ROBUST SPEECH RECOGNITION BY TOPOLOGY PRESERVING ADAPTATION

by

M. Kemal Sönmez

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1998

Advisory Committee:

Professor John S. Baras, Chairman/Advisor
Professor Prakash Narayan
Professor Benjamin Kedem
Associate Professor Adrian Papamarcou
Associate Professor Shihab Shamma

ABSTRACT

Title of Dissertation: ROBUST SPEECH RECOGNITION BY TOPOLOGY PRESERVING ADAPTATION

M. Kemal Sönmez, Doctor of Philosophy, 1998

Dissertation directed by: Professor John S. Baras
Electrical Engineering Department

The performance degradation as a result of acoustical environment mismatch remains an important practical problem in speech recognition. The problem carries a greater significance in applications over telecommunication channels, especially with the wider use of personal communications systems such as cellular phones which invariably present challenging acoustical conditions. Such conditions are difficult to model analytically for a general speech representation, and most existing data-driven models require simultaneous ("stereo") recordings of training and testing environments, impractical to collect in most cases of interest. In this dissertation, we propose an invariance principle for non-parametric speech representations in acoustical environments. We stipulate that the topology of the codevectors in a vector quantization (VQ) codebook as defined in terms of class posterior distributions will be preserved in a certain information-theoretic

sense, and make this invariance principle our basis in deriving normalization algorithms that correct for the acoustical mismatch between environments. We develop topology preserving algorithms in two frameworks, constrained distortion minimization (VQ with a topology preservation constraint) and information geometry (alternating minimization with a topology preservation constraint) and show their equivalence. Finally, we report results on the Wall Street Journal data, the Spoken Speed Dial corpus and the TI Cellular Corpus. The algorithm is shown to improve performance significantly in all three tasks, most notably in the more difficult problem of cellular hands free microphone speech where the technique decreases the word error for continuous ten digit recognition from 23.8% to 13.6% and the speaker dependent voice calling sentence error from 16.5% to 10.6%.

Dedication

Anneme

Table of Contents

List of Tables

List of Figures

ROBUST SPEECH RECOGNITION BY
TOPOLOGY PRESERVING ADAPTATION

M. Kemal Sönmez

December 3, 1998

This comment page is not part of the dissertation.

Typeset by LaTeX using the dissertation class by Pablo A. Straub, University of Maryland.

Chapter 1

Introduction

1.1 Overview

Automatic Speech Recognition (ASR) technology has established itself as a viable research discipline in the last twenty years, and with the relatively more recent developments in Hidden Markov Model (HMM) based technology, is coming to the forefront to take its place in the information technology revolution. Systems with the capability to recognize speaker independent continuous speech of up to 60,000 words exist. However, these state-of-the-art recognizers exhibit a marked sensitivity to mismatches in training and testing acoustical environments. This sensitivity degrades performance in many speech recognition tasks in which a high-quality microphone and wide-band, high sampling-rate recording are not available. This includes the majority of present and potential applications of speech recognition such as information retrieval, command and digit recognition, voice dialing over telephone, and is currently one of the most important practical problems in speech recognition. It carries a greater significance in applications over telecommunication channels, especially with the wider use

of personal communications systems such as cellular phones which invariably present challenging acoustical conditions.

This phenomenon has stimulated growing interest in robust speech recognition making it an active research field in the last decade. There have been many recent significant advances towards rendering the speech recognition systems less vulnerable to noise and distortion in the general direction of uniformly error-free recognition over a wide range of environments (see e.g. [20], [16]). The studied approaches range from microphone arrays [12, 37] and auditory-based representations of speech features [15, 35] to approaches based on filtering of features [3, 18]. Some of the most successful approaches to environmental compensation have been based on modifying the feature vectors that form the core of the speech representation for the speech recognition system. These modifications may be based on empirical comparisons of high-quality and degraded speech data or on analytical models of the degradation [2].

In this dissertation, we propose an invariance principle for non-parametric speech representations in acoustical environments. We stipulate that the topology of the codevectors in a vector quantization (VQ) codebook as defined in terms of class posterior distributions will be preserved in a certain information-theoretic sense. We make this invariance principle our basis in deriving normalization algorithms that correct for the acoustical mismatch between environments.

We develop algorithms in two frameworks, constrained distortion minimization (VQ with a topology preservation constraint) and information geometry (alternating minimization with a topology preservation constraint). The derived stochastic gradient algorithms turn out to be the same for both approaches,

although the two different frameworks shed light on different aspects of the algorithm's properties.

We report robust speech recognition results for the Topology Preserving Adaptive VQ (TPAVQ) algorithm in three different tasks that include cepstrum reconstruction, speech and speaker recognition with data from a variety of environments such as different microphones, telephone lines, and cellular telephones. The algorithm is shown to be affective in all the tasks.

1.2 Goals

0ur goals in the dissertation can be summarized as follows:

- Development of a non-parametric, feature independent formulation of the acoustical environment robustness problem.
- Development of a distortion model for a generic feature space in which the invariance principle of topology preservation can explicitly be enforced.
- Derivation of normalization algorithms for the developed distortion model without imposing impractical requirements such as availability of stereo data.
- Development of analytical tools for the study of the algorithm to demonstrate convergence and to gain understanding of critical parameters.
- Extensive experimental demonstration of the performance of the algorithm with various types of environments and recognition tasks.

In fulfilling these goals, we introduce an environment adaptation technique based on adaptive VQ by topology preserving transformations. It uses a priori

information about likely acoustical environments in the form of environment codebooks derived off-line from the reference environment codebook, *and* may adapt on-line to the test environment to improve recognition. The technique requires neither simultaneously recorded speech from the training and the testing environments nor EM-type batch iterations during testing. Instead of using stereo recorded data, the integrity of the updated VQ codebooks with respect to acoustical classes is maintained by endowing the codebooks with a topology and using transformations which preserve the topology of the reference environment.

1.3 Contributions

We may sum up the major contributions of this dissertation as follows:

- We propose using the topology of the distribution of the feature vectors as an invariant for the distortion model transformation and develop a framework which incorporates the preservation of topology in a workable manner.

- We demonstrate how the topology preservation can be included as a constraint in distortion minimization in VQ and how the minimization can be carried out with a stochastic approximation algorithm.

- We introduce an information geometry framework for topology preservation where the class posteriors of the testing environment are constrained to be within ε I-divergence proximity of the class posteriors of the reference environment.

- In the information geometry framework, we develop a constrained alternating minimization algorithm which preserves topology. We prove its local

convergence and study the relation of its global convergence to the rate at which the smoothing parameter is updated in the stochastic gradient optimization.

- We demonstrate empirically in an extensive manner the effectiveness of the approach in three different tasks which incorporate different acoustical conditions. The most notable turns out to be how the algorithm performs with cellular telephone speech which presents the most ardous challenge.

1.4 Organization

The organization of the dissertation is as follows:

In Chapter 2, we review the HMM speech recognition paradigm and existing approaches to robust recognition, detailing the VQ-class dependent compensation/adaptation techniques in the literature.

Modeling of distortion as a topology preserving transformation and the introduction of Topology Preserving AVQ (TPAVQ) are the key topics in Chapter 3. We use topology constrained distortion minimization as the basis for TPAVQ and describe shrinking of the neighborhoods during minimization to preserve the global order initially and to learn the local statistics as data accumulate.

In Chapter 4, we introduce an information geometry framework for topology preservation, the central result of which is a Phytagorean-like theorem for probability distributions on manifolds defined by the model family and observed data. We develop a constrained alternating minimization algorithm. We prove its convergence and show its equivalence to the constrained distortion minimization algorithm in Chapter 3.

Extensive experimental studies on three different tasks and environment mismatches are reported in Chapter 5. The corpora used are the Wall Street Journal, the Spoken Speed Dial and the TI Cellular Corpus. For example, the voice calling experiment on the TI Cellular Corpus shows TPAVQ decreases the word error for continuous ten digit recognition of cellular hands free microphone speech with land line trained models from 23.8% to 13.6% and the speaker dependent voice calling sentence error from 16.5% to 10.6%.

Chapter 2

Problem of Robust Speech Recognition

2.1 Introduction

In this chapter, to introduce the problem of robust speech recognition, we first review the basics of the current state-of-the-art speech recognizers. They are based on Hidden Markov Models (HMMs) which model feature vectors that include short-term spectral information as well as energy and their derivatives. The robustness of speech recognition systems to the noise and distortion introduced either by the acoustical environment or the transducer/telephone channel is fundamentally determined by the robustness of the features used. The algorithms developed in this dissertation operate on the feature space to make it a more robust representation for speech, and are largely independent of the particular temporal modeling, such as a specific type of HMMs. Therefore, the results should apply to a large class of speech recognition and speaker identification/verification systems.

The chapter is organized as follows: We review two basic aspects of speech recognition: (i) the front-end signal processing and (ii) HMM generation in Sec-

tion 2.2. The robustness problem is introduced in Section 2.3 followed in the remaining sections by a review of existing approaches to the solution of the robustness problem, grouping them as simple filtering methods, model-based methods, and data-driven methods. Finally, we compare and contrast the proposed algorithms in this thesis with other approaches and discuss their advantages and drawbacks.

2.2 Review of Speech Recognition

2.2.1 Signal Processing for the Speech Feature Vectors

Speech waveform by itself has an enormous amount of information which, for most computational approaches rule out the possibility of using it directly for classification. Speech recognition systems use a parametric representation to generate a feature vector space and pose speech recognition as a dynamic pattern recognition problem in the feature space. The most common features which have empirically proven to be most discriminative are based on short-term envelope of the spectrum. Log-spectra or cepstra obtained through either the discrete Fourier transform (DFT) or by estimating a linear prediction coefficients (LPC) model constitute the main feature vector in the majority of state-of-the-art recognizers. These feature vectors are augmented by their differenced versions to capture short-term dynamics in the speech waveform. Power or pitch information may also be included to generate a concatenated vector of speech features.

Below, we give an overview of the front-end processing used in the speech recognition experiments in this work.

Front-end Signal Processing

It involved the following steps:

1. Sampling of the speech waveform at 8 KHz.

2. Pre-emphasis filtering, $H(z) = 1 - 0.96z^{-1}$, to emphasize the spectral features with respect to the effect of the glottus.

3. Hamming windowing using a 30 ms window at 20 ms intervals (10 ms overlap), 240 samples per frame.

4. Estimation of 14 LPC coefficients by Levinson-Durbin recursion.

5. Sampling of the log-spectra of the estimated LPC model at mel-frequency intervals.

6. Computation of the differenced features as a filtered difference between the next and the previous frames.

7. Estimation of the auxiliary features such as power, voicing, and speech effort.

8. Concatenation of the 14 log-spectra, 14 derivatives and the auxiliary features to generate a 34-dimensional feature vector.

9. Karhunen-Loeve transform of the 34-dimensional space to reduce the dimensionality to 16 by selecting axes along which the variation is maximal after the whitening transform.

10. Modeling of the 16-dimensional feature stream by HMMs.

2.2.2 Hidden Markov Models

HMMs have become the most effective and widely-used statistical tool to model the temporal characteristics of the speech feature vectors. HMMs provide a reliable and intuitive way to recognize speech for a variety of applications. They also merge well with language models, an indispensable knowledge source for speech recognition. HMMs have been detailed in many sources [19, 32], therefore we only give a brief overview in this section.

Hidden Markov Models are a doubly stochastic process in which a hidden process is observed through random functions of its states. These states are assumed to be representing the states of the process that generates the speech. The model is fully specified by two sets of parameters:

1. Transition probabilities, which describe the temporal structure of state transitions.

2. Output probabily density functions, which describe the probability of observing speech features given the model is in a certain state. For the recognizer in this work, the output probability functions are multivariate normal densities.

Recognition with an HMM corresponds to identifying a sequence of hidden states which, through the output probabily density functions, produced the observed feature vector sequence. There are three problems related to this approach:

1. Scoring problem: Given a HMM and a sequence of observed feature vectors, what is the likelihood that the HMM produced the observations ? The algorithm that solves this problem is the forward-backward algorithm [32].

2. Decoding problem: Given a HMM and a sequence of observed feature vectors, what is the most likely sequence of hidden states in the model that generated the observed feature vector sequence ? The solution is provided by the Viterbi algorithm [38].

3. Estimation problem: Given a HMM and a sequence of observed feature vectors, what set of parameters has the maximum likelihood of producing the observations ? Baum-Welch algorithm [4] (or the forward-backward algorithm) addresses this problem.

2.3 Robust Speech Recognition

The HMM modeling of log-spectra has proved to be a very successful approach to recognition of speech when the training and the testing environments are identical. However, drastic performance degradations result when the speech data to be recognized come from an acoustical environment different from the one in which the training of the HMMs was carried out. This difference may be due to ambient noise, different microphone characteristics, most notably the wide range of non-linear effects in carbon-button handsets as opposed to higher-quality characteristics of electret handsets, or the distorting effects of a telecommunication channel. The approaches to rectify the robustness problem most closely related to the framework in this thesis are reviewed in this section. We start by a look at simple feature vector filtering techniques CMN and RASTA. We review normalization algorithms in two groups; model-based techniques in which an analytical model of the environment parameterizes the normalization algorithm, and data-driven algorithms where the corrections are general enough to correct

for a wide-range of distortions, but estimation is more costly (by requiring stereo data) due to the non-parametric nature of the approaches. Finally, we put the algorithms developed in this thesis in the context of the existing algorithms.

2.4 Cepstral Mean Normalization and RASTA

Cepstral Mean Normalization (CMN) [2] is a simple algorithm that consists of subtracting the mean of the entire stream of N feature vectors during both training and testing from each of the vectors x_n to generate the normalized stream y_n.

$$y_n = x_n - \frac{1}{N}\sum_{n=1}^{N} x_n \tag{2.1}$$

In this way, wherever the feature vector distribution has been translated in the acoustical space due to a global shift (as a result of linear filtering by the channel, for example), it is translated back to the origin. Thus, the variability among the acoustical environments in reduced. RASTA [17], effectively achieves the same goal by applying a high-pass filter to the log-spectra stream. The SRI DECIPHER system, for example, uses the high-pass filter described by the difference equation

$$y_n = x_n - x_{n-1} + 0.97y_{n-1}. \tag{2.2}$$

CMN and RASTA are routinely used in most of the speech recognition systems in addition to other compensation algorithms. Their main advantage are their simplicity and uniform gain, but when the distortion is more complex than that of a single global shift in the feature space, they arenot as affective and there is considerable room for improvement. This fact is demonstrated by the results of our speech recognition experiments with the algorithms developed in

this thesis.

2.5 Model-based Compensation Methods

One way of compensating for the effects of acoustical mismatch is explicit modeling of the environments and the form of the distortion function, usually in the form of linear filtering and additive noise. In this section, we review Codeword Dependent Cepstral Normalization (CDCN) [1], Parallel Model Combination (PMC)[14], and Vector Taylor Series (VTS) [29].

