:

   ae ao eh ey er ih iy ow uh ah

   These classes were chosen to best match the Hillenbrand data set.

3. **NTIMIT:** The NTIMIT data set is obtained by sending the data from TIMIT through 8 different telephone channels.

In our study, we used Mel scale cepstral coefficients (MFCC) and formant frequency values as our feature sets. The three formant frequency values for each vowel in the TIMIT data set are estimated using a formant estimation method proposed by Welling and Ney [42].

4.2 Mechanism underlying CBPCA

In this section, we study some issues related to the mechanisms which cause CBPCA to adapt to speaker variability. In particular, we study how correlated different phonemes are and how well we can estimate the dominant principal components without a large number of speakers. Finally, we investigate how different phoneme groups contribute to the variations among speakers.

4.2.1 Correlation among classes

The underlying assumption of correlation-based principal component adaptation is that classes for the same speaker are correlated and that speaker variances can be described by the eigenvectors of the speaker-space covariance matrix (Φ). Because the adaptation of unseen classes depends on theses correlations, we expect a direct correspondence between these measurements and the ability of the proposed model to adapt to speaker differences [2].

Table 4.1: Average correlation coefficients among classes for different data sets and different features.

Task	Feature	Whole matrix	Within phoneme	Strip diagonal
Hillenbrand	formant	0.5058	0.4825	0.6737
TIMIT	formant	0.1893	0.1800	0.4378
TIMIT	MFCC	0.1894	0.2034	0.4329
TIMIT (BW=4kHz)	MFCC	0.1612	0.1925	0.3055
NTIMIT(BW=4KHz)	MFCC	0.1402	0.1857	0.2786

Table 4.1 presents the average absolute value of the off-diagonal correlation coefficients. The values are computed for different data sets and different features as indicated in the table. Large value of the correlation coefficient indicates there's strong correlation. In the table, the rows with "BW=4kHz" indicates the signals in those tasks have a bandwidth of 4000 Hz.

In Table 4.1, "whole matrix" corresponds to the values averaged over all the off-diagonal correlation coefficients; "within phoneme" are the average of the correlation coefficients for the same phoneme; "strip-diagonal" are the average of correlation coefficients which represent the correlation of the same dimension of feature but across different classes; these are the coefficients that are parallel to the overall diagonal, as illustrated in Figure 4.1.

The results show that the correlations are very strong. In particular the Hillenbrand data has a very high "strip diagonal correlation". Because the Hillenbrand data do not contain any contextual variation, the speaker space constructed from these measurements is dominated by speaker variances, whereas a combination of speaker and contextual variance is expected to dominate in the TIMIT-based corpora. The results presented in the next section on normalization effects using formants as well as the vowel classification results presented in the following chapter show significant adaptation gained from this high correlation among classes.

[2] The experiments on vowel classification (Chapter 5) and hidden Markov modeling (Chapter 6) confirm this observation.

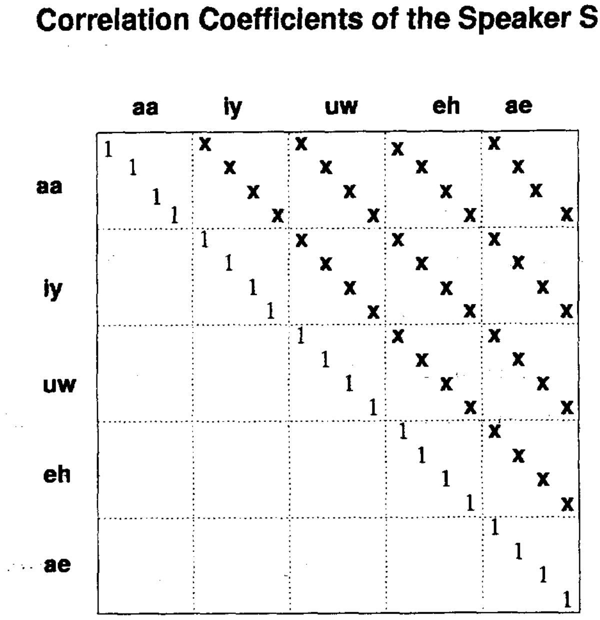
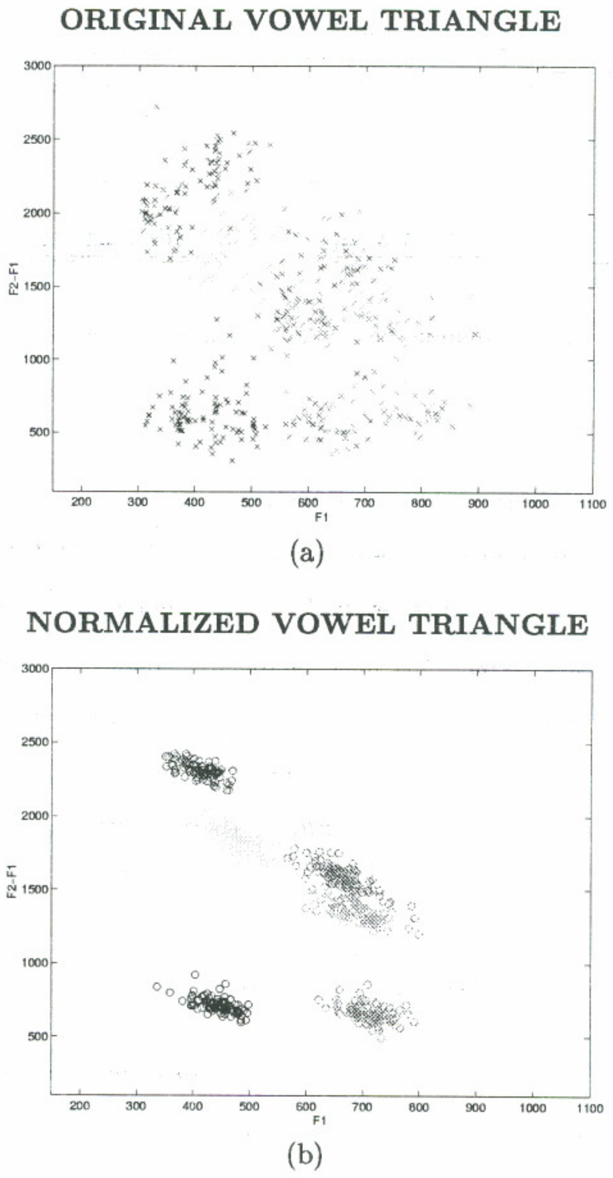

Correlation Coefficients of the Speaker Space

Figure 4.1: Illustration of the correlation coefficients representing the correlation of the same dimension of feature but across different classes. The "X" represents the "strip diagonal" coefficients.

The correlation decreases as more sources of variance enter the feature measurements. For example, the vowel classes from the TIMIT data contain both contextual variance and formant estimation errors. This results in a decrease in the correlation, which in turn affects the potential adaptation gain. Channel differences also influence the correlation as shown by the NTIMIT results. On NTIMIT, the MFCC features are energy normalized and channel normalization is implemented with cepstral mean subtraction.

4.2.2 Normalization effects on formants

In phonetic studies [27], a vowel triangle is often used to describe the relationships between vowels for a single speaker. That is, the locations of the vowels in the plane formed by plotting F1 versus F2-F1 is triangular in shape. The position of this triangle shifts between speakers, but its overall shape is rather constant. In this experiment, we examine the effects of normalization using CBPCA on the formants using the vowel triangle generated from the Hillenbrand data set.



ORIGINAL VOWEL TRIANGLE

(a)

NORMALIZED VOWEL TRIANGLE

(b)

Figure 4.2: The vowel triangle for all speakers before (a) and after (b) normalization. After the normalization, the vowels are more separable.

Figure 4.2 shows the vowel triangles (F1 versus F2-F1) before and after normalization for the following vowels in the Hillenbrand data:

`iy ih eh ae uh uw`

During normalization 30 eigen-directions are used. The figures show that normalization reduced the variance of each class and the vowels are therefore more separable. Here we can see the effect of the high correlation among classes. Since the constructed speaker space models only speaker variances the normalization is able to extract the speaker variance which results in the classes being almost totally separable. This indicates the potential of the proposed normalization model.

4.2.3 Estimation of eigenvectors

The adaptation and normalization of the speaker's features are based on the dominant eigenvectors of the speaker space covariance matrix (Σ_0). Instead of estimating the full covariance matrix, which requires data from a large number of speakers, we estimate the top n eigenvectors directly using singular value decomposition [26].

The assumption behind this estimation method is that, if the variance in the speaker space can be described by only a few dominating eigenvectors, it needs much less data samples to estimate these eigenvectors than that needed to estimate the full covariance matrix. We test the relative accuracy of this estimation method on a five-dimensional simulation.

In this simulation we generate five dimensional features with two dominant eigen-directions. The simulation experiment is designed to study the estimation accuracy of the dominant eigenvectors when only a small number of data samples are available.

The simulation experiment is conducted as follows: First we generate 500 samples of 5 dimensional data with 2 dominant eigenvectors. Using all data points we compute the full covariance matrix (Σ_0) and associated eigenvectors (Φ). Next we randomly select $n = 10, 20, 40, 80$ examples from the original 500 samples and estimate the eigenvectors using singular value decomposition. The resulting eigen-directions are then compared to those computed using all 500 data points. The comparison between the original eigenvectors and the estimated eigenvectors are represented by the norm of the difference of each eigenvector and it's estimation. The results are normalized by the norm of each eigenvector (the eigenvectors are uniform and orthogonal). The results presented are the average taken over 100 such trials.

Table 4.2: Simulation results for the estimation of the principal components.

	eigenvector order				
#samples(n)	1	2	3	4	5
Eigenvalue	59.1501	23.8899	0.3151	0.0850	0.0297
10	0.2907	0.6246	0.9022	1.2635	1.2252
20	0.1607	0.1898	0.8573	1.0881	1.1952
40	0.1474	0.1606	0.8224	1.0234	1.2013
80	0.0819	0.0829	0.8601	1.0501	1.2358

Table 4.2 depicts the results obtained. With limited data samples, the higher eigen-components always get bad estimations, but the dominant eigenvectors can be estimated relatively well. These results show that with a relatively small number of data samples, we can estimate the principal components reasonably well. We therefore do not need to estimate the full covariance matrix (Σ_0), and thus avoid issues of sufficient data and computational requirements associated with computing the corresponding eigenvectors (Φ).

4.2.4 Relative contributions of different groups of phonemes

The following experiment studies the relative contribution of different phoneme groups in adaptation and normalization. We construct the speaker space using all phonemes in the TIMIT dataset and compare the adaptation results when only a subset of phonemes are used.

Table 4.3 list the phoneme groups used in this experiment.

For each phoneme group we compute the phoneme classification performance using (a) only the phonemes in the specific group to estimate the eigenvectors (ϕ_i^r) and associated speaker-dependent attributes (α_i^r) and (b) using all phonemes. Table 4.4 depicts the results obtained. For each phoneme group classification results are presented for (a) phonemes in

Table 4.3: Phoneme groups used to study the most important contributors to normalization.

phoneme group	phonemes
sonorant	aa ae ah ao aw ax axr ay eh el em en er ey ih ix iy l ow oy r uh uw ux w y
consonant	b ch d dx dh f g hh jh k m n ng p s sh t th v z zh
vowel	aa ae ah ao aw ax axr ay eh er ey ih ix iy ow oy uh uw ux
nasal	m n ng em en
stop	b d g p t k dx
fricative	ch dh f hh jh s sh th v z zh

the group, (b) phonemes that do not belong to the specific group and (c) all phonemes. Here the rows labeled as "other" refer to the classification results for phonemes that do not belong to the group.

Using all phonemes to construct the speaker space results in the greatest performance gain (35.20% − > 39.15%). The contribution seems to be equally spread between the sonorant and consonant phoneme groups. However when each of these broad categories are broken into smaller sub-groups we can see that the relative contribution stems mostly from the vowel and nasal sub-groups. When only fricatives or stops are used they do not help to estimate adaptation parameters well, and this ultimately results in a performance degradation (e.g. using fricative only as adaptation data the other phonemes' classification performance went down: 31.45% − > 30.27%).

4.3 Physical interpretations of the speaker space

In order to understand the physical characteristics utilized by our adaptation model, we investigate its relationship with the speech production process. The experiments presented in this section show that our adaptation model can capture some known variances among speakers.

Table 4.4: Classification results (in percentage) using different phonemes groups to estimate the eigenvectors of the speaker space covariance matrix.

	baseline	adaptation	
phoneme groups		all phonemes	phoneme group
sonorants	31.69	35.09	35.19
other	39.55	44.18	43.60
all phonemes	35.20	39.15	38.95
consonants	39.55	44.18	45.22
other	31.69	35.09	33.27
all phonemes	35.20	39.15	38.62
vowels	32.99	36.85	37.05
other	36.76	40.77	40.22
all phonemes	35.20	39.15	38.91
nasal	35.20	34.26	35.36
other	36.39	39.74	36.66
all phonemes	35.20	39.15	36.52
fricatives	51.03	54.72	56.40
other	31.45	35.46	30.27
all phonemes	35.20	39.15	35.27
stops	34.65	37.09	38.17
other	35.31	39.54	30.57
all phonemes	35.20	39.15	31.75

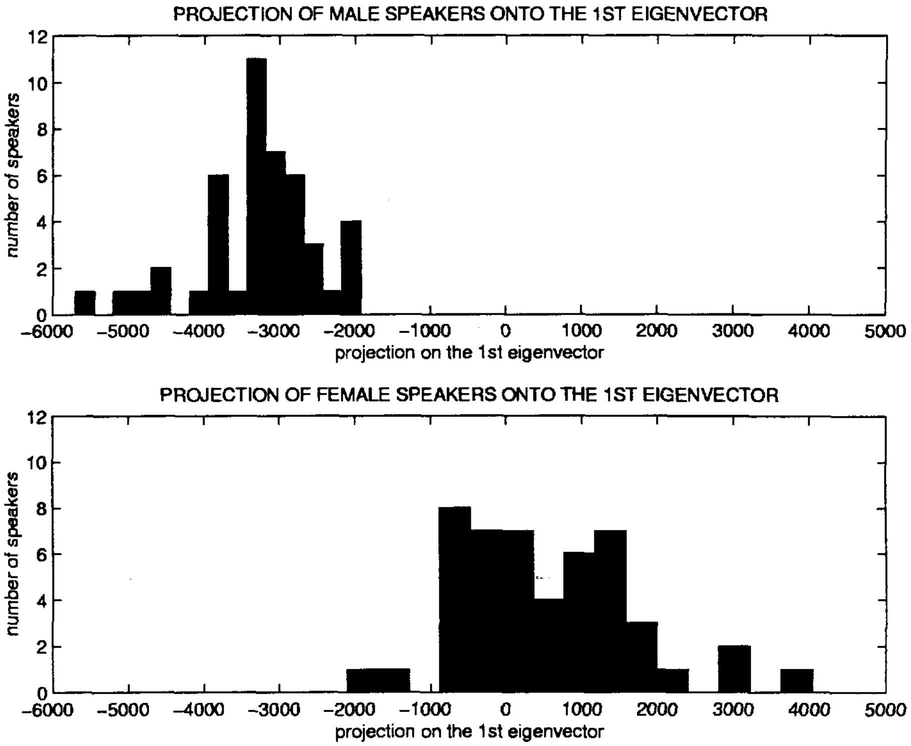


Figure 4.3: The distribution of projections from male and female speakers onto the 1st eigenvector using the Hillenbrand data set. The upper figure shows the distribution of projections for male speakers. The lower figure shows the distribution of projections for female speakers. This projection almost separates male and female speakers.

4.3.1 Gender differences

Since the most distinct difference among speakers is the gender difference, we test if our model can capture this variance.

We study this problem by investigating the distributions of projections (p^r) from speakers' features (μ^r) onto the first eigenvector (ϕ_1) of the speaker space covariance matrix:

$$p^r = \phi_1^T \mu^r \tag{4.1}$$

The distributions of p^r for male and female speakers from the Hillenbrand data set are plotted in Figure 4.3. The figure shows that the projection almost separates the male and female groups. This suggests that the first principal component is highly correlated with the gender of the speaker. Since the main physical difference between male and female speakers is vocal tract length, this result suggests that the first principal component of the speaker space also correlates to the vocal tract length.

Similar experiments using the TIMIT data set and MFCC features yield the same conclusion.

4.3.2 Pitch

It is also known that male and female speech differ in pitch. The following experiments are conducted to study the effects of shifting a particular speaker's pitch without changing the spectrum. In this experiment, 10 vowels taken from the TIMIT data set are generated for two synthesized voices (one male "mwm" and one female "tll") [3]. The speech is synthesized by concatenating pre-recorded speech . The synthesizer allows pitch shifting and duration change. Each voice is then modified to change the pitch without changing the spectrum. The effects of the resulting projection on the first eigenvector of the speaker space are shown in Figure 4.4. MFCC features are used in this experiment. This study shows that the first eigenvector is also related to the pitch of a speaker.

In Figure 4.4, the top graph shows the distribution of the projections on the 1st eigenvector from male speakers and the lower one shows the same for female speakers in the TIMIT training set using MFCC features. The symbol "o" depicts the position of the original synthesized voice in the distribution and the symbol "*" represent the same voice with changed pitch. While changing the pitch, the male voice is transformed by using the female pitch and the female voice using the male's pitch. The figure shows that after the pitch change both voices move towards the neutral region. During the pitch-modification process the spectral envelope of the original speech remains unchanged. However, the cepstral coefficients (MFCC) do to some degree contain pitch information. The resulting change of position of the speaker in the distribution of the projection onto the first eigenvector reflects this information. This result shows that the first eigenvector is correlated with the pitch value of the speaker. Another observation from Figure 4.4 is that, even after the pitch is changed to that of the female speaker "tll", the position of

[3] using Festival [3, 34] with OGI extensions [40]

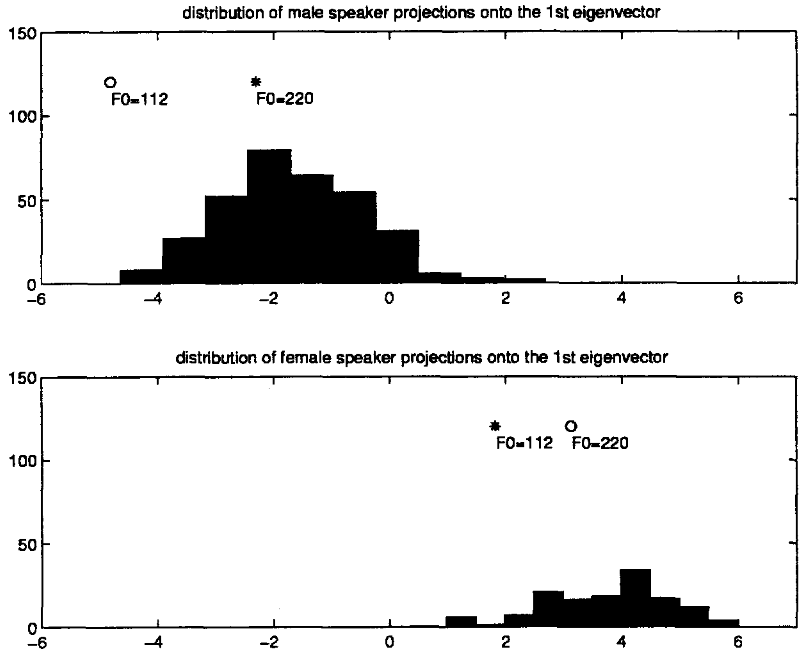


Figure 4.4: Effects of changing speaker's pitch on the projection of the first eigenvector. The histograms are the distribution of the projections onto the first eigenvector of male (or female) speakers in the TIMIT data set. The symbol "o" depicts the original synthesized voice in the distribution and the symbol "*" represent the same voice re-synthesized with changed pitch. The upper graph shows the effects of changing a male voice by using a female (higher) pitch. The effect is to move the voice's projection towards the female speakers' distribution. The lower graph shows the effects of changing a female voice by using a male (lower) pitch. Consequently, the changed voice's projection moves towards the male speakers' distribution.

speaker "mwm" in the distribution is still much different from that of "tll". This shows that pitch is not the only factor that defines the difference between the two speakers.

From experiments in this section, we know that the first eigenvector correlates with the gender (independent of whether pitch information is included or not); therefore, we expect to reduce the influence of gender variations by using the 1st eigenvector in adaptation. Experimental results presented in Section 5.5 validates this observation.

4.3.3 Other physical meanings of speaker space

Other studies of physical meaning [25] found a weak correlation between the speaking rate and the second eigenvector. Also, the third eigenvector of the speaker space was found to correlate with short steady state portions of vowels relative to the onsets and off glides. However, since we are using different features from Kuhn [25], we can not verify the suggestions (the physical meaning of the second and third eigenvectors depends on the features used). There are many different variances in the feature representation, so the eigenvectors can be a combination of various variance sources, and it is difficult to identify which one each eigenvector represents.

4.4 Comparison between EMAP and CBPCA

In order to compare EMAP with CBPCA, we examine the mathematical relationship between the two methods and also compare the two methods on a basic vowel classification task.

We denote by Φ the eigen-matrix of Σ_0 in Equation 2.8. Using Φ to rotate the data results in the following covariance matrix:

$$\Phi^T \Sigma_0 \Phi = \Lambda \tag{4.2}$$

where Λ is a diagonal matrix with the diagonal elements corresponding to eigenvalues of Σ_0. Linear transformation of Equation 2.8 with Φ results in:

$$\Phi^T \mu' = \Phi^T \Sigma \Phi (\Phi^T \Sigma \Phi + L\Lambda)^{-1} \Phi^T \mu_0 + \Lambda (\Phi^T \Sigma \Phi + L\Lambda)^{-1} L \Phi^T a \tag{4.3}$$

Assuming diagonal covariances for the speaker-dependent part (Σ), we define the following diagonal matrix:

$$\Phi^T \Sigma \Phi = \Lambda' \tag{4.4}$$

The term $\Phi^T \Sigma \Phi + L\phi^T \Sigma_0 \Phi$ in Equation 4.3 will therefore also be diagonal:

$$\Phi^T \Sigma \Phi + L\phi^T \Sigma_0 \Phi = A \tag{4.5}$$

The EMAP equation can now be written as follows:

$$\Phi^T \mu' = \Lambda' A^{-1} \Phi^T \mu_0 + \Lambda A^{-1} L \Phi^T a \tag{4.6}$$

Equation 4.6 can be viewed as the EMAP equation in the transformed speaker space. In contrast CBPCA uses the eigenvector matrix to move the speaker's features as follows:

$$\mu' = \mu_0 + \Phi\alpha \tag{4.7}$$

where α represents the projections of the speaker difference vector onto the eigenvectors:

$$\alpha = \Phi^T (a - \mu_0) \tag{4.8}$$

From Equation 4.6 and 4.7 we can see that although EMAP and CBPCA share a similar linear form, they are not fundamentally the same, with EMAP always taking into account the data count L which describes the prior information in the equation.

4.4.1 Similarity of the two methods

- Both are designed to use correlation among classes to achieve adaptation for classes that are not present in adaptation data.
- Both use the between-speaker covariance matrix Σ_0 to model the correlation.

4.4.2 Differences between the two methods

- EMAP corresponds to a linear interpolation between the speaker- independent model mean and the speaker-dependent model mean; CBPCA moves the speaker-independent model mean toward the speaker dependent mean directly along the directions defined by Φ.

- EMAP takes into account the covariance within each class; CBPCA only uses the between-speaker covariance matrix Σ_0 in the modeling. It does not use the prior information of speaker-specific variances for each class.

- EMAP requires a large number of speakers to estimate Σ_0 accurately; CBPCA requires much less speakers to estimate the major eigenvectors of Σ_0.