2.5.1 Codeword Dependent Cepstral Normalization (CDCN)

CDCN[1] attempts to model the distortion imposed on cepstrum coefficients caused by an environment mismatch, by linear filtering, and additive noise. It is also assumed that the "clean" cepstral vectors are distributed with a Gaussian mixture in a reference environment and the secondary environments' cepstral vectors are distorted versions of the clean cepstral vectors. In the first stage of the algorithm, parameters of the distorting function are estimated via maximum likelihood using an EM-algorithm. Once these parameters are known, they are used to obtain a Minimum Mean Square Estimate of the clean cepstral vectors.

CDCN is an affective algorithm at relatively higher Signal to Noise Ratios (SNR). It is computationally expensive due to the batch EM iterations needed to run during testing as well as training. Also it does not apply to a general set of feature vectors since it is the cepstra alone that are targeted in its development.

2.5.2 Parallel Model Combination (PMC)

PMC [14] starts from a similar assumption about the environment to that of CDCN. Previous knowledge of noise and channel vectors are assumed; these need to be estimated in advance via various approximations. The goal is to transform the mean vectors and covariance matrices of the acoustical distributions of the Hidden Markov Models (HMMs) of clean speech so that they fit the noisy feature vector stream at hand solving the mismatch problem. Estimation of channel and noise vectors beforehand is a serious limitation for PMC, as its effectivity relies heavily on the availability of sufficient amount of isolated noise samples.

2.5.3 Vector Taylor Series (VTS)

VTS [29] uses a detailed analytical model of the environment and iteratively learns the parameters of the environment and the correction factors which transform the noisy cepstral sequence so that it fits the distribution of the reference environment better. It can work with very small amounts of data due to the detailed characterization of the cepstral space. It suffers from what all the detailed model-based approaches suffer from, a high bias when the distortion is more involved or qualitatively different than that of the model.

2.6 Data-driven Compensation Methods

Data-driven techniques take the view that the complex distortions brought about by environment mismatches can be learned from exemplars of feature vectors transformed by the transducer/channel effects. Additive correction factors are usually sufficient to model the effects of such mismatches on log-spectra and cep-

stra. There is a fundamental drawback to the existing data-driven techniques in that they require "stereo" data to estimate their parameters, that is, they require speech recorded from the same source simultaneously in both the training and the testing environments. For most practical situations, this is an unacceptable limitation, since it is costly and often times simply impractical to collect stereo data from all environments of interest in an application.

In this section, we review, two representative techniques: CMU's Fixed Codeword Dependent Cepstral Normalization (FCDCN) [1] and SRI's Probabilistic Optimum Filtering (POF) [30].

2.6.1 Fixed Codeword Dependent Cepstral Normalization (FCDCN)

FCDCN [1] uses a VQ codebook to represent the reference cepstrum distribution and computes codeword dependent corrections for cepstral vectors based on stereo data. Its main limitation is its reliance on stereo data for every environment that can be encountered in testing.

2.6.2 Probabilistic Optimum Filtering (POF)

POF [30] starts from a similar approach, but the codeword-dependent corrections are not additive but consist of multi-dimensional transversal filters. This way, it is not limited to static corrections as in FCDCN but can actually model temporal correlations in the environments. It learns the filter parameters by minimizing the norm of the difference between the training environment feature vectors and the testing environment feature vectors on a stereo database. It also suffers from

the stereo database requirement. Both of the data-driven techniques impose very little structure on the mismatch, therefore pay the price in needing a very specific set of data for estimation.

2.7 Maximum Likelihood Linear Regression (MLLR)

It is hard to classify MLLR [22] as either model-based or data-driven, because while it does assume a model of distortion for the reference environment HMM model in terms of the model's means and covariance matrices, the model simply consists of a generic Affine transformation independent of the underlying feature vectors and/or the specific type of channel characteristics. The parameters of the Affine transform are estimated by maximizing likelihood. It was originally introduced as a speaker adaptation technique, although it has proven effective in combating environment mismatches as well [39].

2.8 The Proposed Algorithm and Related Prior Work

In this thesis, we take a data-driven, non-parametric approach to the modeling of feature vectors in the acoustical space. In this regard, the developed normalization algorithm, Topology Preserving Adaptive Vector Quantization (TPAVQ) uses codeword dependent corrections similar to techniques FCDCN and POF discussed in this chapter. It was pointed out that FCDCN and POF required stereo recordings of speech from both the training and testing environments,

since they made very little assumptions about the distortion in the acoustical environment. TPAVQ, on the other hand, is based on an invariance principle that stipulates that the neighborhood relations in the codebook must be preserved in all the acoustical environments. The way this principle is enforced is; the topology of the training environment is determined first by computation of the class posteriors, and this topology is imposed as an optimization constraint during the estimation of the VQ codebooks for the testing environments. The codebooks estimated this way correspond to each other on a per codevector basis, and therefore class integrity can be preserved during the estimation of the codeword dependent corrections *without* stereo data. This allows the technique to be used in a variety of environments where non-stereo recorded data are easily available, but stereo data are costly or impractical to collect, such as speech over cellular telephone channels.

Chapter 3

Topology Preserving Adaptive Vector Quantization

3.1 Introduction

There is one underlying observation about the nature of speech representations which has lead to the development of normalization techniques in this thesis: In any class-based representation of speech there is a structure of classes in the feature space which is invariant to acoustical variations in the environment in which it was produced, such as the vocal tract or prosodic characteristics of the speaker and the spectral distortion and noise present in the medium. More specifically, with respect to the metric on the representation feature space, there are clusters of classes with similar characteristics. For example, at the phone level, one would expect classes representing vowels to be closer to each other than those that represent fricatives. Moreover, the degree to which human understanding of speech is robust suggests that this structure is invariant to the characteristics of the speaker or the recording/transmission apparatus used.

Short-term stationary modeling of speech features with tools such as Gaus-

sian mixture models (GMM) or vector quantization (VQ) representations do not readily provide the means to express or use to advantage this invariance. In this chapter, we introduce a way to express the topology of a VQ codebook (or a GMM) in terms of a posterior probability mass function defined over the classes. We also show how to quantify the preservation of this topological structure in order to use the invariance described above as a constraint in estimation of models. The starting point for the framework is the introduction of a class-dependent transformation model for feature space distortion with an additional topology preservation constraint. The topology preservation constraint is expressed in terms of the posterior class probability mass functions. The class biases are not free parameters but as a set have to ensure that the posterior class probabilities of the testing environment have to be in ε proximity of the posterior class probabilities of the reference environment with respect to the Kullback-Leibler divergence. Estimation under such a constraint is proposed as a solution to the class labelling problem of non-stereo data and allows derivation of practical normalization algorithms.

The chapter is organized as follows: First, we review the feature space for our speech representation, then the unconstrained class-dependent bias model for distortion with its easy estimation in the presence of stereo data and the estimation problem with non-stereo data. Next, we introduce our probabilistic description of the topology of a VQ codebook via Bennett-style [1] high-rate assumptions. Once we have the mathematical tool to express the topologies, we re-introduce the class-dependent bias model *with* the topology preservation constraint between the reference and the secondary environments. This model

[1]The adjective was coined by Gersho and Gray in [13].

allows for estimation with unlabeled channel (non-stereo) data. Then follows the constrained optimization which results in an adaptive vector quantization algorithm reminiscent of the Self-Organizing Feature Map [21] in which the topology matrix is not an arbitrary lower-dimensional mesh but specified through Kullback-Leibler divergences between the high-rate posterior class pmfs. The smoothing parameter associated with the probabilistic topology description is linked to a model complexity framework in the final section which informally suggests the connection between topology preservation and correct model complexity selection as data accumulate. This allows us to come up with a "cooling schedule" for the smoothing parameter via an optimal trade-off between bias and variance in a kernel density estimation framework.

3.2 A Distortion Model for the Feature Space

Our goal is to develop a general model of distortion applicable to a broad range of feature spaces. Accordingly, we develop distortion models on $\mathbb{R}^p$ without explicit reference to the individual dimensions which may include spectral information such as cepstral coefficients, voicing estimates or their derivatives. This is a decision of simplicity for the sake of generality over using more domain knowledge for a specific feature vector. Using a specific degradation model for a specific feature suffers from the limitations of model-based approaches, namely the bias associated with the limited complexity model far outweighs the gain in variance during estimation. In the next section, we outline one such feature vector, specifically the one used for the experiments in this thesis.

3.2.1 The Feature Space

The output space of the discrete-time models in speech recognition is typically a vector space whose components are short-term stationary speech attributes. In the front-end of the HMM speech recognizer used in this work, a broad range of features such as frame energy, voicing, spectra and their derivatives are concatenated to form a 34-dimensional feature vector. Principal component analysis is applied to this vector space to reduce dimensionality to 16 by selecting a subset of axes along which statistical variation is maximal. We denote the resulting principal component vector space by $\mathcal{F}$. Vector quantization is applied to $\mathcal{F}$, therefore, members in a class are related by both static and dynamic features which determine the way they are affected by the environment.

3.2.2 Smooth class dependent translation distortion model

Let us start by considering a VQ based density estimation technique for the reference environment. Given a training set Z from the reference environment

$$\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_N)\} \in \mathcal{F}^N, \tag{3.1}$$

we estimate a vector quantizer Q_Λ, which is a mapping from $\mathcal{F}$ to a finite set of reproduction vectors (code vectors) $\Lambda = \{\lambda_1, \ldots, \lambda_K\} \in \mathcal{F}^K$, *i.e.*

$$Q_\Lambda : \mathcal{F} \longrightarrow \{\lambda_1, \ldots, \lambda_K\} \subset \mathcal{F} \tag{3.2}$$

$$Q_\Lambda(z) = \lambda_{w_\Lambda(z)}, \quad \forall z \in \mathcal{F} \tag{3.3}$$

where

$$w_\Lambda(z) = \arg \min_{k \in \{1,\ldots,K\}} d(z, \lambda_k). \tag{3.4}$$

The codebook Λ is estimated by minimizing the the distortion

$$D = E[d(\mathbf{Z}, Q_\Lambda(\mathbf{Z}))] \tag{3.5}$$

which, in practice, is replaced by the long-term sample average, in our case, the mean-squared error (MSE) over the training set

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d(Z_i, \lambda_{w_\Lambda(Z_i)}) = \frac{1}{N}\sum_{i=1}^{N} |Z_i - \lambda_{w_\Lambda(Z_i)}|^2. \tag{3.6}$$

The quantizer Q_Λ partitions $\mathcal{F}$ into Voronoi cells R_k^Λ

$$R_k^\Lambda = \{z \in \mathcal{F} : Q_\Lambda(z) = \lambda_k\} \tag{3.7}$$

Model 1 *Let Λ and Θ be the codebooks for the reference and the secondary environments, respectively. Then, the codevectors are related through a translation*

$$\theta_i = \lambda_i + \delta_i, \quad i = 1, \ldots, K \tag{3.8}$$

and the overall tranformation, $T(\cdot)$ is given by

$$T(z) = z + \sum_k \left(\frac{\phi_{\sigma^2}(|z - \lambda_k|)}{\sum_{i=1}^{K} \phi_{\sigma^2}(|z - \lambda_i|)} \right) [\theta_i - \lambda_i], \quad \forall z \in \mathcal{F} \tag{3.9}$$

where a Gaussian smoothing of the membership function is introduced by placing multivariate Gaussians

$$\phi_{\sigma^2}(r) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{3.10}$$

at the codevectors with σ determining the smoothness of the transformation. Note that, as we shall soon see, this is equivalent to assuming an underlying kernel density for the VQ codevectors with σ as the smoothing parameter.

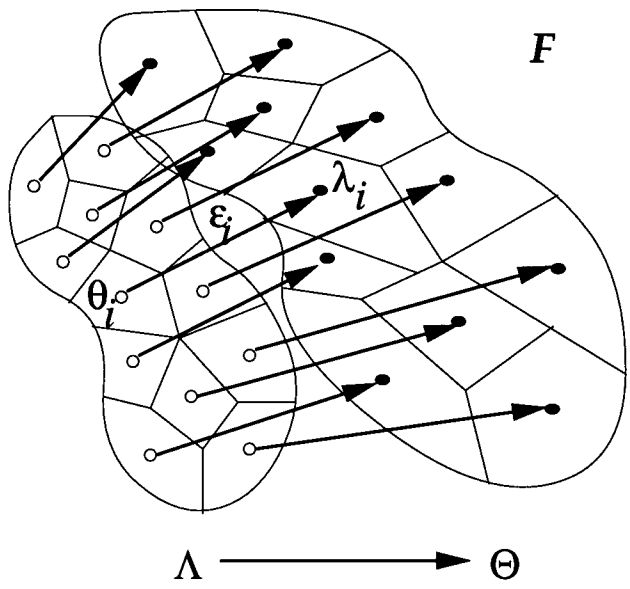

Figure 3.1: Non-parametric, class dependent translation model for the mismatch between two acoustical environments

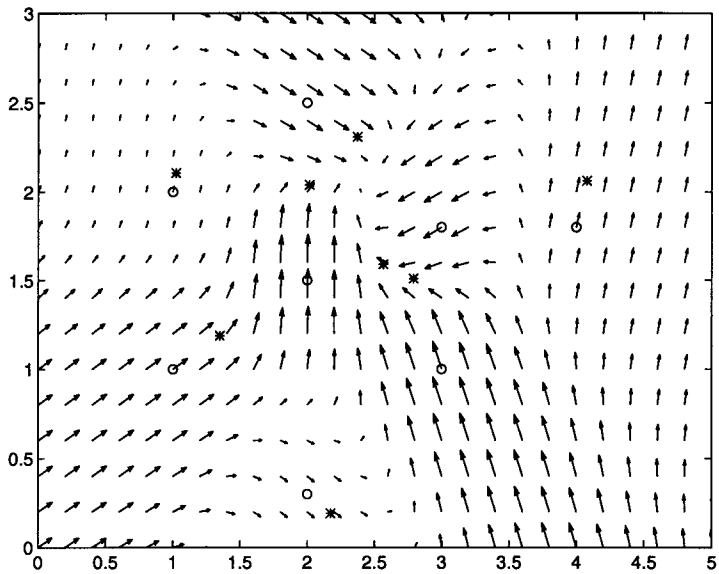

Figure 3.2: The difference vector field between two acoustical environments $\Lambda(\circ)$ and Θ $(*)$in a 2D feature space.