- EMAP requires inversion of a large matrix $(\Sigma + L\Sigma_0)$ for each speaker; CBPCA does not require any matrix inversion.

4.4.3 Advantage of each method

- EMAP can be very effective when there is enough data to estimate Σ_0. Furthermore, it has the nice asymptotic property that adaptation approaches speaker-dependent performance as the adaptation data increases.

- CBPCA does not have to estimate the full Σ_0. It therefore needs to estimate only a smaller number of parameters, and can provide successful adaptation with less training data.

4.4.4 Comparison between experimental results

In this section the effects of our proposed method (CBPCA) and EMAP are compared on a TIMIT vowel classification task. The features used here are the frame-level 14th order MFCC feature with dimension 15 (including the normalized energy). Frame-level features are used to meet the needs of EMAP. EMAP needs a large amount of data to estimate the adaptation parameters accurately, i.e. for Σ and Σ_0 in Equation 4.3. Table 4.5 summarizes the classification results obtained using supervised adaptation. A single mixture Gaussian classifier is used in the experiment as baseline classifier. In Table 4.5, the column "% of variance" describes the percentage of the total variance described by the eigenvectors used in the experiment.

Table 4.5 shows that with good knowledge of the adaptation data (supervised or good baseline system), CBPCA can achieve better adaptation results. The reason is that CBPCA is designed to move the models parameters towards the target directly while

Table 4.5: Comparison results of adaptation performance on vowel classification accuracy using CBPCA versus EMAP.

Method	# param	% variance	error rate	error reduction
baseline			44.7%	
EMAP	150		33.7%	17.9%
CBPCA	3	44%	40.4%	9.6%
CBPCA	10	63%	37.6%	15.9%
CBPCA	20	75%	36.8%	17.9%
CBPCA	30	82%	34.8%	22.0%
CBPCA	100	98%	27.0%	39.6%

EMAP results in an interpolation between models parameters that describe the training data and model parameters of the adaptation data.

Next, we test the performance using different amounts of adaptation data. On the TIMIT data set, each speaker speaks 8 sentences. We use $n, (n < 8)$ sentences as adaptation data and the remainder ($8 - n$ sentences) as test data. This experiment is conducted in supervised adaptation mode. Figure 4.5 depicts the classification results obtained using varying amounts of adaptation data. The irregular curve in the figure is caused by the fact that the adaptation data and test data for this experiment are not chosen by full jack-knifing, so that the test set differs for different n.

The results show that CBPCA consistently outperforms EMAP in this task. The potential reason for this is that: (1) estimate of full correlation matrix in EMAP is not sufficiently accurate; (2) EMAP has much stronger parametric assumptions than CBPCA (that observations of different phoneme's are independent,) which is not valid. The results also indicate that given accurate information about the adaptation data, CBPCA has the advantage of "faster" adaptation.

4.5 Summary and Conclusions

The preliminary experiments in this section show strong evidence that the principal components of the speaker space are highly correlated with the vocal tract length, which contributes a major part of speaker differences in the speech signal. Studies in Section 4.2

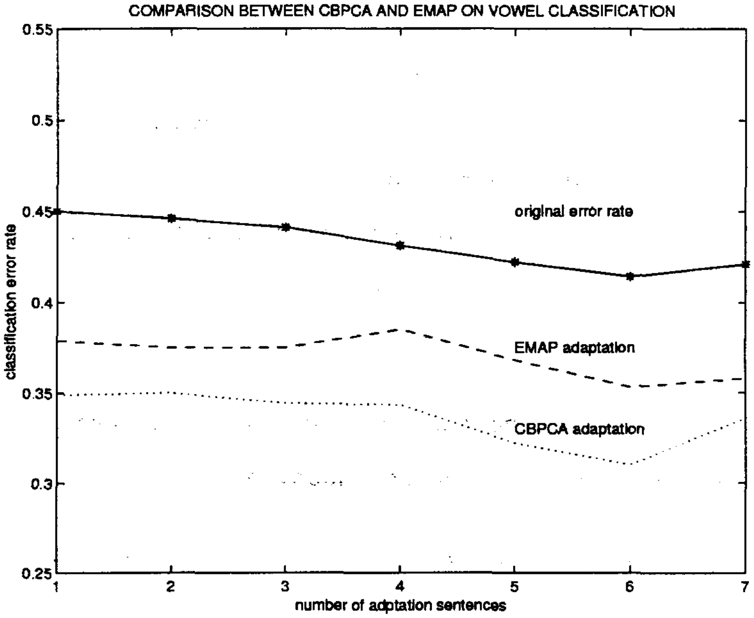


Figure 4.5: Comparing CBPCA and EMAP using varying amounts of adaptation data. The test sentences are different for different numbers of adaptation sentences. 30 eigenvectors are used for CBPCA.

show that features for different classes are correlated, and the dominant principal components can be estimated reasonably well using a relative small number of speakers.

Comparing to the previous work described in Chapter 2, this proposed method does not assume a class-independent transformation, but a class-covariant transformation. The model is constructed in a compact way so that it is able to adapt using limited data and require much less computation. Compared to EMAP, the proposed method uses similar information, reduces the computational complexity of Σ_0, and with same amount of adaptation data, can adapt more accurately. Finally, by analyzing the dominant eigen-directions, one can gain some understanding of the underlying physical meaning of the adaptation.

Chapter 5

Vowel classification

In this chapter, we test the proposed method on vowel classification tasks using Gaussian mixture classifiers (D dimensions, single mixture per class). In these experiments we choose to use the adaptation form (Equation 3.1) rather than the normalization form (Equation 3.2). Since the current paradigm models the deviation of a class from the group mean using a single vector, adjustment of the model and the feature vector are mathematically equivalent. Adaptation and normalization are therefore interchangeable in this paradigm.

5.1 Vowel classification on the Hillenbrand data set

For each of the experiments the feature vector chosen consists of: (1) the average pitch $F0$, (2) the formant values $F1$ through $F4$ sampled at the 20% , 50%, and 80% intervals of each segment, and (3) the log duration of the vowel segment. In these experiments a jack-knifing method was used to create a series of test sets. When constructing the speaker space and corresponding principal components (ϕ_i), we use all but the data from one speaker. The remaining speaker is then used for testing. During testing, a subset of all vowels (for example, 3 out of 10) for the test speaker are used to estimate the speaker attributes α^r, and the remaining vowels are tested. Each speaker in the data set was chosen to be the test speaker once, and all vowels for the same speakers are rotated through the estimation and test set as explained above. This experiment uses the supervised static adaptation mode. Results presented here are the average of the jack-knifing experiments. In this experiment, results are obtained by using a single mixture Gaussian classifier.

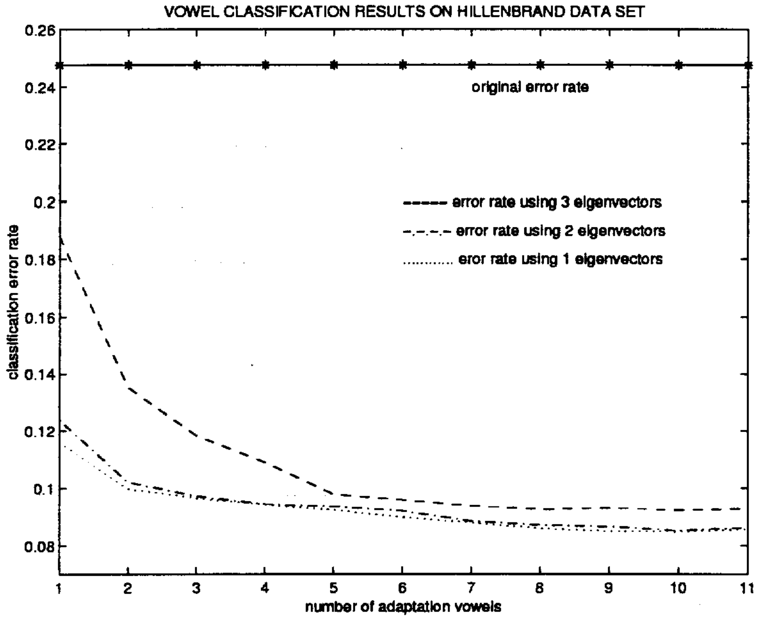


Figure 5.1: The classification results of adaptation on Hillenbrand data using different amounts of adaptation data and different numbers of eigenvectors.

The results are shown in Figure 5.1. Normalization can indeed be seen to improve the classification accuracy when even only one principal component is used. As expected, the performance improved with additional vowels used during the adaptation; however, with only a few eigenvectors and a few vowels used for adaptation, the error rate can be reduced significantly.

In this example, the first eigen vector describes most of variance (>70%), the rest of the eigenvector describe much less variance (4% for the second eigenvector). It requires more data samples to estimate the adaptation parameters that corresponds to the second and third eigenvectors, etc. The fact that using more than two eigenvectors does not improve the classification performance indicates that there are too few examples to estimate the adaptation parameters accurately.

5.2 Vowel classification on the TIMIT data set

In this experiment, we intend to test the effectiveness of our adaptation model on a more realistic data set and using a more widely used feature set, namely Mel Frequency Cepstral coefficients (MFCC). Cepstral-based analysis (as used in MFCCs), is more robust than formant tracking, while sharing some of its attractive features (insensitivity to pitch variations, emphasis on perceptually important aspects of speech). The task includes classification of 10 broad-category context- dependent vowels in TIMIT:

obs<ae>obs obs<ao>obs obs<eh>obs obs<ey>obs

obs<er>obs obs<ih>obs obs<iy>obs obs<ow>obs obs<uh>obs obs<ah>obs

where **obs<ae>obs** represents /ae/ in the context of an obstruent on each side. The following obstruents are allowed:

p k t b d g hh hv z s th zh sh f ch jh

We choose these vowels to match the vowels in the Hillenbrand data set and choose the context to get enough samples of each vowel for each speaker from the TIMIT training set while restricting the contextual variation.

For each of the experiments the feature vector chosen consists of: (1) the average 14 MFCC coefficients over a segment; (2) the average energy over the segment and (3) the log duration of the vowel segment.

The restriction to obstruent contexts obviously is less successful in limiting contextual variation than in the previous experiment. Therefore, the principal components of speaker space will not only represent the differences among speakers but also the differences caused by the variance in context. The classification results are depicted in Figure 5.2.

The following observations can be made from these results:

- Using CBPCA for adaptation can achieve performance gain under more difficult conditions (i.e. with other variance sources present in the feature representation)

- By using only a few principal components, CBPCA adaptation can achieve significant improvement in accuracy (24.6% error reduction).

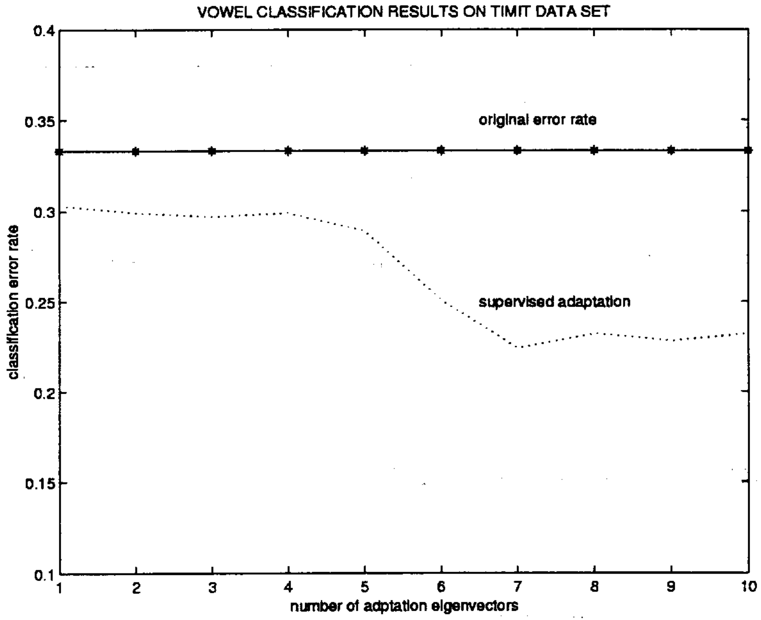
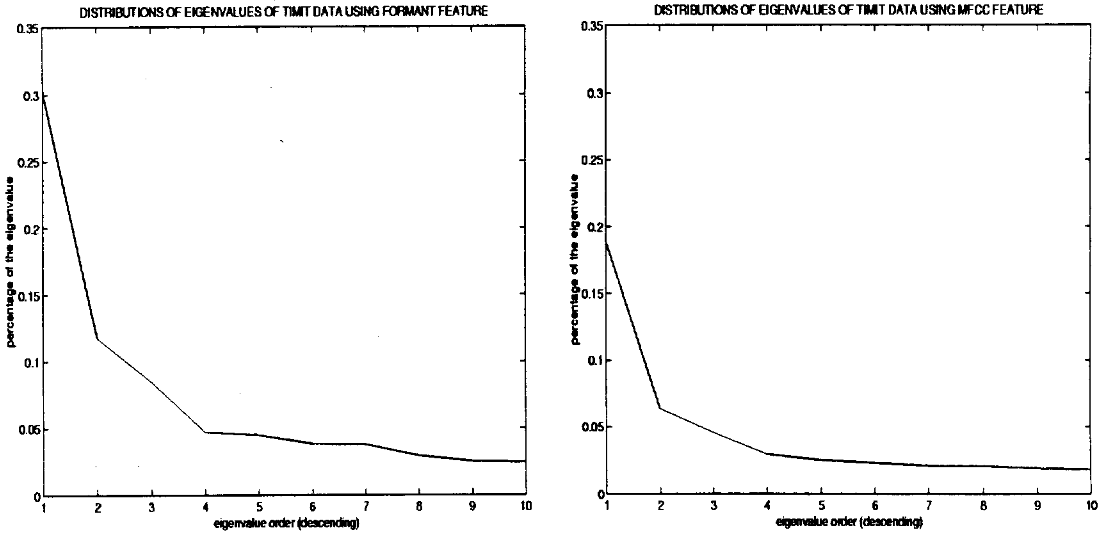


Figure 5.2: The classification results of adaptation on TIMIT data using different numbers of eigenvectors.



(a) (b)

Figure 5.3: On TIMIT Data: The distribution of the eigenvalues. Figure (a) shows the percentage of each eigenvalue (in descending order) of the sum of all the eigenvalues when using formant features. Figure (b) shows the percentage of each eigenvalue (in descending order) of the sum of all the eigenvalues when using MFCC features.

Again, note that increasing the number of eigen-components beyond 7 does not further increase recognition accuracy. This suggests that only a small number of principal components carry significant speaker information. Figure 5.3 plots the relative contribution of each eigenvector (principal component) to the variance of the constructed speaker space. From the classification results presented in Figure 5.2 and the relative contribution, we can see that the lower principal components do not contribute significant extra information.

5.3 Unsupervised adaptation

So far all the adaptation experiments we have conducted are done in supervised mode. This means that during the adaptation of the models, speech data with appropriate labels are available for adaptation. In a realistic scenario, the correct words/phonemes would not be known for a new speaker. Instead, recognition results from a first pass (using speaker-independent models) are used as labels instead. This is called unsupervised adaptation.

In our correlation model, we can estimate the adaptation parameters (α_i^r) reasonably well even if the baseline system made mistakes. This conclusion is based on the fact that the most confusable classes are the ones that share similar acoustic features. If this assumption is true, the proposed method should perform well in an unsupervised adaptation mode.

We performed unsupervised adaptation on the TIMIT data set. In this process, the speaker-dependent attributes α^r are calculated by using the examples that are classified by the baseline classifier instead of using the hand-labels. The results are presented in Figure 5.4 and summarized in Table 5.1:

Table 5.1: Classification performance (error rate) comparison between unsupervised adaptation and supervised adaptation. In the results presented, 6 eigenvectors are used.

data set	baseline	supervised	unsupervised
TIMIT	33.3%	25.1%	28.4%

This result shows that unsupervised adaptation can improve the classification performance at the cost of some reduced performance, and shows the potential for a real-world

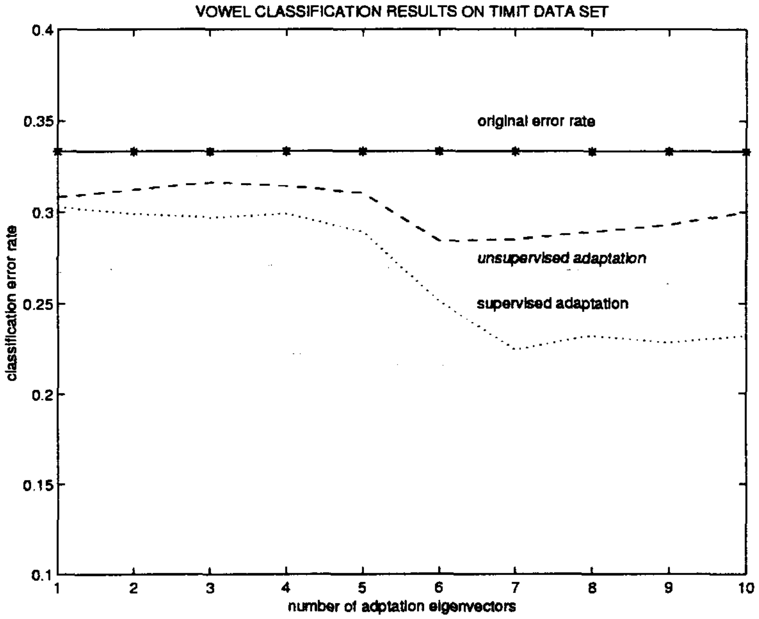


Figure 5.4: The classification results of adaptation on TIMIT data using different numbers of eigenvectors in different adaptation modes.

implementation of the proposed method.

These results also show that our assumption that classes which share similar feature spaces can share similar adaptation parameters is correct and that the proposed method is suitable for unsupervised adaptation.

5.4 The effects of normalization using different features

In this section, we compare the effects of normalization/adaptation using CBPCA on different features. The features we compare are "trajectory features" and "steady state" features. Trajectory features are features that describe the trajectories of the spectral movements. "Steady state" features describe the spectral property at the steady states of a phoneme - that is, the portion of a phoneme where the spectrum stays relative steady. Our experiments are performed on the Hillenbrand data set. The results (classification error rate) are shown in Table 5.2:

From Table 5.2 we can see that adaptation on systems based on trajectory features

Table 5.2: Comparison results of adaptation performance on vowel classification error rate using trajectory features versus steady state features.

feature type	steady state			trajectory		
data set	baseline	norm	error reduction	baseline	norm	error reduction
Hillenbrand	30.4%	20.3%	33.2%	24.8%	8.5%	65.7%

gives a substantially higher gain in performance. This suggests that dynamic information carried by trajectory features contains relatively more speaker-dependent information that can be captured with our adaptation model.

5.5 Gender-dependent modeling and CBPCA adaptation

Gender-dependent modeling uses training data from speakers of the same gender to obtain the acoustic models for that specific gender. The resulting gender-dependent models only describe the variations within each speaker group with the same gender. This experiment compares the effect of normalization to that of gender-dependent modeling to test if the adaptation merely achieves the same effect as the gender- dependent models. The experiments are conducted on both the Hillenbrand and the TIMIT data sets.

The results (classification error rate) shown in Table 5.3 indicate that normalization can indeed obtain better performance improvement than gender-dependent classifiers, with the advantage that no gender information is required during training or testing. This provides further evidence that the proposed method can capture speaker variations not modeled by the normal modeling techniques such as gender-dependent modeling.

Table 5.3: Comparison of classification performance (error rate) using normalization versus gender-dependent models. The classifiers are single mixture Gaussian classifiers. Adaptation outperforms gender-dependent models on both tasks.

data set	baseline	gender-dep.	normalization
Hillenbrand	24.8%	9.7%	8.5%
TIMIT	33.3%	30.3%	28.4%

5.6 Summary

In this chapter, we conducted various classification experiments to evaluate the effectiveness of the proposed method. With only a few eigenvectors and a few vowels used for adaptation, CBPCA adaptation reduced the classification error rate significantly. Without using explicit gender information, CBPCA adaptation can achieve better performance than gender-dependent modeling. CBPCA also proved to be effective in unsupervised adaptation. These results suggest that it is possible to use the proposed model in a real speech- recognition system. This will be the topic of Chapter 6. g

Chapter 6

CBPCA in HMM speech recognition system

The classification experiments in Chapter 5 demonstrate that CBPCA can reliably extract speaker-dependent characteristics. In this chapter we investigate an application of this method, in an HMM recognition system. Section 6.1 briefly introduces some background on HMM recognizers. Section 6.2 describes the baseline HMM recognition system result. Section 6.3 describes the process and results of using CBPCA for speaker clustering. Section 6.4 compares CBPCA and MLLR-based adaptation results. Finally Section 6.6 investigates the possibility of using CBPCA in adaptive training of an HMM recognizer.

6.1 Acoustic Modeling using Hidden Markov Models

In this section, we briefly introduce the well-known and widely used Hidden Markov Model approach, and formulate it in mathematical terms appropriate for this thesis.

In current speech technology, the speech signal is converted to a parametric representation before it is used in recognition. This process produces a sequence of parameter vectors representing the speech waveform. The parameterization can be performed for a fixed short time window (e.g. 20 milliseconds), or for a much bigger section of speech (e.g. segments of phonemes, words, etc). In this chapter, we only discuss the most common frame-based representation.

If an utterance of speech is represented by a series of T parameterized speech frames (referred to as observation vectors O),

$$O = o_1, o_2, ..., o_T \tag{6.1}$$

the object of the pattern matching stage is to determine the most likely sequence of words which produce this observation sequence.

Assuming the observation vectors are produced by a Markov process, the pattern matching can be achieved by using Hidden Markov Models (HMMs). The underlying assumption of an HMM model is that the speech signal can be characterized as a parametric random process, with successive observations conditionally independent. Given this assumption, the parameters of the stochastic process can be estimated in a precise and well-defined manner, as described below.

6.1.1 Hidden Markov Models representations

Markov Model and HMM

A Markov model (Figure 6.1) is a finite state machine which makes a state transition once every time step. At each time t a state j is entered, and an observation vector o_t is generated according to a probability density function associated with the state $b_j(o_t)$. Furthermore, the transition from state i to state j is also probabilistic and is governed by the probability a_{ij}. Hence, the likelihood of generating observation O using any sequence of states can be computed.

A hidden Markov model (HMM)[22] has the following two components:

1. A first order Markov chain that models the sequentially evolving temporal nature of the process.

2. A finite set of output probability distributions that model the statistical behavior of the steady states of the process. Each state of the Markov chain is associated with an output probability distribution, which can be characterized by either a discrete probability function or a continuous probability density function.

When used in speech recognition, the state process $S = (s_1, s_2, ..., s_T)$ is not observed,

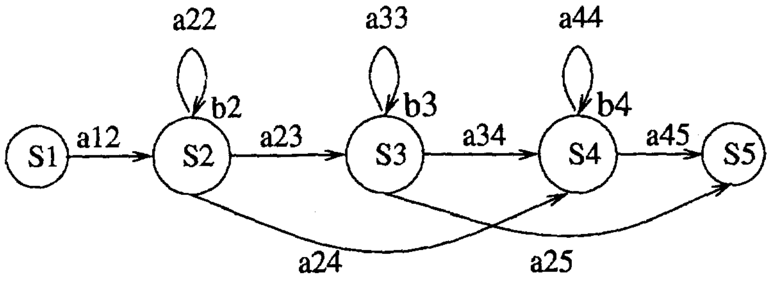


Figure 6.1: Example configuration of a HMM structure with non-emitting beginning and exit state.

hence it is referred to as a *hidden* sequence. Therefore, the models are referred to as *Hidden Markov Models.*

Hidden Markov Model Parameters

As illustrated in Figure 6.1, an HMM consists of a set of N states $s_1, s_2, ..., s_N$ with the following properties:

- The number of states in the model, denoted by N, where the states are labeled $s = s_1, s_2, ..., s_N$. State i at time t is denoted as $s_t = i$.