3.2.3 Class dependent normalization with labeled ("stereo") data

First, let us review the case of labeled data from multiple acoustical environments for which an easy normalization technique exists and has been extensively used in literature. Assume we are given the following training set of mutually labeled, "stereo", training vectors from the reference and secondary environments:

$$(\mathbf{Z}, \mathbf{X}) = \{(Z_1, X_1), (Z_2, X_2) \ldots, (Z_N, X_N))\} \in (\mathcal{F} \times \mathcal{F})^N, \tag{3.11}$$

that is,

$$X_i = T(Z_i), \quad \forall i = 1, \ldots, N. \tag{3.12}$$

Consider the VQ mapping, Q_Λ of 3.2 which partitions $\mathcal{F}$ into Voronoi cells R_k^Λ

$$R_k^\Lambda = \{z \in \mathcal{F} : Q_\Lambda(z) = \lambda_k\}. \tag{3.13}$$

Observe that due to 3.12, we have a decomposition for the secondary environment as the direct images of the reference partition under T:

$$\hat{R}_k^\Theta = T(R_k^\Lambda) \tag{3.14}$$

This need not be a partition since it does not necessarily hold that

$$\hat{R}_j^\Theta \cap \hat{R}_k^\Theta = \emptyset \quad \forall j \neq k. \tag{3.15}$$

Still, this decomposition lends itself to estimation of secondary class codevectors via the VQ class optimality condition:

$$\hat{\theta}_i = \mathbf{E}[x | x \in \hat{R}_i^\Theta] = \mathbf{E}[x | T^{-1}(x) \in R_i^\Lambda] \tag{3.16}$$

$$\hat{\theta}_i \approx \frac{1}{|\hat{R}_i^{\Lambda}|} \sum_{\{(z,x)\in(\mathbf{Z},\mathbf{X})|z\in\hat{R}_i^{\Lambda}\}} x \tag{3.17}$$

In accordance with the distortion model of 1, we have thus estimated parameters for the following normalization technique:

Normalization 1 *Let Λ and $\hat{\Theta}$ be the codebooks for the reference and the secondary environments, respectively. Then, the codevectors are normalized through class-dependent translations:*

$$\hat{\delta}_i = \hat{\theta}_i - \lambda_i, \quad i = 1, \ldots, K \tag{3.18}$$

which result in the overall normalization transformation

$$T^*(x) = x - \sum_k \left(\frac{\phi_{\sigma^2}(|z - \theta_k|)}{\sum_{i=1}^{K} \phi_{\sigma^2}(|z - \theta_i|)} \right) \hat{\delta}_i, \quad \forall x \in \mathcal{F} \tag{3.19}$$

Labeled data are costly and in many important cases impractical to collect, e.g. via a cellular connection in a moving vehicle. This fact limits the usefulness of the simple estimation in 3.17 for Normalization 1. What are widely and much more cheaply available are unlabeled data, that is, independent recordings from various environments. The lost information in the unlabeled case makes the estimation problem an incomplete observation problem. Therefore, for a stable solution we have to augment the problem with additional information, that is, constraints on the solution. That is exactly what we do in the next section in the form of putting constraints on the neighborhood relations of the codevectors via a probabilistic description of topology which are preserved between the reference and the secondary environments.

3.3 A Probabilistic Description of Topology

We need a constraint on the allowable solutions for the secondary environment codebook to augment the ill-defined problem which results from unlabeled data. The new proposal in this section is to capture the neighborhood relations between codevectors in a form which can be used to solve the incomplete observation problem. Our main goal is to quantify the neighborhood relations between the codevectors. The VQ codebook as introduced above does not provide a ready apparatus for accomplishing this goal. To that end, we introduce the following estimator for the density of Z:

$$f_{\Lambda}^{\sigma^2}(z) = \frac{1}{K}\sum_{i=1}^{K}\phi_{\sigma^2}(|z-\lambda_i|), \quad \forall z \in \mathcal{F} \tag{3.20}$$

where

$$\phi_{\sigma^2}(r) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{3.21}$$

This is a classical Gaussian kernel estimator of the density of *quantized* Z with smoothing parameter σ^2. [36] In using the codevectors only, rather than Z, with equal weighting of all the Voronoi cells, we have implicitly made the following "high-rate" (large K) assumption:

Assumption 1

$$P(z \in R_k^{\Lambda}) = \frac{1}{K}, \quad k = 1, \ldots, K, \tag{3.22}$$

that is, the Voronoi cells have the same probability mass and edge effects or "overload noise" [2] *can be neglected for* K *large.*

[2] Overload noise is the distortion due to the quantization of data outside the coverage of the finite number of code vectors.

With $f_\Lambda^{\sigma^2}$ as a model for the density, now we may consider posterior class probabilities. Note that $f_\Lambda^{\sigma^2}$ is a Gaussian mixture model with K equally likely classes and a common variance σ^2. The reason we chose not to introduce it as such is that σ^2 will truly be used as a smoothing parameter in the following development rather than a variance.

Now, let us introduce the classes as an unobserved random vector augmenting the data Z:

$$(\mathbf{Z}, \mathbf{C}) = \{(Z_1, C_1), (Z_2, C_2), \ldots, (Z_N, C_N)\} \in (\mathcal{F} \times \{1, \ldots, K\})^N, \tag{3.23}$$

and define the class mapping as

$$c(Z) = C, \quad \forall (Z, C) \in \mathcal{F} \times \{1, \ldots, K\}. \tag{3.24}$$

Then, the posterior class probabilities are given as

$$p_\Lambda^{\sigma^2}(k|z) \equiv P(c(Z) = k | Z = z) = \frac{\phi_{\sigma^2}(|z - \lambda_k|)}{\sum_{i=1}^K \phi_{\sigma^2}(|z - \lambda_i|)} \quad k \in \{1, \ldots, K\} \tag{3.25}$$

These posterior class probabilities are the first step towards a useful description of topology. Given a data point z, we can deduce how close the classes i and j are are by looking at probabilities $p_\Lambda^{\sigma^2}(i|z)$ and $p_\Lambda^{\sigma^2}(j|z)$. Yet, the posteriors in 3.25 are in terms of the data Z. The next development will use an approximation to make them viable topological descriptions.

At this point, we introduce a Bennett-style high-rate approximation for Q_Λ:

Assumption 2 *Given ε and σ^2, one can select K large enough so that*

$$|f_\Lambda^{\sigma^2}(z) - f_k^\Lambda| \leq \varepsilon, \quad \forall z \in R_k^\Lambda, \quad k = 1, \ldots, K, \tag{3.26}$$

where $f_k^\Lambda > 0$, that is, the Voronoi cells are small enough so that the smoothed density on a given cell is approximately constant.

With the high-rate assumption, Assumption 2, the following result immediately follows:

Lemma 1

$$p_{\Lambda}^{\sigma^2}(k|z) \to p_{\Lambda}^{\sigma^2}(k|Q_{\Lambda}(z)) \tag{3.27}$$

as $K \to \infty$.

Proof: Write the posteriors in terms of $F_{\Lambda}^{\sigma^2}$:

$$p_{\Lambda}^{\sigma^2}(k|z) = \frac{f_{\Lambda}^{\sigma^2}(z|c(z)=k)\frac{1}{K}}{f_{\Lambda}^{\sigma^2}(z)} \tag{3.28}$$

Now, $Q_{\Lambda}(z) = \lambda_{w_{\Lambda}(z)}$, therefore $z \in R_{w_{\Lambda}(z)}^{\Lambda}$. Then, by the high-rate approximation Assumption 2, as $K \to \infty$,

$$f_{\Lambda}^{\sigma^2}(z) \to f_{w_{\Lambda}(z)}^{\Lambda} \tag{3.29}$$

But, $\lambda_{w_{\Lambda}(z)} \in R_{w_{\Lambda}(z)}^{\Lambda}$ as well:

$$f_{\Lambda}^{\sigma^2}(\lambda_{w_{\Lambda}(z)}) \to f_{w_{\Lambda}(z)}^{\Lambda} \tag{3.30}$$

as $K \to \infty$, which immediately imply

$$p_{\Lambda}^{\sigma^2}(k|z) \to p_{\Lambda}^{\sigma^2}(k|\lambda_{w_{\Lambda}}(z)) = \frac{\phi_{\sigma^2}(|\lambda_{w_{\Lambda}(z)} - \lambda_k|)}{\sum_{i=1}^{K} \phi_{\sigma^2}(|\lambda_{w_{\Lambda}(z)} - \lambda_i|)}. \tag{3.31}$$

Lemma 1 provides us with a framework in which we can define a convenient form of code vector neigborhood (topology) preservation between a reference environment and a secondary environment. This is accomplished through 3.31, in which the approximation posterior pmf, which we shall use as a probabilistic description of the topology from now on, is shown to be dependent on the data only through the code vector indices. Our goal in the next section is to introduce topology preservation as the invariance of the class posterior as approximated by Lemma 1.

3.4 Topology Preserving Class Dependent Translation Model for Distortion

Let

$$\Theta = \{\theta_1, \ldots, \theta_K\} \in \mathcal{F}^K \tag{3.32}$$

be the set of code vectors to be estimated for the secondary environment. According to our model,

$$\theta_i = \lambda_i + \varepsilon_i, \quad i = 1, \ldots, K \tag{3.33}$$

Therefore, we need to estimate ε_i, or equivalently θ_i by using a training set from the secondary environment

$$X = \{X_1, X_2, \ldots, X_N)\} \in \mathcal{F}^N, \tag{3.34}$$

in a manner similar to that of 3.6. In this case, however, vector quantization of X must be carried out in a way that ensures the integrity of the ordering of the indices is preserved. Without ordered indices of the reference and the secondary environments, a normalization of the sort 1 is not possible..

The unconstrained distortion model of 1 offers no assistance in the satisfaction of this requirement. In this section, we introduce topology preservation as an invariance between reference and secondary environments, which allows us to estimate Θ while keeping the integrity of the cross class labels intact.

First, let us define the sets of class posterior probabilities for the codebooks Λ and Θ, $\mathcal{P}_\Lambda^{\sigma^2}$ and $\mathcal{P}_\Theta^{\sigma^2}$, respectively.

$$\mathcal{P}_\Lambda^{\sigma^2} = \{p_\Lambda^{\sigma^2}(\cdot|k), k = 1, \ldots, K\} \tag{3.35}$$

where the pmfs $p_\Lambda^{\sigma^2}(\cdot|k)$ are

$$p_\Lambda^{\sigma^2}(\cdot|k) \equiv p_\Lambda^{\sigma^2}(\cdot|\lambda_k) = \frac{\phi_{\sigma^2}(|\lambda_{(\cdot)} - \lambda_k|)}{\sum_{i=1}^K \phi_{\sigma^2}(|\lambda_{(\cdot)} - \lambda_i|)} \quad k = 1, \ldots, K \tag{3.36}$$

and similarly for Θ.

Finally, we have arrived at a description of code vector neighborhood by means of which we can express the topology invariance. In the presence of distortion and noise, the code vectors will be transformed in the feature space, but the transformation will leave neighborhood relationships essentially the same. To this end, we require,

$$\mathcal{P}_{\Lambda}^{\sigma^2} \equiv \mathcal{P}_{\Theta}^{\sigma^2} \tag{3.37}$$

With (3.37) as the constraint of topology preservation between environment codebooks Λ and Θ, we can now state the Topology Preserving Class-dependent Translation Model (TPCDTM):

Model 2 *Let Λ and Θ be the codebooks for the reference and the secondary environments, respectively. Then, the codevectors are related through a translation*

$$\theta_i = \lambda_i + \delta_i, \quad i = 1, \ldots, K \tag{3.38}$$

subject to

$$\mathcal{P}_{\Lambda}^{\sigma^2} \equiv \mathcal{P}_{\Theta}^{\sigma^2} \tag{3.39}$$

and the overall tranformation, $T(\cdot)$ is given by

$$T(z) = z + \sum_{k} \left(\frac{\phi_{\sigma^2}(|z - \lambda_k|)}{\sum_{i=1}^{K} \phi_{\sigma^2}(|z - \lambda_i|)} \right) [\theta_i - \lambda_i], \quad \forall z \in \mathcal{F} \tag{3.40}$$

where $\varepsilon > 0$.

In the next section, we will see how the additional constraint enters the distortion minimization to obtain a normalization algorithm.

3.5 Topology Preserving Adaptive Vector Quantization

In this section, we will see how the distortion model TPCDTM, Model 2, developed in the previous section can be used to solve the index integrity problem and derive normalization algorithms for the unlabeled ("mono") data case.

For the constrained optimization, consider the minimization problem

$$\Theta^*(\sigma^2) = \min_{\Theta \subset \mathbb{R}^K} \mathbf{E}[||x - \theta_{c_\Theta(x)}||^2] \tag{3.41}$$

subject to

$$p_\Lambda^{\sigma^2}(\cdot|l) \equiv p_\Theta^{\sigma^2}(\cdot|l) \quad l = 1, \ldots, K$$

Note that the objective function is a σ^2-smoothed version of the distortion in 3.5 which converges to D as $\sigma^2 \to 0$. At the limit, the membership function becomes a step function whence the first term tends to D and the topology constraint disappears as the effective neighborhood radius shrinks to zero. Therefore, the solution we are interested in is

$$\Theta^*(0) \equiv \lim_{\sigma^2 \to 0} \Theta^*(\sigma^2) \tag{3.42}$$

Equation 3.41 can be re-written via the double expectation formula as

$$\Theta^*(\sigma^2) = \min_{\Theta \subset \mathbb{R}^K} \mathbf{E}[\mathbf{E}[||x - \theta_k|^2 || w_\Theta(x)]] \tag{3.43}$$

subject to

$$p_\Lambda^{\sigma^2}(\cdot|l) \equiv p_\Theta^{\sigma^2}(\cdot|l) \quad l = 1, \ldots, K$$

or equivalently

$$\Theta^*(\sigma^2) = \min_{\Theta \subset I\!R^K} \mathbf{E}[\sum_k p_\Theta^{\sigma^2}(k|w_\Theta(x))||x - \theta_k||^2] \tag{3.44}$$

subject to

$$p_\Lambda^{\sigma^2}(k|l) \equiv p_\Theta^{\sigma^2}(k|l) \quad l = 1, \ldots, K.$$

Upon substitution of the reference measure in place of the secondary environment measure in the minimization as required by the constraint, this leads to

$$\Theta^*(\sigma^2) = \min_{\Theta \subset I\!R^K} \mathbf{E}[\sum_k p_\Lambda^{\sigma^2}(k|w_\Theta(x))||x - \theta_k||^2]. \tag{3.45}$$

Given the training set $\mathbf{X}$, we can estimate $\Theta^*(\sigma^2)$ by

$$\hat{\Theta}^*(\sigma^2) = \min_{\Theta \subset I\!R^K} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} p_\Lambda^{\sigma^2}(k|w_\Theta(X_i))||X_i - \theta_k||^2 \tag{3.46}$$

Note that the cost function 3.46 is an augmentation of distortion by contributions from all the codevectors in proportion to their topological relation to the winning codevector in the *reference* environment as reflected in the class posterior. Minimization of the final expression 3.46 is accompished by a stochastic gradient descent algorithm. The incremental update is given by

$$\Theta^{(t+1)} = \Theta^{(t)} - \eta(t) \frac{\partial}{\partial \Theta} \mathbf{E}\left[\sum_k p_\Lambda^{\sigma^2(t)}(k|w_\Theta(x))||x - \theta_k||^2\right] \tag{3.47}$$

which can be written with respect to the individual θ_j as

$$\frac{\partial}{\partial \theta_j} \sum_{k=1}^{K} p_\Lambda^{\sigma^2}(k|w_\Theta(x))||x - \theta_k||^2 = 2p_\Lambda^{\sigma^2}(j|w_\Theta(x))(x - \theta_j). \tag{3.48}$$

This results in the following algorithm:

Algorithm 1 *Topology Preserving Adaptive VQ:*

0. Set the initial codebook to that of the reference environment.

$$\Theta^{(0)} = \Lambda \tag{3.49}$$

1. Update with the stochastic gradient at the current smoothness $\sigma(t)$

$$\theta^{(t+1)} = \theta^{(t)} - \eta(t) p_{\Lambda}^{\sigma^2(t)}(j|w_{\Theta}(X_t))(X_t - \theta_j). \tag{3.50}$$

2. Repeat 1 until convergence as $\sigma^2(t) \to 0$ *and* $\eta(t) \to 0$

Minimization of the final expression via a stochastic gradient descent results in our estimation algorithm which is described in this section.