- The left most (entry) state (s_1 in the figure) of the model and the right most (exit) state (s_5 in the figure) of the model are non-emitting states. The non-emitting states allow simple construction of model sequences and incorporation of language models.

- States s_2 through s_4 are emitting states. Each emitting state s_j has an associated output probability density function $b_j(o_t)$.

- The initial state distribution is defined as

$$\pi = \{\pi_i | \pi_i = P(s_1 = i), 1 \leq i \leq N\} \tag{6.2}$$

  where $P(s_1 = i)$ represents the initial probability of state i at time step $t = 1$.

- A transition matrix A defines allowable transitions between states. a_{ij} is the probability of moving from state i at time step t to state j at time step $t + 1$.

$$a_{ij} = P(s_{t+1} = j | s_t = i) \tag{6.3}$$

with the following constraints:

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{6.4}$$

- The output probability distribution associated with each state is denoted as:

$$B = b_i(o_t) = P(o_t | s_t = i), 1 \leq i \leq N \tag{6.5}$$

The output probability density functions $b_j(o_t)$ can in theory be any probability distribution. However, in practice, they are usually assumed to be mixtures of Gaussian densities.

$$b_j(o_t) = \sum_{k=1}^{K} c_{jk} b_{jk}(o_t) \tag{6.6}$$

where K is the number of component densities, c_{jk} is the mixture weight, where $0 \leq c_{jk} \leq 1$ and

$$\sum_{k=1}^{K} c_{jk} = 1$$

and b_{jk} is a multivariate Gaussian density function.

$$b_{jk}(o_t) = \mathcal{N}(o_t; \mu_{jk}, \Sigma_{jk}) = \frac{1}{(2\pi)^{n/2} \Sigma_{jk}^{1/2}} e^{-\frac{1}{2}(o_t - \mu_{jk})' \Sigma_{jk}^{-1} (o_t - \mu_{jk})} \tag{6.7}$$

Here d is the dimension of the observation vector o_t and μ_{jk} and Σ_{jk} are the mean and covariance matrix respectively for the kth Gaussian component of the jth state. The covariance matrix could be a full $n \times n$ matrix or a diagonal covariance matrix.

An HMM can be represented by using the compact notation $\lambda = (\pi, A, B)$ as defined in Equations 6.2, 6.3 and 6.7. Specification of an HMM involves in the choices of number of states (N), the number of mixtures components per state (K), and the estimation of its parameters λ given the observation data.

Assumptions in HMM modeling

The assumptions made in the generally used HMM speech modeling are the following:

1. The distribution of observation vectors (the output probabilities) is assumed to depend only on the present state.
2. The speech signal is assumed to be stationary over a single frame of speech.
3. The observation vector distributions are adequately modeled by the parameters of the HMM.

The independence assumption is obviously not true, since there are certainly correlations between adjacent frames. However, it is computationally expensive to integrate the correlations between frames and difficult to train a model which accounts for these adequately.

The second assumption is also not strictly true. However, it is generally accepted that spectral analysis based on this assumption is adequate for the purposes of speech recognition.

The third assumption might also not be true since we do not know the underlying model. Fortunately, an arbitrary density function can be approximated to any degree of accuracy by a sufficiently large number of component densities.

6.1.2 HMM parameter estimation

The objective of a speech recognizer is to find the word sequence which is most likely to produce the given set of observation vectors (O).

If an observation sequence is known to represent a single word (w_i) from a limited set of possible words (vocabulary V), the task is to compute the most probable word given the observation data O.

$$\max_i p(w_i|O) \tag{6.8}$$

Using the Bayes rule we can rewrite the *a posterior* probability in terms of the data likelihood as follows:

$$p(w_i|O) = \frac{p(O|w_i)p(w_i)}{p(O)} \tag{6.9}$$

This task can then be reduced to determining $p(O|w_i)$ assuming that $p(w_i)$, the probability of occurance of word w_i, can be computed through a language model using *a priori* knowledge, and that $p(O)$ are common to all words under consideration, and therefore do not affect the choice of the w_i.

Using hidden Markov models, we can generate a class-conditional observation probability $p(O|w_i)$, given that an acoustic model λ representing each w_i exists. This can be achieved by computing the likelihood $\mathcal{L}(O|\lambda)$ of taking any path through the acoustic model λ and producing the given sequence of observation vectors. $\mathcal{L}(O|\lambda)$ takes into account all possible state sequences,

$$\mathcal{L}(O|\lambda) = \sum_{s \in \mathcal{S}} \pi_{s_1} b_{s_1}(o_1) \prod_{t=2}^{T} a_{s_{t-1}} a_{s_t} b_{s_t}(o_t) \tag{6.10}$$

Computing $\mathcal{L}(O|\lambda)$ defines the parameter estimation problem in HMM modeling. The most common parameter estimation method is *maximum likelihood* estimation. In maximum likelihood (ML) estimation, the parameter vector λ of a model is derived as a function of the observation vector O by maximizing the likelihood function $\mathcal{L}(O;\lambda)$, which is defined as the probability that the model λ has generated the observation sequence O. Therefore:

$$\lambda_{ML} = \arg\max_{\lambda} \mathcal{L}(O;\lambda) \tag{6.11}$$

In HMM training, the parameter estimation is achieved using the *expectation-maximization* (EM) algorithm to provide ML estimates in an iterative fashion. Detailed derivations are given in many papers such as [2, 33, 37, 43]. Here we just highlight the results:

- maximization of the auxiliary function $\mathcal{Q}(\lambda, \tilde{\lambda})$ will result in maximization of the likelihood $\mathcal{L}(O;\lambda)$, where

$$\mathcal{Q}(\lambda, \tilde{\lambda}) = \mathcal{E}\{log\mathcal{L}(O, S, K; \tilde{\lambda})|O;\lambda\} \tag{6.12}$$

  where S is the Markov state sequence corresponding to the observation sequence O.

When only the parameters μ and Σ in the HMM models are optimized, the auxiliary function becomes:

$$\mathcal{Q}(\lambda, \tilde{\mu}, \tilde{\Sigma}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} \gamma_{ik}(t) log \mathcal{N}(o_t; \tilde{\mu}_{ik}, \tilde{\Sigma}_{ik}) \tag{6.13}$$

where $\gamma_{ik}(t)$ is the state occupation probability of occupying state s mixture k at time step t. In this equation N represents the total number of states in the system. $\tilde{\mu}$ and $\tilde{\Sigma}$ are the estimated μ and Σ respectively.

Differentiating Equation 6.13 with respect to $\tilde{\mu}$ and $\tilde{\Sigma}$ and setting to 0 to find the maximum of the likelihood function yields the following iterative solution:

$$\tilde{\mu}_{ik} = \frac{\sum_{t=1}^{T} \gamma_{ik}(t) o_t}{\sum_{t=1}^{T} \gamma_{ik}} \tag{6.14}$$

and

$$\tilde{\Sigma}_{ik} = \frac{\sum_{t=1}^{T} \gamma_{ik}(t)(o_t - \tilde{\mu}_{ik})(o_t - \tilde{\mu}_{ik})^T}{\sum_{t=1}^{T} \gamma_{ik}} \tag{6.15}$$

6.2 Baseline system

For all the experiments described in this chapter, the baseline recognition system is an HMM alphabet-digit system. The major parameters are shown in Table 6.1. There are 650 speakers in the training data; each speaker speaks approximately 30 sentences of alphabet letters and digits. There are about 6 words in each sentence. The development set has 77 speakers and the final test set has 82 speakers. The system has $N = 599$ tied states with $K = 10$ mixtures per state. Each tied-state triphone model is represented using the standard 3-state left-to-right model configuration.

All results in this chapter are obtained from the final test set using the models which performed best on the development set. The baseline system's word error rate is 13.2%.

6.3 Speaker clustering

As illustrated in Equations 3.2 and 3.3, the speaker-dependent subspace of speaker space can be extracted by computing each speaker's projections onto the principal components (a_i^s). Since similar projections (a_i^s) translate into speakers having similar voices, we can generate speaker clusters using these projections. Since each speaker cluster contains

Table 6.1: Baseline system parameters and performance.

task	alphabet a and digits
# speakers	train: 650
# speakers	dev: 77
# speakers	test: 82
# states	599
# mixtures per state	10
word error rate	13.22%

similar speakers, the model trained on data from one cluster should perform better than the baseline system which was trained on data from all speakers.

In this experiment, speaker clusters are generated using five speaker-dependent attributes. Three speaker clusters are generated using vector quantization. The baseline system is further trained on each cluster's data. During evaluation, the resulting "cluster-dependent" recognizers are used to compute likelihoods for each test sentence. The model which produces the highest likelihood is used to generate the answer. Gender-dependent result is also computed for comparison. The results (word error rate) for 10 mixture models are presented in Table 6.2.

Table 6.2: Comparison between speaker clustering and gender-dependent system.

	baseline	gender-dependent	clustering
word error rate	13.22%	12.07%	11.69%
% error reduction		8.7%	11.6%

Speaker clustering resulted in a 11.6% reduction in word error. This result is similar to that of gender-dependent training in large-vocabulary systems, but achieved without any knowledge of gender in the training data. On this particular task, using 3 speaker clusters results in better performance than gender-dependent models.

The result shows that the speaker clusters generated this way indeed reflect speaker similarities. This also provides further evidence that the constructed speaker space describes speaker variation. The advantage of using the proposed method is that it needs very limited data to estimate the low-dimensional speaker-dependent parameters on which to do clustering. These clusters may also be used in speaker segmentation for tasks such as the NIST "broadcast news" task in which speech from many different speakers is present. Finally, when a very large training corpus is available, additional improvements may be achievable with more than two speaker clusters. Multiple speaker clusters may more accurately describe the natural distribution of speaker groups and the proposed method can discover these groups.

6.4 Speaker adaptation for HMM using CBPCA

In this section we study the behavior of adaptation using CBPCA and compare it with that of traditional MLLR.

6.4.1 Speaker adaptation results

Unsupervised adaptation results for both CBPCA and MLLR are presented first. The alignment from the corresponding baseline system for all the test sentences are used to estimate the adaptation parameters.

The recognition results (word error rate) using adaptation with CBPCA are presented in Table 6.3. In the experiments, 5 principal components are used hence 5 adaptation parameters $(\alpha_i^r, i = 1..5)$ are estimated.

Table 6.3: Adaptation results using CBPCA.

# mixture	1	2	4	6	8	10
baseline(%)	23.67	18.01	15.56	14.21	13.40	13.22
adaptation(%)	22.37	17.39	14.93	13.62	12.71	12.53
% error reduction	5.5	3.4	4.0	4.2	5.1	5.2

Results in Table 6.3 show that CBPCA can perform adaptation with a relatively small number of adaptation parameters. These results also show that CBPCA can result in consistent error reduction for different mixture models.

The comparative results with a class-dependent offset MLLR are presented in Table 6.4. In these comparisons, all 30 sentences are used as adaptation data.

Table 6.4 shows that MLLR and CBPCA adaptation perform similarly. MLLR adaptation uses many more adaptation parameters than CBPCA (1014 versus 5), and may

Table 6.4: Comparison between MLLR and CBPCA adaptation results.

# mixture	1	2	4	6	8	10
baseline(%)	23.67	18.01	15.56	14.21	13.40	13.22
MLLR offset(%)	22.74	17.75	15.38	14.07	12.99	12.74
CBPCA (%)	22.37	17.39	14.93	13.62	12.71	12.53

have a disadvantage when only a small amount of adaptation data are presented as we demonstrate in the following section.

6.4.2 Results with limited adaptation data

Since we only need to estimate a few adaptation parameters for CBPCA, we can adapt successfully when only limited adaptation data is available. The experiments in this section investigate the behavior of adaptation using CBPCA and MLLR in this case. In the following experiments, we use a small part of the data in the development set as adaptation data and test on 20 previously unseen sentences for each speaker.