In the optimization algorithm, the limiting and gradient descent minimization are carried out together in such a way as to adapt instantaneous model complexity to adjust bias and variance so as to minimize instantaneous mean squared error.

This may be understood more easily in terms of the notion of an effective number of parameters. In the TPAVQ algorithm, the parameters are tied to each other by the class posterior pmf which effectively decides which codevectors will be updated depending on which index belongs to the winner. At the initialization and early stages of the algorithm when σ is very large, $\sigma \to \infty$, the class posterior is a uniform distribution. This results in *all* the codevectors being updated towards the data point X_t regardless of who the winner is, that is, the parameters, the class means, are not free. In fact, when $\sigma \to \infty$ the distribution of the acoustic feature vectors tends to a uniform distribution for which *one* parameter (the overall mean, for example) suffices to describe. As $\sigma \to 0$ slowly, the codevectors become more and more independent of each other as the updates will now depend strongly on who the winner is. One can therefore speak of an

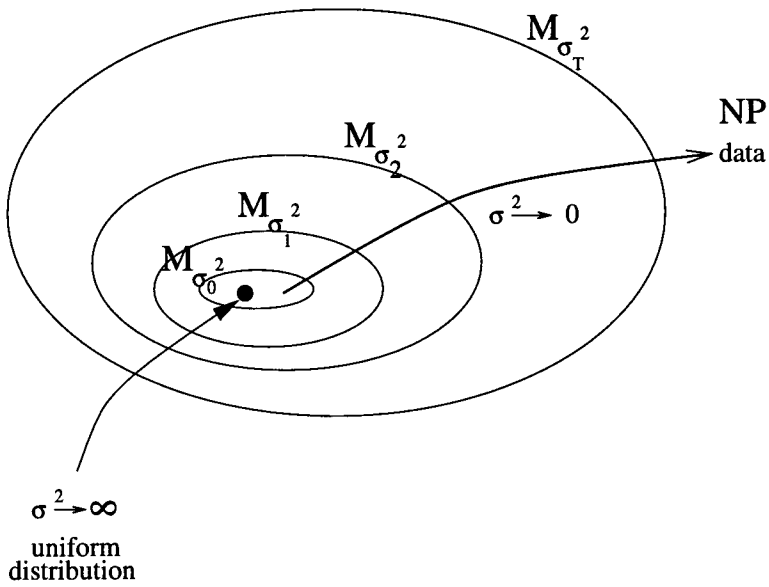

Figure 3.3: Model spaces smoothed by σ: $\infty > \sigma_0 > \sigma_1 > \sigma_2 > \cdots > \sigma_T > 0$; $M_\infty = \{(\frac{1}{K}, \ldots, \frac{1}{K})\} \subset M_{\sigma_0} \subset M_{\sigma_1} \subset M_{\sigma_2} \subset \cdots \subset M_{\sigma_T} \subset NP \equiv$ data

effective number of parameters, $K_{\text{eff}}(t)$, which starts from 1 and converges to K, the number of codevectors as σ is decreased slowly from a very large value.

$$\begin{aligned} 1 < \quad K_{eff}(\sigma(t)) \quad < K \\ \infty > \qquad \sigma(t) \qquad > 0 \end{aligned} \tag{3.51}$$

Corresponding to the sequence of $K_{eff}(t)$ is a sequence of model spaces, $\mathcal{M}_{\sigma(t)}$

$$\mathcal{M}_\infty = \{(\frac{1}{K}, \ldots, \frac{1}{K})\} \subset \mathcal{M}_{\sigma_0} \subset \mathcal{M}_{\sigma_1} \subset \mathcal{M}_{\sigma_2} \subset \cdots \subset \mathcal{M}_{\sigma_T} \subset \mathcal{M}_0 \tag{3.52}$$

depicted in Fig. 3.3. $\mathcal{M}_\infty$ consists of the uniform distribution and $\mathcal{M}_0$ is the models allowed by the set of VQ codevectors with the hard decision. Then, we can think of the distribution as a Gaussian kernel estimator of the quantized $\mathbf{X}$ with K_eff number of points (kernels) with σ as the smoothing parameter. Therefore, by quantizing the acoustic feature vectors we have reduced the problem of choice of $\sigma^2(t)$ to one of optimal smoothing parameter selection depending

on the number of data points, a well-worked problem in kernel density estimation [36]. Let K denote the multi-variate kernel,

$$\int_{\mathbb{R}^p} K(\mathbf{x})d\mathbf{x} = 1, \tag{3.53}$$

which, in our case, is a multi-variate normal

$$K(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} e^{-\frac{1}{2\sigma^2}\mathbf{x}^T\mathbf{x}}. \tag{3.54}$$

Let us define the following two variables [36]

$$\alpha = \int t_1^2 K(\mathbf{x})d\mathbf{x} \tag{3.55}$$

and

$$\beta = \int K(\mathbf{x})^2 d\mathbf{x}. \tag{3.56}$$

The bias of a density estimator, $\hat{f}$ can be written via a vector Taylor series expansion [36]

$$\text{bias}\hat{f}(\mathbf{x}) \approx \frac{1}{2}\sigma^2\alpha\nabla^2 f(\mathbf{x}). \tag{3.57}$$

Also, the variance of the same estimator is given by

$$\text{var}\hat{f}(\mathbf{x}, t) \approx \frac{1}{t}\frac{1}{\sigma^d}\beta f(\mathbf{x}) \tag{3.58}$$

Now, taking mean integrated square error (MISE) as our complexity optimizer via bias-variance trade-off,

$$\text{MISE}_\sigma(\hat{f})(t) \approx \int \left([\text{bias}\hat{f}(\mathbf{x})]^2 + \text{var}\hat{f}(\mathbf{x}, t)\right) d\mathbf{x} \tag{3.59}$$

we can write MISE as

$$\text{MISE}_\sigma(\hat{f})(t) \approx \frac{1}{4}\sigma^4\alpha^2 \int \left(\nabla^2 f(\mathbf{x})\right) d\mathbf{x} + \frac{1}{t}\frac{1}{\sigma^d}\beta \tag{3.60}$$

which needs to be minimized with respect to σ

$$\sigma_{opt}(t) = \arg\min_{\sigma} \mathrm{MISE}_{\sigma}(\hat{f})(t). \tag{3.61}$$

Minimization of the mean integrated square error, which is the combination of bias and variance terms, results in:

$$\sigma_{opt}(t) = \left(\frac{4}{p+2}\right)^{\frac{1}{p+4}} \left(\frac{1}{t}\right)^{\frac{1}{p+4}} \tag{3.62}$$

This suggestion for a cooling schedule relies on a high-rate approximation, and produces a very slow rate for σ to shrink. This theoretical rate cannot be tolerated in practice, and generally much faster schemes are used. Interesting things to note are: As the dimensionality of the space, p, increases, the rate must further decrease. This makes intuitive sense; with large dimensional feature vectors topological variations to avoid during optimization increase exponentially, therefore σ must shrink much slower.

3.6 Extension to Multiple Environments

The off-line codebook adaptation described in the previous section is carried out for a set of representative environments as shown in Figure 3.4. Once the codebooks are available, they are simply used as a basis in which to express the data from an unknown environment. For a discrete density HMM, the technique may be regarded as codebook adaptation, for a continuous density HMM, such as the one used in this work, it is necessary to put it in the form of a compensation algorithm as follows. Let the incoming speech feature vector(t-th frame of the utterance) from the unknown test environment be denoted as $\mathbf{x}(t)$. Then, the

compensated feature vector, $\hat{\mathbf{x}}(t)$ is computed as

$$\hat{\mathbf{x}}(t) = \mathbf{x}(t) + \sum_h P_h \sum_k p_k^h(t)[\mathbf{x}_k^{ref} - \mathbf{x}_k^h(t)] \tag{3.63}$$

where the probability that the t-th frame belongs to Voronoi region k in the codebook h, $p_k^h(t)$, and the probability that the utterance belongs to environment h, P_h are estimated as

$$p_k^h(t) = \frac{e^{-\beta(\mathbf{x}_k^h(t)-\mathbf{x}(t))^2}}{\sum_k e^{-\beta(\mathbf{x}_k^h(t)-\mathbf{x}(t))^2}} \tag{3.64}$$

$$P_h = \frac{e^{-\alpha\sum_n(\mathbf{x}_w^h(t)-\mathbf{x}(t))^2}}{\sum_h e^{-\alpha\sum_n(\mathbf{x}_w^h(t)-\mathbf{x}(t))^2}}. \tag{3.65}$$

The initial codebook selection is a fast adaptation using a priori knowledge about likely representative environments. A new testing environment may not always fit the available codebooks to give a satisfactory performance. In such cases, on-line adaptation to the new environment may be accomplished by utilizing the testing environment's data during compensation via the same topology preserving minimization algorithm. This is shown in the block diagram of the compensation in Figure 3.5. In this way, even if the initial match between the environments is not as good, it gets better as the codebooks get updated.

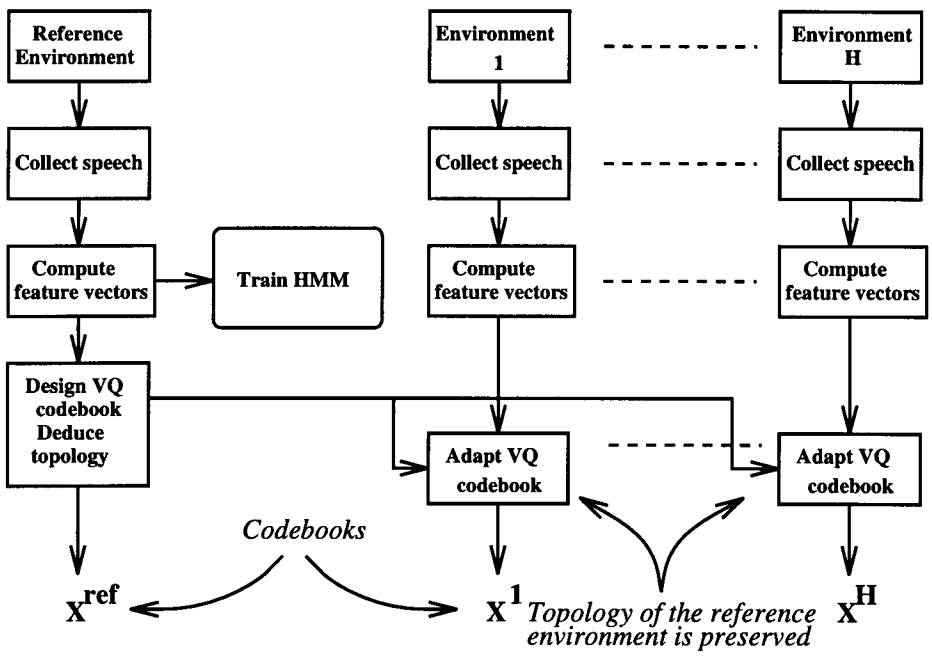

Figure 3.4: Off-line topology preserving VQ codebook adaptation for a set of representative environments.

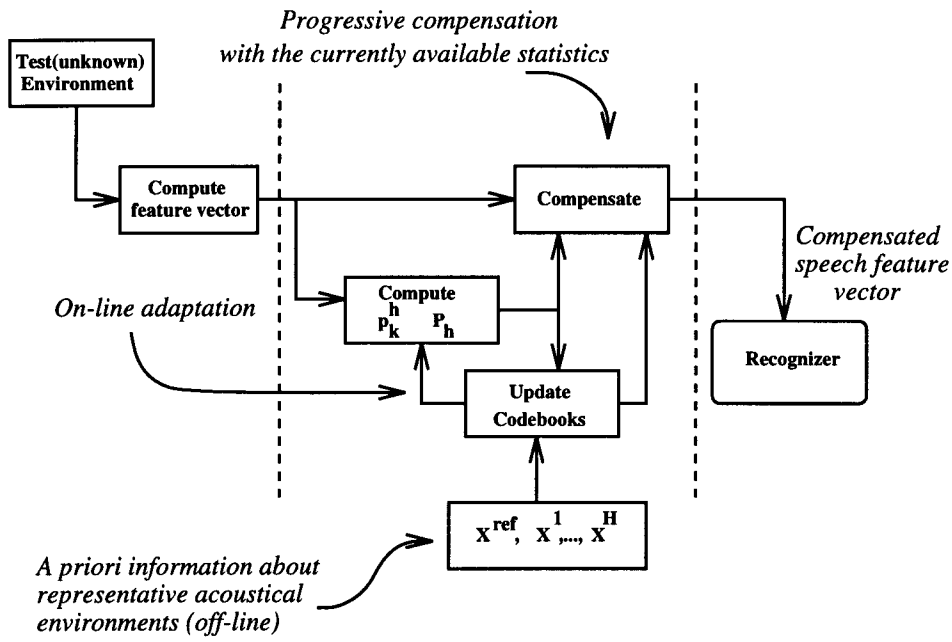

Figure 3.5: Compensation with both a priori information about likely environments and on-line adaptation.

Chapter 4

Information Geometry of Topology Preserving Adaptation

4.1 Introduction

In the preceding chapter, we started with the motivating problem of estimating a transformation for robust speech recognition and developed TPAVQ as a constrained vector quantization algorithm that enforced a novel definition of a specific type of topological invariance between acoustical environments. The preservation of topology was realized by using the conditional class posterior probabilities of the reference environment *a priori* in the parameter estimation of the secondary environment. The VQ/kernel density estimation framework proved useful in deriving a stochastic gradient descent algorithm which minimized the distortion plus a penalty for topology discrepancy expressed in terms of the Kullback-Leibler divergences between the class posteriors of the reference and testing environments.