In the experiments, CBPCA adaptation uses five adaptation parameters and the class-dependent offset requires $39 \times 26 = 1014$ parameters. We adapted an HMM with K (K $=1,2,4,6,8,10$) mixture components.

The results are presented in Figure 6.2 and Figure 6.3.

Figure 6.2 and Figure 6.3 show that, when less data is used for adaptation, MLLR's performance declines significantly, while CBPCA-based adaptation maintains similar performance gain. This result strongly supports our conclusion in Section 3.1 that we can use CBPCA successfully in situations where only limited adaptation data are available.

6.5 Normalized training for HMM

In this section, we describe our attempt to use the proposed model in normalized training of an HMM system.

Since the proposed CBPCA normalization model can reduce the variance across different speakers, training recognizers using the normalized features can concentrate the models on variances other than the speaker variations. Alternatively, we can also shift

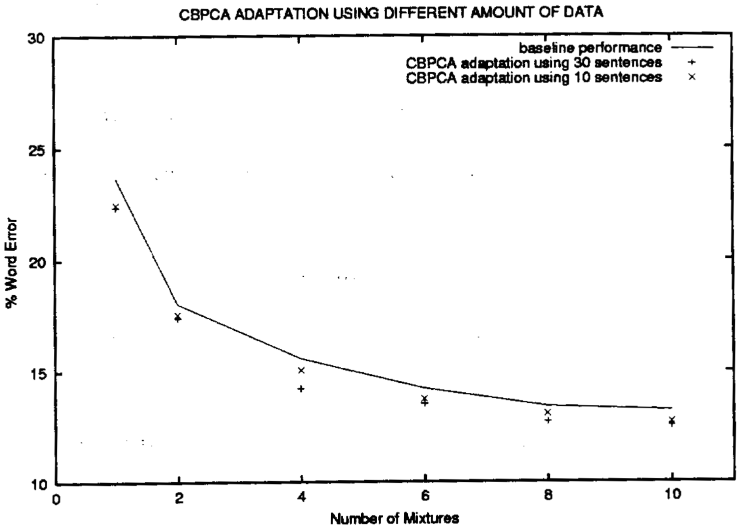


Figure 6.2: Adaptation results using different amounts of adaptation data using CBPCA.

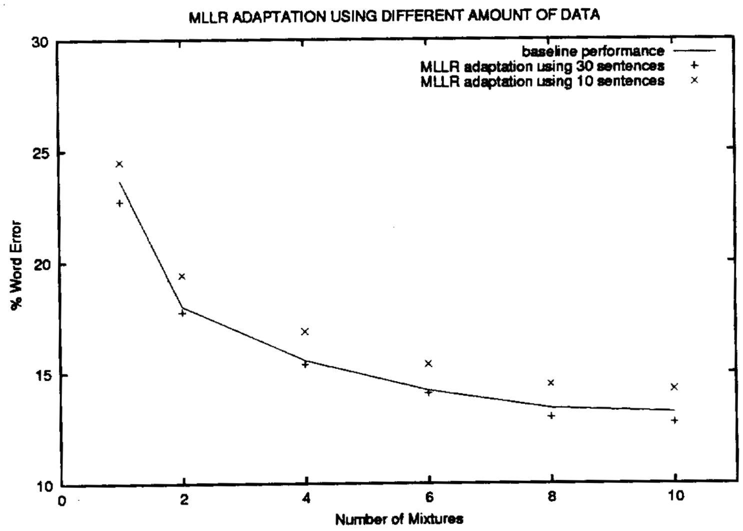


Figure 6.3: Adaptation results using different amounts of adaptation data using MLLR.

the model for each speaker during training, which will result in a more compact model (Section 6.6).

In this experiment, we normalize the features of each speaker in the training set using Equation 3.2. The normalized data are then used to train a new HMM system. This is a form of speaker adaptive training (SAT) in which the features from each speaker are normalized using the method presented in this research. The results (word error rate) are shown in Table 6.5: (These results were obtained with a slightly different feature set compared to that used previously, and also from a different test set, word error rates are therefore not directly comparable.)

Table 6.5: Recognition results from normalized training on alpha-digits.

mixture	baseline	normalized training (unsupervised)
4	13.71%	12.67%
10	11.38%	10.86%

There are several interesting observations in these results:

- Normalized training results in models that have smaller variance with the same complexity, as expected.
- Models trained with normalized features achieve proportionally better results at lower complexities. This can be explained by the lower model variance resulting from normalization. This may be an important implementational advantage for real-world systems, since fewer resources may be needed due to the use of CBPCA.
- In the most complex configuration, the system performed only slightly better than the baseline system. This might be attributed to the hard decision of class boundaries when aligning classes during the feature normalization stage.

6.6 Speaker adaptive training

In this section, we explore speaker adaptive training (SAT) using CBPCA, in which we adapt the model during the training process.

In Section 2.2, the general idea of speaker-adaptive training was introduced. In practice, this method normally is used by combining the estimation of the adaptation parameters with the embedded training algorithm in a unified parameter-estimation process. Instead of moving the features as in Section 6.5, the models are adjusted to fit each training speaker before training on that specific speaker; then the models are adjusted back using the inverse transformation. The advantage of this method is that there is no need to make the hard boundary decisions when aligning the classes to the data. During the embedded training process, each frame of training data will automatically be associated with the likelihood of it belonging to a particular model, and the corresponding parameters are adjusted accordingly.

The most important difference between adaptive training based on the ML criterion and CBPCA is that by using CBPCA, the training process incorporates the information of correlation in the acoustic models which the ML- based SAT normally does not capture.

6.6.1 SAT training based on Maximum Likelihood

Speaker adaptive training [2] is based on the assumption that speech features contain two types of information. One type is the speaker-independent information which reflects phonetic variations among all speakers. The other component of speech is a speaker-dependent variation which reflects each speaker's characteristics. This component can be represented as the transformation (or a filtering) of the speaker-independent parameters through a transformation G.

Assume a set of acoustic models λ that represent phonetic phenomena of speech that are invariant to speaker-dependent variations. The effect of each training speaker on these acoustic models is considered as a transformation G that describes the particular speaker. Therefore, the acoustic models would best match the speaker-specific data if the data can be represented as the transformation of the speaker-independent model with the speaker-specific transformation G.

In the SAT process, training data are collected from several different speakers. To estimate the model parameters that best fit the underlying speaker-independent acoustic models, the derivation of the acoustic models involves simultaneously estimating the parameters λ, which represent the phonetically relevant speech variations, as well as the speaker-dependent transformation G. Therefore the speaker-independent training problem is transformed into the following: given the set of transcribed observations $O = O^1, O^2, ..., O^R$ that are collected from R training speakers, estimate the speaker transformations $G = G^1, G^2, ..., G^R$ for each speaker, as well as the speaker-invariant acoustic model λ which maximizes the likelihood of the training data O. The maximization formulation is:

$$\begin{aligned}(\tilde{\lambda}, \tilde{G}) &= \arg\max_{\lambda, G} \mathcal{L}(O; \lambda, G^1, G^2, ..., G^R) \\ &= \arg\max_{\lambda, G} \prod_{r=1}^{R} \mathcal{L}(O^r; \lambda, G^r) \qquad (6.16)\end{aligned}$$

In this formulation, each observation $o_t^{(r)}$ provides evidence of the speaker-invariant observations via the transformation G^r. If the inverse transformation G^{-1} for each speaker exists, it is possible to normalize each speaker's training data with respect to the speaker's characteristics. It is obvious that this SAT formulation can be used in a feature-based approach, by training models on the normalized features. Here we will concentrate on the model-space approach, because it provides more flexibility and allows for the development of an EM-based algorithm for the maximum-likelihood parameter optimization.

6.6.2 SAT formulation

In conventional SAT the transformation is often modeled as an additive bias:

$$\hat{\mu}_k^r = \mu_k + \beta^r \qquad (6.17)$$

so that the speaker-independent Gaussian mean μ_k is transformed to best represent the speaker-dependent mean $\hat{\mu}_k^r$ of the r'th speaker. The additive bias β^r depends on the characteristics of the speaker and they can be shared among all Gaussians of the system or be specific for a cluster of Gaussians, states or phones.

The objective of speaker adaptive training is to estimate the parameters of the Gaussian densities and the parameters in the linear transformation for all training speakers according to the transformation (Equation 6.17) so as to maximize the joint likelihood of the training data according to Equation 6.16, in which $G^r = \beta^r$. The maximum likelihood estimates can be computed using the same Baum-Welch algorithm. Following Equation 6.13, we have:

$$\begin{aligned} Q_N(\lambda, \tilde{\mu}_{ik}, \tilde{\Sigma}_{ik}, \tilde{\beta}^r) &= \sum_{r=1}^{R}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T_r} \gamma^r_{ik}(t) log\mathcal{N}(o^r_t; \tilde{\mu}_{ik} + \tilde{\beta}^r, \tilde{\Sigma}_{ik}) \\ &= \sum_{r=1}^{R}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T_r} \gamma^r_{ik}(t) dlog(2\pi) + log|\tilde{\Sigma}_{ik}| + \\ &\quad (o^r_t - \tilde{\mu}_{ik} - \tilde{\beta}^r)^T \tilde{\Sigma}^{-1}_{ik} (o^r_t - \tilde{\mu}_{ik} - \tilde{\beta}^r) \end{aligned} \tag{6.18}$$

d is the dimension of the feature vector o^r_t.

The effects of different speakers in the training data are explicit in the summation over R. Maximization of Q_N will lead to the estimation of the parameters for the speaker-independent model and the speaker-dependent parameters in β^r for each speaker so that the overall likelihood for all the training data is maximized. These estimates are derived by differentiating Equation 6.18 with respect to the parameters to be optimized and setting to 0 to find the maximum. The resulting optimization solution can be realized in the following steps:

The SAT process starts from the speaker-independent model which is represented by parameters μ_{ik} and Σ_{ik}.

For each speaker r, we estimate the parameter β^r using:

$$\tilde{\beta}^r = {\sum_{i=1}^{N}\sum_{k=1}^{K} \gamma^r_{ik}\Sigma^{-1}_{ik}}^{-1} \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma^r_{ik}\Sigma^{-1}_{ik}(o^r_{ik} - \mu_{ik}) \tag{6.19}$$

Then, using the estimated β^r we compute the model parameters μ and Σ:

$$\tilde{\mu}_{ik} = \frac{\sum_{r=1}^{R} \gamma^r_{ik}(o^r_{ik} - \tilde{\beta}^r)}{\sum_{r=1}^{R} \gamma^r_{ik}} \tag{6.20}$$

$$\tilde{\Sigma}_{ik} = \frac{\sum_{r=1}^{R} \gamma^r_{ik} S^r_{ik} + (o^r_{ik} - \tilde{\beta}^r - \tilde{\mu}_{ik})(o^r_{ik} - \tilde{\beta}^r - \tilde{\mu}_{ik})^T}{\sum_{r=1}^{R} \gamma^r_{ik}} \tag{6.21}$$

6.6.3 Speaker adaptive training using CBPCA

When implementing speaker adaptive training using CBPCA, the speaker dependent and class dependent bias β^r can be estimated in the standard CBPCA process. The rest of the parameter-adaptation process is similar to that of the standard SAT (Equation 6.20 and Equation 6.21).

Here the calculation of β^r is:

$$\beta^r = \phi\alpha^r \tag{6.22}$$

where α^r is computed by solving Equation 3.8.

The advantage of using CBPCA is that the speaker-specific parameters can be calculated for the phoneme classes not presented in the adaptation data while in normal SAT, those classes are not adapted. Another potential advantage of using CBPCA-based SAT it that the resulting system would inherently incorporate the information of class correlations, and therefore might be more concise than a normally trained system using ML-based SAT.

Experimental results

The results of our experiments comparing CBPCA-based SAT with the baseline performance (word error rate) are given below. These results are also depicted in Figure 6.4.