In this chapter, we show that we can pose topology preservation in a new framework. We cast our discussion in terms of information and likelihood which

allows us to derive a constrained alternating minimization algorithm closely related to the TPAVQ. The alternating minimization framework lets us study the convergence properties of topology preservation. Specifically, we present a geometrical view by developing tools which allow us to introduce projections and Pythagorean theorems of information divergences in similar ways to those in Euclidean spaces.

The goals of the chapter are; posing topology preservation in an alternating minimization framework, developing a topology preserving alternating minimization algorithm, and investigating its relationship with TPAVQ developed in the previous chapter. The theoretical development in the chapter is as follows: Three fundamental results of Csiszár [10] form the basis of the information divergence geometry, the framework in which we study the convergence properties of topology preservation. In Section (4.2), we re-introduce information divergence for abstract measures, present and comment on three basic theorems of information geometry that determine the conditions under which information projections exist and divergence can be used in a way similar to the use of Euclidean distance in Euclidean spaces to generate Pythagorean-like theorems. In Section (4.3), we review the Csiszár-Tusnády alternating minimization framework of which (Generalized) Expectation Maximization algorithm ((G)EM) can be derived as a special case. In the sequel, this derivation of the EM algorithm as an iterative projection algorithm between two manifolds of probability distributions serves as the framework in which we develop our convergence argument for topology preserving adaptation. Section (4.4) introduces topology preservation in an alternating minimization setting. We develop a GEM-like algorithm, and analyse the convergence characteristics of topology preserving adaptation. The setting

can be described via two probability manifolds: (i) the set of probability distributions realizable by the model family parameterized by Θ, $\mathcal{M}$: and (ii) the set of probability distributions that are consistent with the observed data, $\mathcal{D}$:

There are two aspects of topology preservation via alternating minimization: *local* and *global*. First question (local) is whether the topology preserving alternating projections converge to a local minimum of the I-divergence. For proving the convergence of topology preserving alternating minimization, we use results from the first two sections to note that the Pythagorean equation holds for the "triangle" formed by the distribution in the model manifold $\mathcal{M}$, its I-projection on the data manifold $\mathcal{D}$ and the topology preserving adaptation distribution since $\mathcal{D}$ is a linear family.

The completion of the result can be sketched as follows: The topology preservation assumption allows one to bound the divergences in the iterations of the EM and the topology preserving adaptation. This results in a decreasing sequence of topology preserving adaptation iterations that follows EM iterations for topology preservation with small enough ε.

The second question (global) in topology preservation involves whether the local minimum the algorithm converges to, preserves the topology between the reference and the testing environments. This is precisely the goal of topology preservation, and we demonstrate how this works depending on how fast the smoothing parameter is varied.

Finally, in Section (4.5), we show how alternating minimization allows derivation of incremental algorithms for alternating minimization, and how this results in stochastic gradient algorithms equivalent to TPAVQ. We summarize our development and findings in Section (4.6).

4.2 Information Divergence Geometry of Probability Distributions

In this section, we introduce the information divergence geometry of probability distributions developed by Csiszár in [10]. In particular, we reproduce three of his main results which led to the development of the alternating minimization framework. Information geometry of alternating minimization has been used in neural network theory [5, 8, 7] and in Bolzmann machine learning in particular [6]. Some special cases of this framework relating to the EM algorithm were later re-discovered in other contexts such as "free enegy" in statistical mechanics [31].

Let us describe the main setting. Let P, Q, R be probability distributions (PDs) on a measurable space $(\Omega, \mathcal{B}, \cdot)$, and let $\mathcal{E}$ be the manifold of all such PDs. In the discussion that follows, we will not refer to the space specifically, Ω will either be finite or the Euclidean space with the Borel σ-algebra. Also, we assume the existence of the Radon-Nikodym derivatives or the appropriate absolute continuity conditions as the case requires.

Let us start with an abstract measure definition of divergence:

$$D(P||Q) = \begin{cases} \int \log(p_Q) dP & \text{if } P \ll Q \\ +\infty & \text{if } P \not\ll Q \end{cases} \tag{4.1}$$

In particular, it can be written in the familiar form below when the pdf exists

$$D(P||Q) = \int p(x) \log \frac{p(x)}{q(x)} dx. \tag{4.2}$$

The divergence is non-negative [44]

$$D(P||Q) \geq 0, \quad D(P||Q) = 0 \iff P = Q. \tag{4.3}$$

Yet, it is not a metric and the ε-balls

$$S(R,\varepsilon) = \{P|D(P||R) < \varepsilon\} \tag{4.4}$$

do not even define a basis for a topology.

In order to generate a geometry in which I-divergence plays a role similar to that in Euclidean spaces, we will use the following development due to csiszar[10]. First, let us define a projection with respect to the I-divergence in a similar way to that in Euclidean spaces. Let P, Q, R be PDs in $\mathcal{E}$. Q is called the *I-projection* of P on $\mathcal{E}$ if

$$D(Q||R) = \min_{P\in\mathcal{E}} D(P||R). \tag{4.5}$$

The first result concerns the existence of I-divergence:

Theorem 1 *[10] If the convex set $\mathcal{E}$ of PDs is closed in the topology of the variation distance, $|P-Q| = \int |p_R - q_R| dR$, then each R with $S(R,\infty) \cap \mathcal{E} \neq \emptyset$ has an I-projection on $\mathcal{E}$.*

The closure with respect to the variation distance will suffice for establishing the existence of I-projection for our problem. The next lemma shows how one can use analogs of connected line segments in a PD simplex.

Lemma 2 *[10] If $D(P||Q)$ and $D(Q||R)$ are finite, the "segment joining P and Q" does not intersect the I-sphere $S(R,\varepsilon)$ with radius $\varepsilon = D(Q||R)$, i.e., $D(P_\alpha||R) \geq D(Q||R)$ for each P_α*

$$P_\alpha = \alpha P + (1-\alpha)Q, \quad 0 \leq \alpha \leq 1, \tag{4.6}$$

iff

$$\int \log q_R dP \geq D(Q||R). \tag{4.7}$$

If

$$Q = \alpha P + (1 - \alpha)P', \tag{4.8}$$

then $D(Q||R) < \infty$ *implies* $D(P||R) < \infty$ *and the segment joining* P *and* P' *does not intersect* $S(R, \varepsilon)$ *(with* $\varepsilon = D(Q||R)$*) iff*

$$\int \log q_R dP = D(Q||R). \tag{4.9}$$

The equation (4.8) defines an algebraic inner point of $\mathcal{E}$:

Definition 1 *Q is an algebraic inner point of $\mathcal{E}$ if for all $P \in \mathcal{E}$, there exists $0 < \alpha < 1$ and $P' \in \mathcal{E}$ such that*

$$Q = \alpha P + (1 - \alpha)P'. \tag{4.10}$$

The notion of an algebraic inner point will be used in the conditions for the main theorem.

With a formal definition of a connecting line segment, we are ready to state the main result of the information geometric development, a "Pythagorean" theorem where I-divergence plays the role of squared Euclidean distance. By establishing the conditions under which the Pythagorean relation holds for PDs, Theorem 2 will allow us to bound the I-divergences of I-projections by virtue of the fact that "the hypotenuse is greater than either of the perpendicular sides".

Theorem 2 *[10] A PD $Q \in \mathcal{E} \cap S(R, \infty)$ is the I-projection of R on the convex set $\mathcal{E}$ of PDs iff every $P \in \mathcal{E} \cap S(R, \infty)$ satisfies (4.7) or, equivalently, iff*

$$D(P||R) \geq D(P||Q) + D(Q||R) \quad \forall P \in \mathcal{E}. \tag{4.11}$$

If the I-projection Q is an algebraic inner point of $\mathcal{E}$ then $\mathcal{E} \subset S(R, \infty)$ and (4.7) and (4.11) hold with the equality.

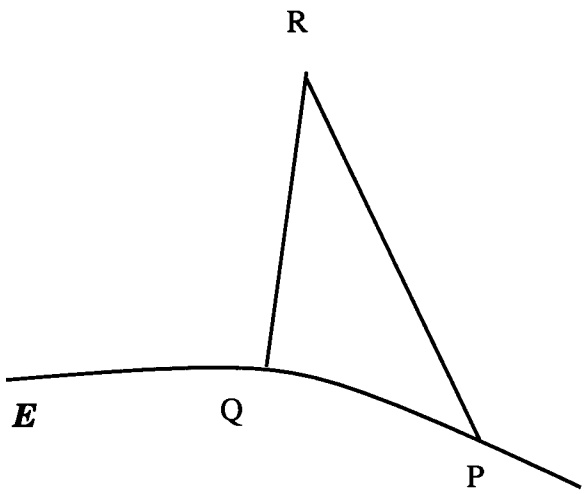

Figure 4.1: Pythagorean equation in a PD simplex

Theorem 2 defines the extent of geometrical results we will need for the topology preservation problem. Specifically, it provides us with a condition defined in terms of the notion of an algebraic inner point for PDs so that the following Pythagorean equation holds:

$$D(P||R) = D(P||Q) + D(Q||R), \quad \forall P \in \mathcal{E}. \tag{4.12}$$

See Figure 4.1. The ability to work with projections on probability manifolds and to compare I-divergences as if they were Euclidean distances allows the development of a framework in which we can analyse topology preservation in its relation to alternating minimization algorithms.

4.3 Alternating Minimization and Generalized EM Algorithms

The classical setting of maximum likelihood from partial observations leads to the iterative EM-algorithm [11]. In this section, we derive EM-like generalized

iterative minimization algorithms by using the geometry of information divergences developed in the previous section.

Let $\mathcal{S}$ be the manifold of probability distributions on two random variables X and C; $\mathcal{S} = \{P(X, C)\}$. Let us start by defining two sub-manifolds of $\mathcal{S}$. First manifold is the set of probability distributions realizable by the model family parameterized by Θ:

$$\mathcal{M} = \{Q_\Theta \in \mathcal{S} | \Theta \subset I\!R^p\} \tag{4.13}$$

The second manifold is the set of probability distributions that are consistent with the observed data:

$$\mathcal{D} = \{P \in \mathcal{S} | \sum_c P(X, C = c) = \delta(X - x)\} \tag{4.14}$$

The maximum likelihood problem of inferring C and Θ from observed X can equivalently be posed as finding the pair of distributions from two manifolds which are closest in terms of information divergence. A global solution to this optimization problem is usually intractable, however there exists a suboptimal iteration in the form of alternating projections back and forth between the two manifolds which will produce a sequence of non-increasing information divergences. Since the divergence is bounded from below, convergence to a local minimum (which, in practice, usually corresponds to a useful parameter setting) is guaranteed. It has been shown that in this framework EM algorithm corresponds to two projections: (i) projection from the model manifold to the data manifold (Expectation step), (ii) projection from the data manifold to the model manifold (Maximization step).

Thus, the search for the most likely parameter and missing values is reduced to a search for distributions in $\mathcal{M}$ and $\mathcal{D}$ which are closest in information divergence:

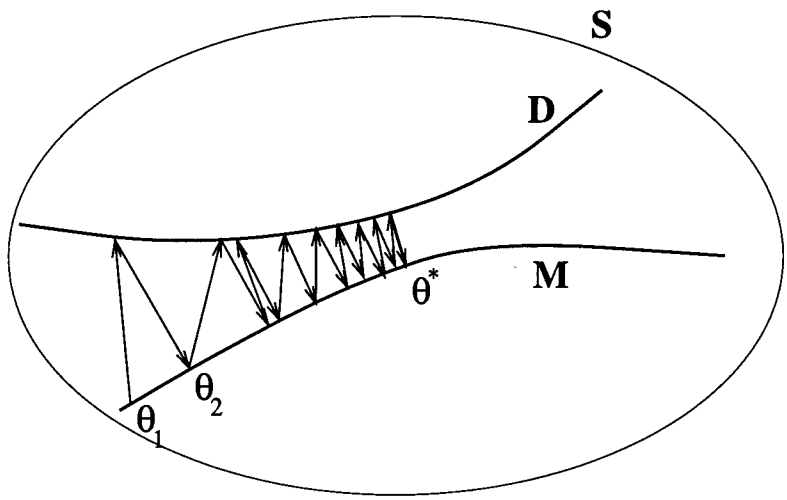

Figure 4.2: EM algorithm in information geometry.

$$D(\mathcal{D}||\mathcal{M}) = \min_{P\in\mathcal{D}} \min_{Q_\Theta\in\mathcal{M}} D(P||Q_\Theta). \tag{4.15}$$

Since this optimization usually proves to be intractable, a suboptimal iterative procedure described in the sequel is used. In particular, the usefulness of a given model Q_Θ is measured in terms of its distance to the data manifold, $\mathcal{D}$. This is determined by the following minimization which also defines the I-projection of Q_Θ on $\mathcal{D}$:

$$D(\mathcal{D}||Q_\Theta) = \min_{P\in\mathcal{D}} D(P||Q_\Theta). \tag{4.16}$$

Definition 2 *P^* is the I-projection of Q on $\mathcal{D}$ if*

$$P^* = \arg\min_{P\in\mathcal{D}} D(P||Q). \tag{4.17}$$

Thus, the alternating minimization search can be described in terms of the following two steps:

Algorithm 2 *Alternating Minimization*

0. Pick an initial model from $\mathcal{M}$ with parameter $\Theta^{(1)}$, $Q_{\Theta^{(1)}}$. Set $i = 1$.

1. I-project $Q_{\Theta^{(i)}}$ onto $\mathcal{D}$ to find $P^{(i)}$ via

$$D(P^{(i)}||Q_{\Theta^{(i)}}) = \min_{P \in \mathcal{D}} D(P||Q_{\Theta^{(i)}}). \tag{4.18}$$

2. Find the optimum model for $P^{(i)}$, $Q_{\Theta^{(i+1)}}$ in $\mathcal{M}$ via

$$D(P^{(i)}||Q_{\Theta^{(i+1)}}) = \min_{Q_\Theta \in \mathcal{M}} D(P^{(i)}||Q_\Theta). \tag{4.19}$$

3. Check convergence, i.e., *if $D(P^{(i)}||Q_{\Theta^{(i+1)}}) < \varepsilon$. If not, set $i = i + 1$, goto 1.*

This procedure is proved in [9] to produce sequences of distributions for which

$$D(\mathcal{D}||Q_{\Theta^{(i+1)}}) \leq D(\mathcal{D}||Q_{\Theta^{(i)}}) \tag{4.20}$$

hold, *i.e.*, a sequence of continuously improved models. It is also proved in [9] that this algorithm is equivalent to the EM-algorithm with (4.18) forming the E-step and (4.19) forming the M-step.