Table 6.6: SAT with CBPCA results compared to baseline and non-SAT adaptation.

	# of mixtures					
method	1	2	4	6	8	10
baseline(%)	23.67	18.01	15.56	14.21	13.40	13.22
CBPCA SAT(%)	22.48	18.31	14.98	13.89	13.38	13.16
% error reduction	3.2	-	3.7	2.2	0.1	0.4

The results in Table 6.6 indicate that CBPCA-based speaker adaptive training results in a larger performance gain at lower complexities. This indicates that the adaptation indeed can reduce the variance among speakers; therefore the resulting models can be more precise while modeling the phonetic variances.

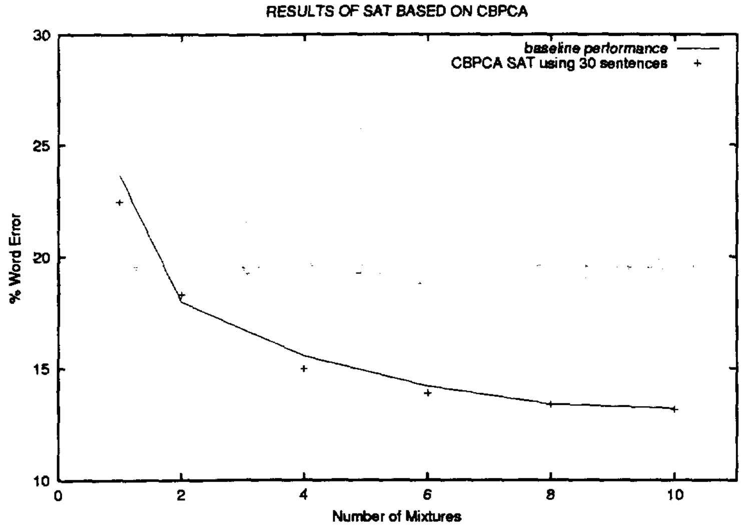


Figure 6.4: Adaptive training results using CBPCA.

The results in Table 6.6 show no obvious advantage in using CBPCA when all the data are presented at the highest complexities.

6.7 Summary

The experiments in this chapter investigated various applications of the proposed method. CBPCA proved to be able to capture speaker differences, therefore providing a new method to successfully perform speaker clustering using very few parameters. Experiments also show that CBPCA can achieve adaptation using limited adaptation data while MLLR failed to perform robustly in this situation. Normalized training using CBPCA yield more compact models than the original baseline model, and speaker adaptive training using CBPCA results in better models that are compact and have better performance.

The adaptation results show that we did not achieve significant performance gain with higher model complexities. This can be attributed to following factors:

- The higher complexity models already have the ability to model most of the variance sources present in the data, including speaker variance.
- The data contains more variations than the Hillenbrand and TIMIT data sets. This

results in decreased class correlations, which in turn affects the ability to adapt based on these correlations.

- The current adaptation model adapts models based on a chosen set of phonemes. The computed offset vector will be the same for all mixtures corresponding to a particular phoneme. This high degree of sharing might not be optimal.

Chapter 7

Summary and conclusions

7.1 Summary and discussion

In this thesis, we introduced a theoretical model of the proposed speaker normalization/adaptation method based on correlations between acoustic classes. This correlation can be used to achieve better speaker adaptation using small amounts of data. Various experiments showed the soundness and feasibility of this method. The mechanism and underlying physical meaning of this model was also explored. One important advantage of the proposed method is that it allows researchers to understand some of the properties of speaker variability and gives insight into some physical aspects of speaker differences.

Comparing to other speaker adaptation techniques, our method has the following unique characteristics which result in its advantage over current methods:

7.1.1 Understanding the underlying mechanism of adaptation

This method provides an environment that allow researcher to study and understand the underlying physical meaning and the mechanism of the adaptation process, and this suggests future extensions and improvements. Most of the previous model-based methods do not provide this kind of environment (vocal tract normalization is expensive to implement in live systems). When we understand the underlying process of adaptation, we can detect the deficiency and its cause, and subsequently design methods to reduce the deficiency.

Generic data driven adaptation techniques (such as MLLR) do not provide these kinds of insights.

The proposed method (CBPCA) describe the speaker variations within a speaker space, and the adaptation and normalization process is conducted within that space. Because of this mechanism of adaptation and normalization, we can study the underlying physical meaning of the process. This helps us to understand the effects on models and features during the adaptation and normalization process. For example, our study shows that the first principal component of the speaker space highly correlates with speaker gender which suggests that it correlates with speaker's vocal tract length. When we perform adaptation based on the first principal component, we know we are reducing gender-related variation during the process. The speaker clustering experiment in Chapter 6 also show that with CBPCA we could use very few parameters to characterize different groups of speakers. This is very different from the data-driven model-based techniques such as MLLR which only result in a data-wise optimal transformation without thoughts of the physical effects of the transformation. Those techniques generally need large amount of adaptation data to perform well as we have seen in Chapter 6.

7.1.2 Adaptation using correlations

The CBPCA method enables us to use correlations among classes to adapt the classes not presented in the adaptation data. Most of previous model-based approaches do not have this ability. In CBPCA, we model a "covariance" relationship between speakers and between classes, comparing to the "invariant" relation proposed in Cox's [7] prediction techniques. Our research allows more flexibility in the relationships between classes for different speakers.

7.1.3 Adaptation using limited data

The proposed adaptation model (CBPCA) has the ability to adapt with limited adaptation data. Data-dependent adaptation methods such as MLLR, require large amount of adaptation data for each adaptation transformation. A small amount of adaptation data often yields little adaptation (small change for the models) for Bayesian methods such as

MAP.

Since our proposed method operates within a concentrated subspace of the speaker variation, it requires estimation for only a few adaptation parameters. Therefore, it allows us to adapt efficiently using limited data. The results in Section 6.4 show strong support for this assumption. With limited adaptation data, MLLR adaptation's performance degrades while CBPCA's performance remains unchanged. This capability of CBPCA can be very useful in applications that only have very few phrases for adaptation, or when it is essential to adapt faster. For example, in a system for a large population when we need to adapt in the context of a single dialog, CBPCA adaptation will be very useful. Even if we get multiple interactions, this will allow us to store a very small amount of adaptation parameters for each speaker. This is a simpler and cheaper method which can result in improved performance.

7.2 Extension of this work

As with every new method, many issues are still unsolved and should be investigated further. During our investigations, we identified the following areas:

7.2.1 Combining CBPCA and MAP

It would be beneficial to combine the advantage of CBPCA with that of MAP. CBPCA could be used to estimate better priors for MAP adaptation. Since CBPCA could adapt the models using small amount of data, this would "accelerate" the MAP adaptation when there's only small amount of adaptation data are available.

7.2.2 Extension of this method to adapt the covariance of the models

The adaptation method studied in this thesis only addresses the adaptation of the means of models. In addition, the adaptation of the covariance of each model could also be investigated within the general CBPCA framework, i.e. in the "concentrated subspace" of the original speaker space.

The underlying assumption of the adaptation formula (Equation 3.4) is that all speaker variances can be described by Φ (and most of variance could be described in the subspace of Σ_0) and the difference between speakers can be described by a shift in the subspace. Other forms of more complicated transformation in the subspace could also be investigated. For example, a matrix transformation of the subspace projections.

7.2.3 Other physical Meanings

Possible physical meanings other than the ones discussed in this thesis could also be further investigated. This might be done with specific features like formants. Large number of speaker should be included to be able to provide quantifiable result. Many different combination of variations should be tested since the higher rank eigenvectors might represent some combinations of variations. This will lead to better understanding of the adaptation process which may lead to better adaptation method.

7.2.4 Speaker Morphing

In the speaker space we constructed, we can transform a speaker's feature into another, by computing a transformation as follows:

$$\mu^p = \mu^q + \sum_{i=1}^{CD} \phi_i \cdot \beta_i^{pq} \tag{7.1}$$

where μ^p and μ^q represent the position of speaker p and q in the speaker space, and β_i^{pq} are the projections of the difference vector between p and q onto the principal components.

Given some training data from speaker p and q, β_i^{pq} can be estimated for the transformation and then used to map speaker p onto speaker q. If the features used in this process capture the major speaker characteristics, we could transform a speaker's voice into another. The potential advantage of this approach is that it would require a small amount of data to morph between speakers.

The advantage of this method is that only limited parameter for the source and target speaker are needed, which equal to that of the number of eigenvectors used in the system. The requirement of this method is that the features used in this scheme must be highly speaker-dependent and accurately estimated. This is of major importance in text-to-speech(TTS) systems where most of the storage goes towards speaker data. By capturing the transformation from a specific speaker to another with small number of parameters, we can have many voices from a single database. In addition, we can "add" a new voice from very limited data.

Bibliography

[1] Ahadi, S. M., and Woodland, P. C. Rapid speaker adaptation using model prediction. *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing, 1995* (1995), 684–687.

[2] Anastasakos, A. *Speaker Normalization Methods for Speaker Independent Speech Recognition.* PhD thesis, Northeastern University, October 1996.

[3] Black, A., and Taylor, P. The festival speech synthesis system: system documentation. *Technical Report HCRC/TR-83*, Human Communication Research Centre, University of Edinburgh,Scotland,UK, January 1997.

[4] Brown, P. F., Lee, C. H., and Spohrer, J. C. Bayesian adaptation in speech recognition. *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing 1983* (1983), 761–764.

[5] Burnett, D. *Rapid Speaker Adaptation for Neural Network Speech Recognizers.* PhD thesis, Oregon Graduate Institute, December 1996.

[6] Cox, S. Speaker adaptation using a predictive model. *Proceedings of the European Conference on Speech Technology* (1993), 2283–2286.

[7] Cox, S. Predictive speaker adaptation in speech recognition. *Computer Speech and Language 9* (1995), 1–17.

[8] Cox, S. A speaker adaptation technique using linear regression. *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing* (1995), 700–703.

[9] Digalakis, V. V., and Neumeyer, G. Speaker adaptation using combined transformation and bayesian methods. *IEEE Transactions on Speech and Audio Processing 4* (July 1996), 294–300.

[10] Eide, E., and Gish, H. A parametric approach to vocal tract length normalization. *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing 1996* (1996), 346–348.

[11] FUKUNAGA, K. *Statistical Pattern Recognition.* Academic Press, Inc, 1990.

[12] GALES, M. The generation and use of regression class trees for MLLR adaptation. *Technical Report CUED/F-INFENG/TR263,* Cambridge University, August 1996.

[13] GAUVAIN, J. L., AND LEE, C. H. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing 2* (April 1994), 291–298.

[14] GRENIER, Y. MICHLET, L., MAURIN, J. C., AND MICHEL, H. Speaker adaptation for phoneme recognition. *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing* (1981), 1273–1275.

[15] GRENIER, Y. Speaker adaptation through canonical correlation analysis. *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing 1980* (1980), 888–891.

[16] HAEB-UMBACH, R. GELLER, D., AND NEY, H. Improvements in connected digit recognition using linear discriminant analysis and mixture densities. *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing 1981* (1993), 239–242.

[17] HERMANSKY, H., MORGAN, N., BAYYA, A., AND KOHN, P. Compensation for the effect of the communication channel in auditory-like analysis of speech(rasta-plp). *Proceedings of the 2nd European Conference on Speech Communication and Technology, Genova, European Speech Communication Association* (1991), 1367–1370.

[18] HILLENBRAND, J., GETTY, L., CLARK, M., AND WHEELER, K. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America 97* (1995), 3099–3111.

[19] HU, Z., BARNARD, E., AND P., V. Speaker normalization using correlations among classes. *Proceedings the 1998 Workshop on Speech Processing, Chinese University of Hong Kong, Hong Kong, China* (1998), II223–II228.

[20] HUANG, X. D., AND LEE, K. F. On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2* (1991), 877–880.

[21] JASCHUL, J. Speaker adaptation by a linear transformation with optimized parameters. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing* (1982), 1657–1670.