In the alternating minimization algorithm, convergence is still guaranteed under the following relaxation of step (2):

Algorithm 3 *Generalized Alternating Minimization*

0. Pick an initial model from $\mathcal{M}$ with parameter $\Theta^{(1)}$, $Q_{\Theta^{(1)}}$. Set $i = 1$.

1. I-project $Q_{\Theta^{(i)}}$ onto $\mathcal{D}$ to find $P^{(i)}$ via

$$D(P^{(i)}||Q_{\Theta^{(i)}}) = \min_{P \in \mathcal{D}} D(P||Q_{\Theta^{(i)}}). \tag{4.21}$$

2. *Find an improved model* $Q_{\Theta^{(i+1)}}$ *in* $\mathcal{M}$ *via*

$$D(P^{(i)}||Q_{\Theta^{(i+1)}}) \leq D(P^{(i)}||Q_{\Theta^{(i)}}) \tag{4.22}$$

3. *Check convergence,* i.e., *if* $D(P^{(i)}||Q_{\Theta^{(i+1)}}) < \varepsilon$. *If not, set* $i = i + 1$, *goto 1.*

Therefore, full M-step minimization (maximization in EM) need not be carried out, and an improved parameter setting will suffice for convergence. Such algorithms include the EM algorithm as a special case and are called Generalized EM (GEM) algorithms [27].

4.4 Information Geometry of Topology Preservation

Let us now recall the missing information problem introduced in Section (3.3). The data for the reference environment and the data for the secondary environment are unlabeled, therefore a "stereo", class to class matching is not possible. The unobserved class identities for correct transformation constitute the missing information. EM completes this missing information for the secondary environment in an arbitrary manner, converging to the closest fixed point defined by the secondary environment data regardless of the class structure of the reference environment. In order to estimate the correct transformation, in the previous chapter we resorted to constraining the VQ estimation with the topology of the reference environment. Similarly, in the information geometry framework, topology preservation constraint supplements the EM algorithm with information about the structure of the posterior class distribution. The goal is to restrain

the optimization towards a region of the probability manifold where the local minima are closer to the desired transformation. In this section, we show how, under the assumption of topology preservation, convergence to the correct (as defined by the topology preserving model assumption) PD can be achieved. We will make use of the information geometry tools developed in the preceding section, particularly a version of the Pythagorean theorem which will form the basis of our argument.

First, recall from the previous chapter how we defined the data manifold:

$$\mathcal{D} = \{P(X,C) = p(C|X)\delta(X-x)|\sum_{k=1}^{K} p(k|x) = 1\} \tag{4.23}$$

The posterior pmf, $p(C|X)$, as in the previous chapter, will be our primary tool for enforcing the topology preservation constraint. In its current form, it is dependent directly on the data, therefore we invoke Lemma 1 again to get the same probabilistic description of the topology in which the approximation posterior pmf is dependent on the data only through the code vector indices.

Recall that the topology of the reference environment is captured in the posterior class pmfs:

$$p_\Lambda(\cdot|k) \equiv p_\Lambda(\cdot|\lambda_k) = \frac{\phi(|\lambda_{(\cdot)} - \lambda_k|)}{\sum_{i=1}^{K}\phi(|\lambda_{(\cdot)} - \lambda_i|)} \quad k = 1,\ldots,K \tag{4.24}$$

And, same as in Chapter 3, the assumption of topology preservation can be written as

$$D(p_\Theta(\cdot|k), p_\Lambda(\cdot|k)) \leq \varepsilon \quad \forall k = 1,\ldots,K, \quad \varepsilon > 0. \tag{4.25}$$

That is, given the codebook index of the closest vector, I-divergences of the class posteriors of the reference and the secondary environments are assumed to be within ε-divergence proximity.

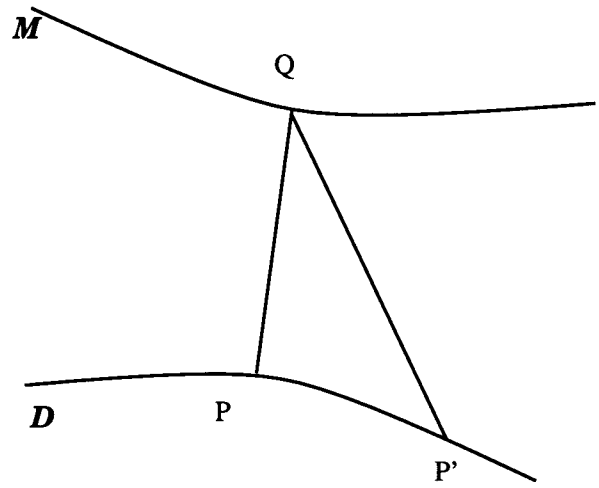

Figure 4.3: Pythagorean theorem in alternating minimization

The Pythagorean theorem we need to establish involves the model PD Q_θ in $\mathcal{M}$, P, the I-projection of Q_θ on $\mathcal{D}$, and the topology preserving PD P' in $\mathcal{D}$ as shown in Figure 4.3. The right triangle thus formed with $Q_\Theta P$ and PP' as the perpendicular sides and $Q_\Theta P'$ as the hypotenuse will serve as the basic picture in the development. Next, we observe that such a Pythagorean theorem holds for the VQ/kernel density framework for the robust transformation estimation problem and it can be used to introduce topology preservation in information geometry.

4.4.1 Pythagorean Theorem of Topology Preservation

First, note that $\mathcal{D}$ is the manifold of pmfs on K points, data determining the observed variable, X, and it is a linear family.

$$\mathcal{D} = \{P(X,C) = p(C|X=x) | \sum_{k=1}^{K} p(k|x) = 1\} \tag{4.26}$$

Therefore, the set of algebraic inner points of the set of pmfs $\mathcal{D} = \{(p_1,\ldots,p_K) | \sum p_i = 1, p_i \geq 0\}$ is $\mathcal{D}^* = \{(p_1,\ldots,p_K) | \sum p_i = 1, p_i > 0\}$. The set of algebraic points of

the manifold of finite probability mass functions are those pmfs with all nonzero masses. Since $\mathcal{D}$ is a linear family, the I-projection of the model manifold distribution $\mathcal{M}$ on $\mathcal{D}$ is in $\mathcal{D}^*$, i.e., that a pmf with at least one zero mass can never be the I-projection. Therefore, The Pythagorean equation holds for P, Q_Θ and any algebraic inner point $P' \in \mathcal{D}$.

$$D(P||Q_\Theta) = D(P||P') + D(P'||Q_\Theta) \quad \forall P' \in \mathcal{D}^*. \tag{4.27}$$

4.4.2 Topology Preserving Alternating Minimization

Now that we have a right triangle to work with, we are ready to introduce the optimization algorithm. The main idea is to bound the sequence of I-divergences in the alternating minimization algorithm by the sequence of I-divergences between the topology preserving data PD's and model PD's using the Pythagorean theorem. Let P^* and Q^*_Θ be the PD's which minimize the I-divergence between $\mathcal{D}$ and $\mathcal{M}$ as shown in Figure 4.4:

$$D(\mathcal{D}||\mathcal{M}) = D(P^*||Q^*_\Theta) = \min_{P\in\mathcal{D}} \min_{Q_\Theta\in\mathcal{M}} D(P||Q_\Theta) \tag{4.28}$$

and let P' be the data manifold PD required by topology preservation,

$$D(P'||P^*) \leq \varepsilon, \tag{4.29}$$

where $\varepsilon > 0$. Now, we can write the Pythagorean equation for these three PDs,

$$D(P^*||Q^*_\Theta) + D(P'||P^*) = D(P'||Q^*_\Theta), \tag{4.30}$$

which implies

$$D(P^*||Q^*_\Theta) - D(P'||Q^*_\Theta) \leq \varepsilon. \tag{4.31}$$

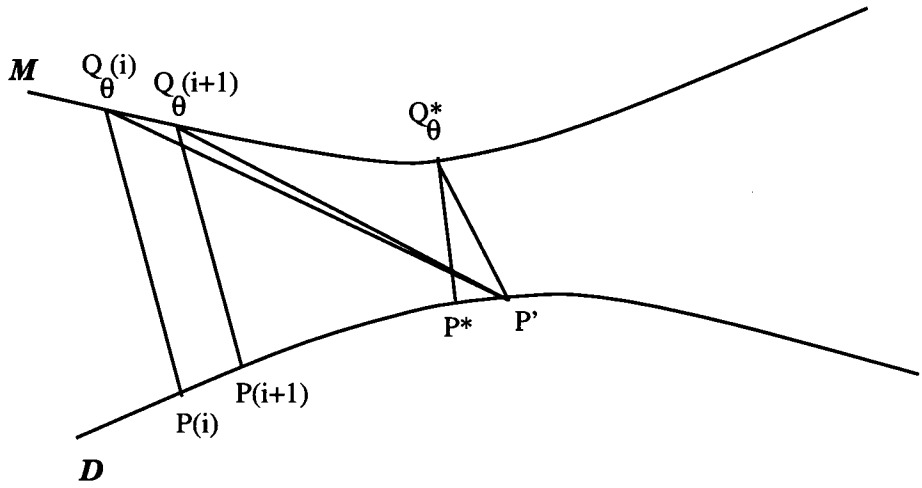

Figure 4.4: Information geometry of topology preserving alternating minimization

Consider the i-th step of the (generalized) alternating minimization algorithm. Let the I-projection of $Q_\Theta^{(i)}$ on $\mathcal{D}$ be $P^{(i)}$:

$$D(P^{(i)}||Q_{\Theta^{(i)}}) = \min_{P\in\mathcal{D}} D(P||Q_{\Theta^{(i)}}). \tag{4.32}$$

By the Pythagorean theorem, the following holds for all I-projections of the alternating minimization estimates

$$D(P^{(i)}||Q_\Theta^{(i)}) + D(P'||P^{(i)}) = D(P'||Q_\Theta^{(i)}). \tag{4.33}$$

Therefore, for all i,

$$D(P^{(i)}||Q_\Theta^{(i)}) \leq D(P'||Q_\Theta^{(i)}). \tag{4.34}$$

Simply put, hypotenuse is longer than the perpendicular sides. With such a bound on the I-projection divergences, we have, thus, proven the convergence of the following algorithm:

Algorithm 4 *Topology Preserving Alternating Minimization*

0. *Pick an initial model from* $\mathcal{M}$ *with parameter* $\Theta^{(1)}$, $Q_{\Theta^{(1)}}$. *Set* $i = 1$.

1. *Use the topology preserving PD* P' *to compute the bound on* $Q_{\Theta^{(i)}}$*'s divergence from* $\mathcal{D}$

$$D(P'||Q_{\Theta}^{(i)}) \geq D(P^{(i)}||Q_{\Theta}^{(i)}) \tag{4.35}$$

2. *Find an improved model,* $Q_{\Theta^{(i+1)}}$ *in* $\mathcal{M}$ *via*

$$D(P'||Q_{\Theta}^{(i+1)}) \leq D(P'||Q_{\Theta}^{(i)}) \tag{4.36}$$

which is a bound for the I-projection from $Q_{\Theta}^{(i+1)}$*:*

$$D(P'||Q_{\Theta}^{(i+1)}) \geq D(P^{(i+1)}||Q_{\Theta^{(i+1)}}) \tag{4.37}$$

3. *Check convergence,* i.e., *if* $D(P'||Q_{\Theta^{(i+1)}}) < \delta$. *If not, set* $i = i + 1$, *goto 1.*

Theorem 3 *Algorithm 4 produces the non-increasing sequence* $\{D(P'||Q_{\Theta}^{(i+1)})\}$ *which bounds the alternating minimization sequence* $\{D(P^{(i+1)}||Q_{\Theta^{(i+1)}})\}$, *and converges to within* ε *of the minimum I-divergence* $D(P^*||Q_{\Theta}^*)$.

By using the Pythagorean identity and the topology preservation assumption, we have obtained a topology-preserving alternating minimization algorithm which produces sequences of PDs in $\mathcal{M}$ and $\mathcal{D}$ with decreasing information-divergences and that converge to a local minimum of the I-divergence.

In the argument above, we proved that under the topology preservation assumption, minimization of the I-divergence between the topology preserving class posterior distribution and the model manifold resulted in an algorithm which converges to a local minimum of the I-divergence between the model manifold and the data manifold. Under the assumptions on the environment distributions, it is easy to see that there will be $K!$ permutations of the class indices and

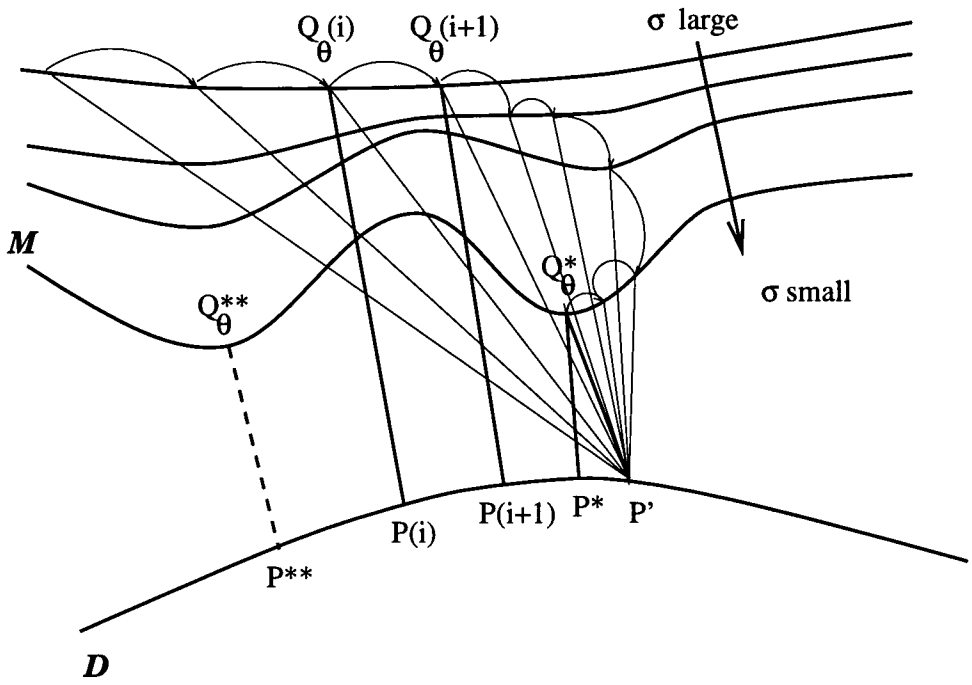

Figure 4.5: Topology preserving alternating minimization with multiple local minima

therefore that many local minima of the I-divergence between the two manifolds each corresponding to a re-naming of the codevector indices. Only one of the $K!$ labelings will provide the exact desired transformation and the fewer deviations from this permutation, the closer the topology of the secondary environment will be to the topology of the reference environment. The goal of topology preservation is to capture, as an invariant of both environments' data, the neighborhood structure of the classes and enforce that structure to drive the normally local algorithm to the desired local minimum, or a close neighbor.