[22] JUANG, B. H., AND RABINER, L. R. The segmental k-means algorithm for estimating parameters of hidden markov models. *IEEE Transactions on Acoustic, Speech, and Signal Processing 38* (September 1990), 1639–1641.

[23] KAMM, T., ANDREOU, A. G., AND COHEN, J. Vocal tract normalization in speech recognition: Compensation for systematic speaker variability. *Proceedings of the 15th Annual Speech Symposium (Johns Hopkins University,)* (June 1995), 175–178.

[24] KIRBY, M., AND SIROVICH, L. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12* (Jan. 1990), 103–108.

[25] KUHN, R., NGUYEN, P., JUNQUA, J.-C., GOLDWASSER, L., NIEDZIELSKI, N., FINCKE, S., FIELD, K., AND CONTOLINI, M. Eigenvoices for speaker adaptation. *Proceedings International Conference on Spoken Language Processing* (1998), 1771–1774.

[26] KUMAR, B., AND MURAKAMI, H. Principal-component imagery for statistical pattern recognition correlators. *Optical Engineering 21* (1982), 43–47.

[27] LADEFOGED, P. *A Course in Phonetics.* Harcourt Brace Jovanovich College Publishers, 1975.

[28] LAMEL, L., KASSEL, R., AND SENEFF, S. Speech database development: design and analysis of the acoustic-phonetic corpus. *Proceedings DARPA Speech Recognition Workshop*, SAIC-86/1546 (February 1986), 100–109.

[29] LASRY, M., AND STERN, R. A posterior estimation of correlated jointly gaussian mean vectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6* (1984), 530–535.

[30] LEE, C. H., LIN, C. H., AND JUANG, B. H. A study on speaker adaptation of continuous density hmm parameters. *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing* (1990), 145–148.

[31] LEE, L., AND ROSE, R. Speaker normalization using efficient frequency warping procedures. *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing 1996* (1996), 353–356.

[32] LEE, L., AND ROSE, R. A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and audio Processing 6* (January 1998), 49–60.

[33] LEGGETTER, V. *Statistical Trajectory Models for Phonetic Recognition.* PhD thesis, Cambridge University, August 1995.

[34] MACON, M., CRONK, A., WOUTERS, J., AND KAIN, A. Ogireslpc: Diphone synthesis using residual-excited linear prediction. *Technical Report CSE-97-007*, Oregon Graduate Institute, Portland,OR, 1997.

[35] NEAREY, T. Static,dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America 85* (May 1989), 2088–2113.

[36] PETERSON, G., AND BARNEY, H. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America 24* (1952), 175–184.

[37] RABINER, L., AND JUANG, B. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[38] ROZZI, W. *Speaker Adaptation in Continuous Speech Recognition via Estimation of Correlated Mean Vector.* PhD thesis, Carnegie Mellon University, April 1991.

[39] SANKAR, A., AND LEE, C. H. A maximum likelihood approach to stochastic matching for robust speech recognition. *IEEE Transactions on Acoustic, Speech, and Signal Processing 4* (1996), 190–202.

[40] SCHALKWYK, J., DE VILLIERS, J., VAN VUUREN, S., AND VERMEULEN, P. Cslush: An extendible research environment. *Proceedings of the European Conference on Speech Technology* (1997), 698–701.

[41] WAKITA, H. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing 25* (April 1977), 183–192.

[42] WELLING, L., AND NEY, H. A model for efficient formant estimation. *Proceedings IEEE International Conference on Acoustic, Speech, and Signal Processing* (May 1996), 797–800.

[43] ZAVALIAGKOS, G. *Maximum A Posteriori Adaptation Techniques for Speech Recognition.* PhD thesis, Northeastern University, August 1995.

[44] ZHAO, Y. An acoustic-phonetic based adaptation technique for improving speaker independent continuous speech recognition. *IEEE Transactions on Speech and Audio Processing 2*, 3 (July 1994), 380–394.

Appendix A

Maximum *A Posteriori* Estimation(MAP) and EMAP

The Maximum a Posteriori(MAP) estimation approach has shown to be an effective method of adaptation. It's advantage , when compared to other linear transformation methods, is that it provides a way of incorporating prior information in the training process and its asymptotic behavior. As the amount of adaptation data increases, the estimation process weighs the prior information less, thus converging to the speaker dependent training.

One major disadvantage of the MAP adaptation approach is that the method only adapts the models for which samples have been observed in the adaptation data. These models may represent only a small fraction of all the models in the recognition system. Extended MAP (EMAP) is one of the schemes that address this problem by incorporating correlations among models, so that observations that contribute to the estimation of one model will also contribute to the estimation of all other models. This appendix reviews the derivation of MAP and EMAP.

A.1 Maximum *A Posteriori* Estimation(MAP)

Within the HMM framework, where the data is modeled as a random variable with Normal distribution, the model parameters can be adapted using maximum *a-posterior* training (MAP).

Considering one Gaussian model for one class, through Bayes law, the *a posteriori* probability of the adapted model mean μ' can be expressed as:

$$p(\mu'|\chi) = \frac{p(\chi|\mu')p(\mu')}{p(\chi)} \tag{A.1}$$

where χ is the set of observation for this class from a single speaker. The definition of $p(\chi|\mu')$ and $p(\mu')$ can be written as:

$$p(\chi|\mu') = c1 \cdot exp[-1/2(\chi - \mu')^T \Sigma^{-1} (\chi - \mu')] \tag{A.2}$$

and

$$p(\mu') = c2 \cdot exp[-1/2(\mu' - \mu_0)^T \Sigma_0^{-1} (\mu' - \mu_0)] \tag{A.3}$$

where $c1$ and $c2$ are independent from μ'.

In order to compute the maximization of the *a-posterior* probability for the adapted model, taking the derivative with respect to μ' of the natural logarithm of $p(\mu'|\chi)$ we have:

$$\nabla_{\mu'} log[p(\mu'|\chi)] = \nabla_{\mu'} log[p(\chi|\mu')] + \nabla_{\mu'} log[p(\mu')] - \nabla_{\mu'} log[p(\chi)] \tag{A.4}$$

Given the initial model mean μ_0 with covariance Σ_0, and the speaker dependent adaptation data mean a and covariance Σ, the first term on the right side of Equation A.4 can be written as:

$$\nabla_{\mu'} log[p(\chi|\mu')] = \sum_{i=1}^{N} \Sigma^{-1}(x_i - \mu_0) = \Sigma^{-1} N(a - \mu_0) \tag{A.5}$$

In Equation A.5, N is the number of examples for the class in the adaptation data.

The second term on the right of Equation A.4 can be written as:

$$\nabla_{\mu'} log[p(\mu')] = -1/2 \nabla_{\mu'} (\mu' - \mu_0)^T \Sigma_0^{-1} (\mu' - \mu_0) = -\Sigma_0^{-1}(\mu' - \mu_0) \tag{A.6}$$

The third term of right side of Equation A.4 is zero since $p(\chi)$ is independent of μ'.

Setting $\nabla_{\mu'} log[p(\mu'|\chi)]$ equal to zero, we get:

$$\nabla_{\mu'} log[p(\mu'|\chi)] = \Sigma^{-1} N(a - \mu_0) - \Sigma_0^{-1}(\mu' - \mu_0) = 0 \tag{A.7}$$

Solving for μ' results in the MAP equation for the updated model mean μ' [11]:

$$\mu' = \frac{\Sigma}{N}[\Sigma_0 + \frac{\Sigma}{N}]^{-1}\mu_0 + \Sigma_0[\Sigma_0 + \frac{\Sigma}{N}]^{-1}a \tag{A.8}$$

With large N we can see that this corresponds to computing the model mean directly from the adaptation data:

$$\lim_{N\to\infty} u' = \frac{1}{N}\Sigma_{i=1}^{N} x_i = a \tag{A.9}$$

In general, however, MAP corresponds to a linear interpolation between the speaker independent model mean and the speaker dependent model mean.

A.2 Extended MAP(EMAP)

In normal MAP, data from one class can not be used to update any other class. In order to solve this problem, an extension of MAP (called EMAP) was introduced by Lasry and Stern [29] and further investigated by Rozzi [38]. The aim of the EMAP algorithm is to use the correlations between classes to estimate the transformation for unseen classes from the available adaptation data.

EMAP differs from MAP in that EMAP updates a concatenated mean of all classes. This concatenated mean vector can be represented as

$$\mu^r = [\mu_1^r, \mu_2^r, ..., \mu_C^r] \tag{A.10}$$

where C represents the total number of classes. Each mean vector has D dimensions. Therefore the concatenated mean vector for speaker r, μ^r is a CD dimensional vector that is assumed to be normally distributed around *a priori* mean μ_0 with covariance matrix Σ_0. Each example of class $j \in 1..C$ across all speakers is assumed to be a random vector of a normal distribution with a mean μ_j and covariance Σ_j.

In this system, correlations amongst classes are modeled by Σ_0, and the matrices Σ_j model the within class variances. We define Σ as a CD by CD block diagonal matrix with Σ_j as the jth block. Σ is block diagonal due to the assumption of independent observations., e.g. individual realizations of phonemes are not correlated.

Similar to Equation A.8 we can now compute the adapted mean vector as follows:

$$\mu' = \Sigma(\Sigma + L\Sigma_0)^{-1}\mu_0 + \Sigma_0(\Sigma + L\Sigma_0)^{-1}La, \tag{A.11}$$

where L is a diagonal matrix with the number of observations per class along the diagonal. The vector a denotes the concatenated mean vector computed from the adaptation data only.

Therefore, through the mean cross covariance matrix Σ_0, EMAP estimation is able to update all classes given any example from any class. We see in chapter 3 that this mean cross covariance matrix can be interpreted as a covariance matrix for a speaker space. Because of the dimensionality it is very difficult to estimate Σ_0 accurately. To estimate this matrix accurately would require full coverage of each class (phoneme) spoken by every speaker in the training data. Furthermore the matrix inversion $(\Sigma + L\Sigma_0)^{-1}$ is also prohibitive. Rozzi [38] extended the basic EMAP algorithm so that the computation is more efficient. These extensions however, still do not solve the problem of accurately estimating Σ_0 from a reasonable amount of data.

Biographical Note

I graduated from Tsinghua University in Beijing, China in 1990. After that, I worked as a research engineer in Tsinghua Mechanical-Electronical Technology Co. on the design of the first full-automatic ultrasonic fault-detector. Later, I worked as a software engineer in the Four-Dimension Co. before I came to OGI to pursue a Ph.D. degree.

I worked on various areas of speech recognition technology during the years in CSLU. The experience includes:

- Confidence measure and reclassification in speech recognition
- Dynamic modeling in the transition area in acoustic modeling
- Building a syllable-based recognizer
- Dynamic modeling in a syllable-based recognizer
- Modeling dynamics by using perceptual effective formants
- Speaker adaptation/normalization in continuous speech recognition

The area that I'm interested in are speech recognition, speech synthesis, spoken language systems.

Publications:

- Z. Hu, E. Barnard and R. A. Cole, "Transition-Based Feature Extraction Within Frame-Based Recognition," the Fourth European Conference on Speech Communication and Technology, 1555-1558, Sept.18-21, Madrid, Spain, 1995.
- Z. Hu, J.Schalkwyk, E.Barnard and R.Cole, "Speech Recognition Using Syllable-Like Units", International Conference on Spoken Language Processing(ICSLP), 1117-1120, Philadelphia (USA), 1996.

- Z. Hu and E. Barnard, “Smoothness Analysis for Trajectory Features,” International Conference on Acoustics, Speech and Signal Processing(ICASSP), 979-982, Munich, Germany, 1997.

- Z. Hu and E. Barnard, “Efficient Estimation of Perceptual features for speech recognition”, Fifth European Conference on Speech Communication and Technology, Athens, Greece, 1997.

- Z. Hu , E. Barnard and P. Vermeulen, “Speaker Normalization using Correlations Among Classes”, Proceedings of the 1998 Workshop on Speech Processing, II223-II228, CUHK,Hong Kong, September 3-4, 1998.