The avoidance of spurious local minima and convergence to a more desirable local minimum is accomplished by varying the σ during optimization (See Figure 4.5). Recall that σ is fixed parameter of the VQ/Gaussian kernel density.

Therefore, changing σ corresponds to having a set of model families changing in complexity.

During topology preserving alternating minimization, we start with a model manifold which essentially only includes the uniform class pmf. As the optimization progresses, σ is reduced and the parameter updates occur by projections to families of models increasing in complexity. The parameters are globally positioned in the vicinity of topology preserving PDs early in the optimization while the I-divergence is close to a convex function, and fine tuning occurs later as local statistics are used for updates. This is a version of simulated annealing for the I-divergences on probability manifolds. Future work involves using ideas from coding to describe the complexity of the model families to match them to minimum description lengths warranted by the data to come up with σ variation schedules similar to the one derived in Chapter 3 via bias-variance trade-off.

4.5 Incremental Alternating Minimization Algorithms, TPAVQ

Alternating minimization of information divergences extends GEM in a way that allows optimization by any available means, opening the possibility of incremental optimization techniques. In this section, we develop those ideas to generate incremental, ("on-line") optimization algorithms that allow us to connect the information divergence-based view elaborated in this chapter with the constrained distortion minimization view of the preceding chapter. The result is that the incremental stochastic gradient descent algorithms for both optimization criteria turn out to be identical.

We will now give a sketch of how incremental EM-algorithms can be derived from the alternating minimization framework.

Let us reproduce Equation (??) from the proof of Lemma ??

$$D(P||Q_\Theta) = \sum_k \tilde{P}(k|X) \log \frac{\tilde{P}(k|X)}{Q_\Theta(k, X = x)} \tag{4.38}$$

and note that it can be decomposed into

$$D(P||Q_\Theta) = -E_{\tilde{P}}[\log Q_\Theta(k, X = x)] - H(\tilde{P}) \tag{4.39}$$

where the first term is the negative of likelihood and the second term is the entropy of the class distribution. Neal and Hinton have re-introduced (4.39) as their objective function in analogy with the "free energy" function in statistical physics [31].

From (4.39), it is readily observed that the minimization is with respect to the two variables, $\tilde{P}$ and Θ, corresponding to the E and M steps, respectively, and that the minimization with respect to Θ (M step) is not affected by the entropy of $\tilde{P}$.

Thus, we can simply look at incremental stochastic gradient algorithms which, during the M-step, maximize (EM) or improve (GEM) the likelihood integrated by the class posterior. The topology preservation requires that

$$\tilde{P}(j|X) = p_\Lambda^{\sigma^2(t)}(j|w_\Theta(X)), \quad \forall j = 1, \ldots, K. \tag{4.40}$$

where $w_\Theta(X)$ is the index of the codevector in Θ closest to X as in Chapter 3. Therefore,

$$\begin{aligned} \min_\Theta \left(-E_{\tilde{P}}[\log Q_\Theta(k, X_t = x_t)]\right) &= \max_\Theta \sum_k p_\Lambda^{\sigma^2(t)}(k|w_\Theta(x_t)) \log Q_\Theta(k, X_t = x_t) \\ &= \max_\Theta \sum_k p_\Lambda^{\sigma^2(t)}(k|w_\Theta(x_t)) \log \left(e^{-\frac{(x_t-\theta_i)^2}{2\sigma^2}}\right) \end{aligned}$$

$$= \max_{\Theta} -\sum_{k} p_{\Lambda}^{\sigma^2(t)}(k|w_{\Theta}(x_t))\frac{(x_t-\theta_i)^2}{2\sigma^2} \quad (4.41)$$

The stochastic gradient is given by:

$$\frac{\partial}{\partial\theta_i}\left(-E_{\tilde{P}}[\log Q_{\Theta}(k, X_t = x_t)]\right) = -p_{\Lambda}^{\sigma^2(t)}(k|w_{\Theta}(x_t))\frac{(x_t-\theta_i)}{\sigma^2} \quad (4.42)$$

resulting in the algorithm:

Algorithm 5 *Topology Preserving Alternating Minimization:*

0. Set the initial codebook to that of the reference environment.

$$\Theta^{(0)} = \Lambda \quad (4.43)$$

1. Update with the stochastic gradient at the current smoothness $\sigma(t)$

$$\theta^{(t+1)} = \theta^{(t)} - \eta(t)p_{\Lambda}^{\sigma^2(t)}(j|w_{\Theta}(X_t))(X_t - \theta_j). \quad (4.44)$$

2. Repeat 1 until convergence as $\sigma^2(t) \to 0$ *and* $\eta(t) \to 0$

But this is equivalent to the Algorithm 1. Therefore, we have established the equivalence of stochastic gradient descent algorithms with topology preserving adaptive VQ and topology preserving alternating minimization. The two frameworks produce similar algorithms for topology preservation yet each provides a distinct look which leads to a convergence proof in the case of alternating minimization.

Chapter 5

Robust Speech Recognition Experiments with TPAVQ

In this chapter, we describe three spech recognition experiments with telephone speech. Firstly, let us make a brief review of speech recognition over telephone lines. We need to distinguish between land-line and wireless speech. In land-line speech, handset is the most important factor due to still wide-spread use of carbon button handsets which tend to have sharply varying non-linear responses. Electret handsets tend to resemble clean, studio speech and generally work well with models trained with clean speech. Carbon button handset speech, on the other hand, degrade the performance of speech recognizers immensely with models trained with clean speech.

We report results on three tasks on three different corpora: Continuous Speech Recognition Corpus (CSR, a stereo database, part of the Wall Street Journal), the Spoken Speed Dial (SSD) Corpus and the TI Cellular Corpus. On the CSR corpus, we investigate the capstral match after normalization for a variety of techniques. The task on the SSD corpus is ten-digit recognition. In addition, we run two tasks on TI Cellular Corpus speaker independent digit

recognition and speaker dependent name recognition.

5.1 Cepstral Normalization on the CSR corpus

5.1.1 Description of the Corpus

The Continuous Speech Recognition (CSR) database is a subset of the Wall Street Journal (WSJ) database, and consists of stereo recordings of read speech from WSJ. It is made up of 10 sentences from 30 speakers, half female, half male, recorded simultaneously (stereo) with the Sennheiser HMD-414, a high-quality, head-mounted, close-talking microphone (CLSTLK) and the Crown PZM6FS, an omnidirectional, desktop microphone (PZM6FS). The training data consists of 20 speakers, in stereo form for algorithms that require stereo data. The testing data are divided into two; first part consisting of 5 speakers is used to adjust algorithm parameters as a development set, and the second part as the final evaluation set, on the basis of which all the algorithms are compared. The stereo form of the testing data is used to generate the scattergrams and compute the scattergram-based deviation ratios.

5.1.2 Task and Results

The task of cepstral normalization can best be understood from Figure 5.1 where the second cepstral coefficients computed from speech recorded by the CLSTLK microphone and by the PZM6FS microphone are plotted frame by frame as a scattergram. The second cepstral coefficient has been shown to be the dimension with the most discriminating power, therefore affects recognition the most. If the two microphones had recorded identical speech, the second cepstral coefficients

would have aligned on the $y = x$ line. Note that there is a significant offset from the $y = x$ line, resulting in very different locations in the acoustical space for the plotted frames. A HMM trained with the data recorded by the CLSTLK microphone will lose performance considerably when tested with data recorded by the PZM6FS microphone.

In Figures 5.2 and 5.3, we show the improvement in cepstral alignment by RASTA and CMN, respectively. Even though, the global offset has been taken out, note that the slope of the line deviates from 1 and therefore many frames at the margins cannot be aligned properly. Contrast this with the results of the TPAVQ algorithm shown in Figure 5.4 in which the data line is observed to be aligned with $y = x$. TPAVQ can translate and rotate the feature vectors, therefore is able to correct in a much more affective manner.

The ratio of the sum of differences of points from the line $y = x$ for a given algorithm to the identical sum for the baseline defines the *adjusted deviation ratio*, a numerical description of the scattergrams. In the table 5.1.2, we show adjusted deviation ratios for the three algorithms on the second and third cepstral coefficients. The ability of TPAVQ to align the cepstra is easily observable.

Technique	$DR(c_2)$	$DR(c_3)$
Baseline	1.0	1.0
RASTA	0.92	0.13
CMN	0.99	0.11
TPAVQ	0.59	0.07

Table 5.1: Adjusted deviation ratios for the CSR corpus

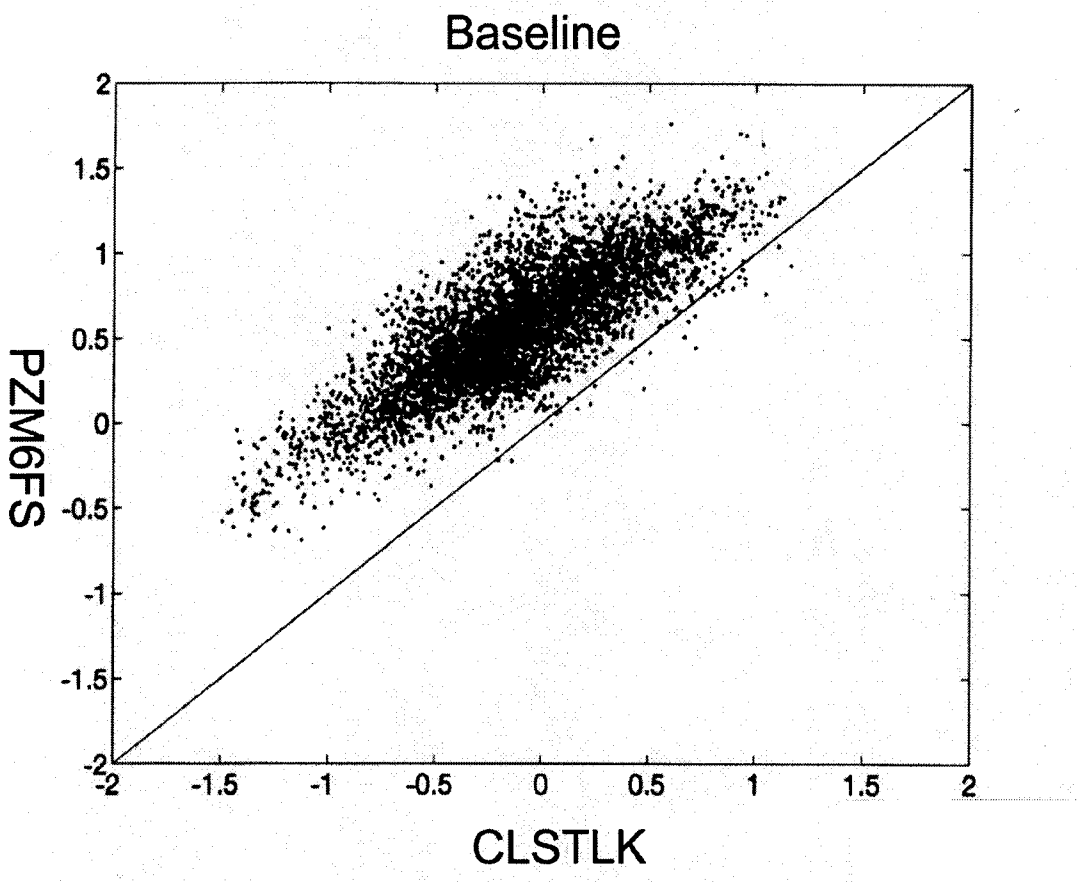

Figure 5.1: Scattergram, baseline

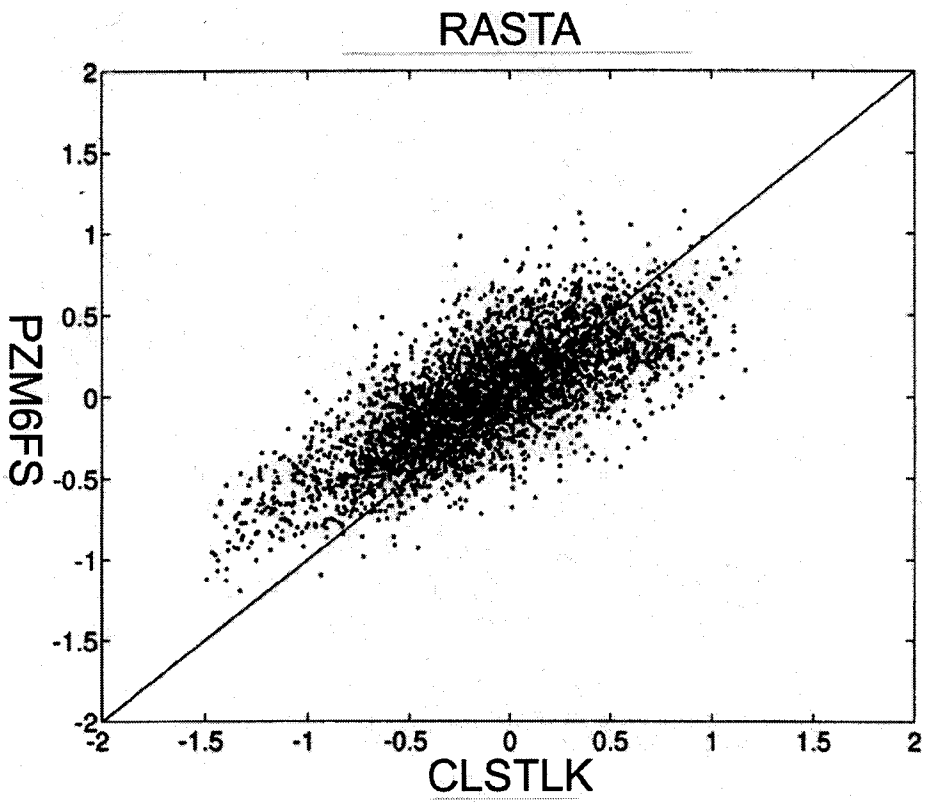

Figure 5.2: Scattergram, RASTA

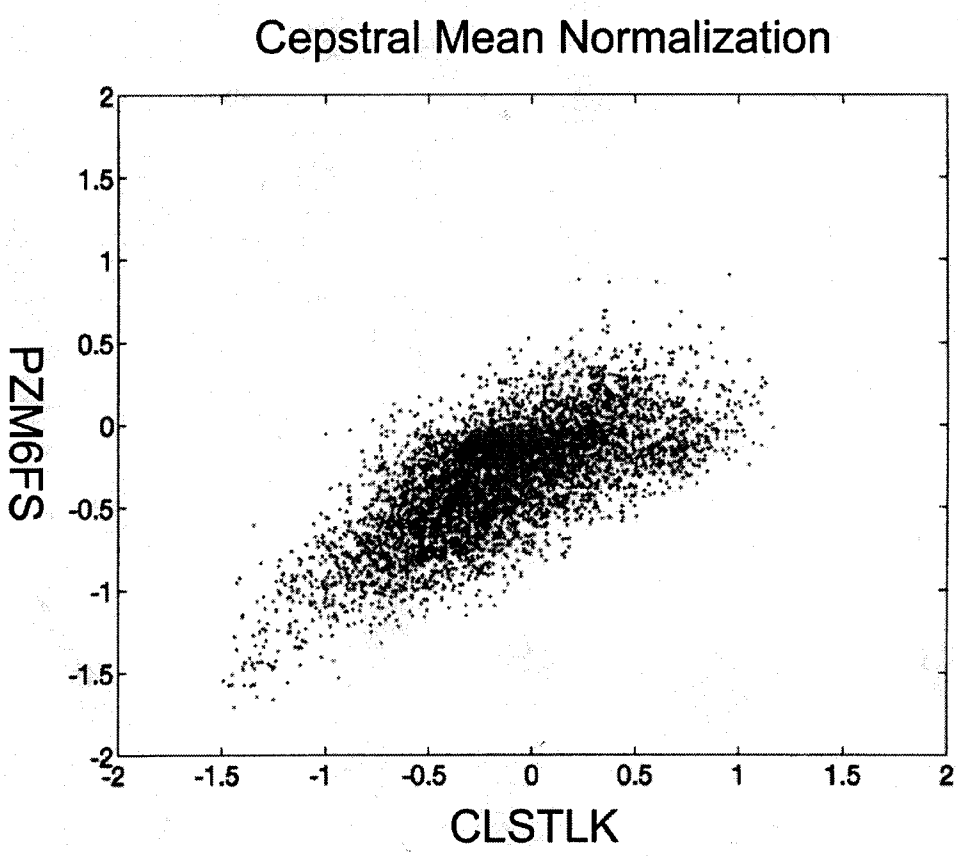

Figure 5.3: Scattergram, CMN

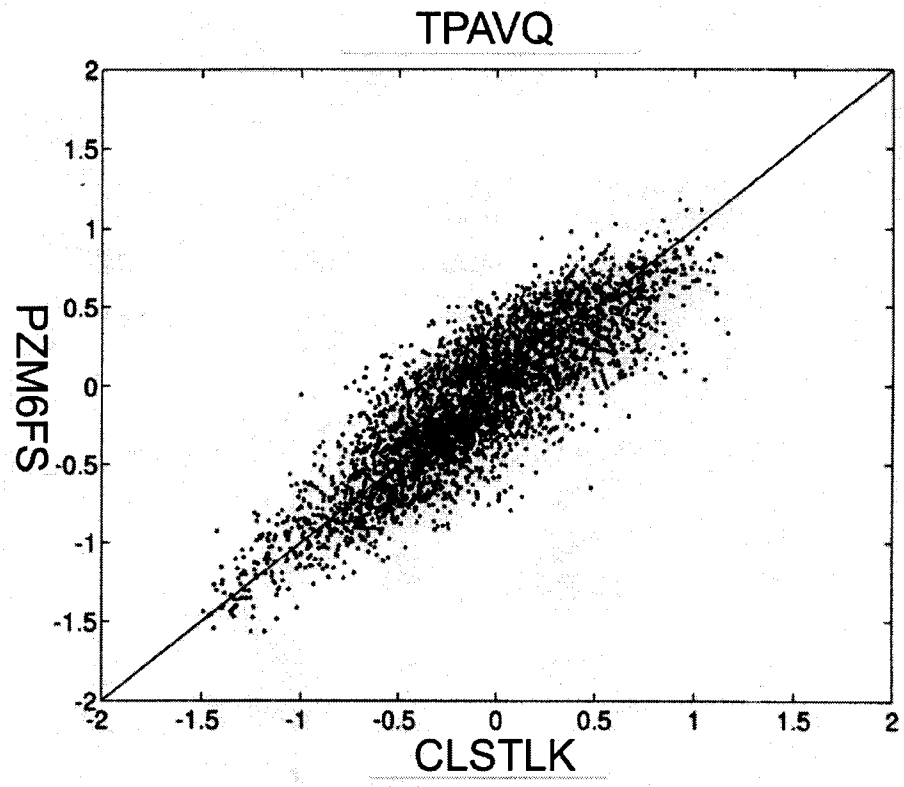

Figure 5.4: Scattergram, TPAVQ

5.2 Normalization of Acoustic Enrollment on the Spoken Speed Dial Corpus

5.2.1 The Spoken Speed Dial Corpus

The Spoken Speed Dial (SSD) corpus consists of names uttered by 10 speakers (5 male and 5 female) from name lists which contain 25 names for each speaker. 20 of the names in each list are unique to each speaker and 5 names are common to all lists. The recordings have been made with four different kinds of handsets:

1. Electret handset
2. Carbon button handset
3. Cordless telephone (electret handset)
4. Speakerphone

5.2.2 Acoustic Enrollment

In acoustic enrollment, speakers are asked to utter each name in their list three times:

file1: `Jerry Junkins`

file2: `Jerry Junkins`

file3: `Jerry Junkins`

The task is to use these three files to train a model for name recognition which will be tested with test utterances of the form:

file: `Call Jerry Junkins .`

The enrollment files are used to generate an HMM for the name for the particular speaker in the following way:

1. Mark the speech frames in file1
2. Compute the feature vectors for the marked frames
3. Generate an HMM using the duration and the acoustic feature vectors in file1: left-to-right HMM with one state per frame, transition probabilities fixed, and mean vectors of the states set to feture vectors of the frame.
4. Update the acoustic feature vector parameters in the HMM using the remaining file2 and file3 by forward-backward.

Once the models for all the names in the list for a particular speaker has been trained, the most likely name according to the HMM scores for the test sequences is declared as the recognized name (and dialed in the real-world application).

5.2.3 Results

Due to its small vocabulary and the abundance of sufficiently distinct names, this task is not too difficult for most speakers under matched training and testing conditions. We show the baseline results in Table 5.2.3. The performance varies greatly depending on the speaker. In particular, S09 tried to confuse the recognizer by varying the duration of the utterances during enrollment and test. S10 is a very soft spoken speaker.

In Table 5.2.3, we show the results obtained with RASTA. RASTA does not seem to improve the performance although the increase in overall error rate from 5.1% to 5.8% may not be statistically significant. In Table 5.2.3, we show the

Spkr.	c/c	c/e	c/cl	e/c	e/e	e/cl	cl/c	cl/e	cl/cl	x/x
S01	1.3	1.3	0.9	1.3	1.3	1.3	4.9	5.8	0.9	2.1
S02	0.3	1.3	1.3	1.0	2.7	4.7	12.3	8.4	7.3	4.4
S03	0.4	2.7	6.0	0.9	3.3	4.7	2.2	8.0	4.7	3.6
S04	0.0	1.3	0.0	0.0	1.0	0.0	3.7	4.3	2.7	1.6
S05	1.3	3.6	3.3	5.0	0.4	12.7	3.7	0.9	2.0	3.4
S06	0.0	0.7	0.7	4.4	3.7	3.3	0.4	1.0	0.7	1.7
S07	0.4	0.7	1.3	3.6	0.0	1.8	1.3	0.0	1.3	1.1
S08	0.0	0.0	0.9	0.0	0.0	4.9	2.7	1.3	0.9	1.2
S09	11.0	17.3	26.7	22.3	10.2	28.7	32.0	20.4	17.3	20.4
S10	2.7	4.4	26.7	3.6	1.8	23.1	10.7	9.3	34.2	12.9
all	1.9	3.1	6.9	4.5	2.4	8.4	8.0	5.6	7.6	5.1

Table 5.2: Baseline system performance for acoustic enrollment

results for TPAVQ. There are drastic improvements in some speakers, and the overall error rate improves by 24%, from 5.1% to 3.9%.

Table 5.2.3 shows the results for TPAVQ for the 8 speakers who are more representative of the average in terms of speech effort, and who did not try to confuse the recognizer. The improvement is more striking in this case; 46% improvement from an error rate of 2.4% down to 1.3%.

Spkr.	c/c	c/e	c/cl	e/c	e/e	e/cl	cl/c	cl/e	cl/cl	x/x
S01	2.2	3.1	0.4	2.7	2.2	0.0	4.9	6.2	0.4	2.5
S02	0.7	1.8	2.0	2.0	1.3	1.3	14.3	11.6	7.3	4.9
S03	0.0	2.0	6.0	0.9	3.7	4.0	7.1	7.7	4.0	3.9
S04	0.0	1.0	0.0	0.3	1.0	0.0	5.7	3.3	2.7	1.7
S05	2.3	0.9	2.7	14.0	0.4	6.7	14.7	0.4	2.0	5.6
S06	0.0	0.7	0.7	3.6	1.3	1.3	0.4	0.7	0.7	1.0
S07	0.4	0.0	1.3	5.8	0.0	2.2	3.6	0.3	1.3	1.5
S08	0.0	0.0	0.9	0.0	0.0	2.2	4.0	0.4	0.9	0.9
S09	14.7	12.9	21.3	23.0	11.6	26.7	39.7	24.4	23.3	22.2
S10	4.4	7.6	32.4	4.4	1.8	25.3	12.0	8.9	30.7	14.2
all	2.7	2.7	7.1	6.2	2.2	7.1	11.6	6.0	7.5	5.8

Table 5.3: RASTA system performance for acoustic enrollment

5.3 Normalization of Cellular Telephone Speech

5.3.1 Description of the Corpus

Results are presented on continuous digit recognition and voice dialing in the TI Cellular Corpus. The corpus consists of data collected over cellular channels by using two types of microphones: a hand-held, close talking microphone and a hands-free, visor mounted microphone together with land-line collected speech data. The land-line and hand-held microphone parts of the corpus are mostly clean telephone speech comparable in quality to VAA corpus. The hands-free microphone part of the corpus, however, is significantly noisier than the rest.

Spkr.	c/c	c/e	c/cl	e/c	e/e	e/cl	cl/c	cl/e	cl/cl	x/x
S01	1.3	0.9	0.4	1.3	1.3	0.9	2.2	1.8	0.4	1.2
S02	0.7	1.3	1.3	0.3	1.3	4.0	3.0	1.8	0.0	1.6
S03	0.4	2.3	4.7	0.4	2.7	4.0	1.8	2.0	4.0	2.5
S04	0.3	1.7	0.0	0.0	1.0	0.0	0.0	2.0	0.0	0.5
S05	1.0	2.7	4.7	1.7	0.4	10.0	2.7	0.9	2.7	3.0
S06	0.0	0.7	0.7	0.0	0.7	0.7	0.0	1.0	0.7	0.5
S07	0.4	0.0	0.9	3.1	0.0	1.8	0.9	0.0	1.3	0.9
S08	0.0	0.0	0.9	0.0	0.0	4.0	2.7	0.0	0.9	0.9
S09	11.0	16.0	26.0	21.0	9.8	28.0	31.7	19.1	17.3	20.0
S10	4.0	3.1	10.0	3.1	1.8	6.0	11.6	6.2	22.0	7.5
all	1.9	2.9	5.0	3.0	1.9	5.9	5.7	3.5	4.9	3.9

Table 5.4: TPAVQ system performance for acoustic enrollment

5.3.2 Speaker-independent Digit Recognition

The first experiment investigates the effectiveness of the compensation algorithm in normalizing the TI cellular speaker independent digit recognition data to improve recognition using models trained on the VAA1 corpus. The codebooks were trained on data sets in the TI cellular and VAA corpora disjoint from the model training and testing sets for which the recognition results were obtained. The results in Table 5.3.2 indicate that the normalization does not disturb the reference environment (VAA) appreciably, nor the land line and hand held environments which are close to the VAA. There is a 43% decrease in the error of the hands free microphone.

Spkr.	c/c	c/e	c/cl	e/c	e/e	e/cl	cl/c	cl/e	cl/cl	x/x
1-8	0.5	1.5	1.8	1.9	1.5	4.2	3.9	3.7	2.6	2.4
1-8	0.5	1.1	1.7	0.9	0.9	3.1	1.6	1.1	1.2	1.3

Table 5.5: Speakers 1-8,TPAVQ system performance for acoustic enrollment

environment	*no. of* utt.'s	*error baseline*	*error w/ CMN*	*error w/ TPAVQ*
VAA2	1390	4.1	4.1	4.2
land line	282	4.5	4.4	4.7
hand held	283	6.0	6.0	6.1
hands free	246	23.8	17.8	13.6

Table 5.6: Results of the speaker independent digit recognition experiment.

5.3.3 Speaker-dependent Voice Calling

A similar experiment was carried out on the speaker dependent portion of the TI cellular database. Table 5.3.3 summarizes the average results for 30 speakers each uttering 10 names in a voice calling application in which the land-line is the reference environment. The reference and clean environments are again not disturbed appreciably and there is a 36% decrease in the error of the hands free microphone.

TPAVQ decreases the word error for continuous ten digit recognition of cellular hands free microphone speech with land line trained models from 23.8% to 13.6% and the speaker dependent voice calling sentence error from 16.5% to 10.6% in the TI cellular corpus.

environment	*no. of* utt.'s	*error baseline*	*error w/ CMN*	*error w/ TPAVQ*
land line	696	3.4	3.4	3.7
hand held	688	4.7	4.8	5.4
hands free	650	16.5	13.4	10.6

Table 5.7: Results of the speaker dependent voice calling experiment.

In all three experiments, TPAVQ has outperformed feature vector filtering via RASTA and CMN decisively. It is important to remember unlike other codeword dependent methods that are superior to filtering, TPAVQ (similar to RASTA or CMN) does not require stereo data.

Chapter 6

Conclusion and Future Work

This dissertation addresses the problem of environmental robustness in current speech recognition technology. Starting with a non-parametric model of the effects of the environment on speech distributions, we proposed a mathematical framework based on adaptive vector quantization or, as we have shown, equivalently on the alternating minimization algorithm for environment compensation. Specifically, the TPAVQ algorithm was proposed as a means of preserving the global topology of the speech distributions under distortion.

In this chapter, we summarize our conclusions and findings based on experiments with various tasks. We also review the major contributions of the work and present several suggestions for future work.

6.1 Summary of Results

The most basic result of the dissertation is the viability of the topology preservation hypothesis. In addition to making intuitive sense, we showed that it can be exploited to derive algorithms which do not need stereo data from the secondary environments and which normalize the noisy and distorted speech successfully.

We also demonstrated that topology preservation lends itself to statistical analysis when formulated in terms of preserving the conditional class distributions in an information geometry. Posed in this way, its convergence was shown in relation to an associated alternating minimization algorithm.

Our experimental results demonstrate that the TPAVQ technique proposed in this thesis produces significant improvements in recognition accuracy. As predicted, the improvements are much higher than those produced by simple filtering of the features such as mean normalization or RASTA. This is due to the finer local compensation of the feature space which, as experimentally demonstrated, is distorted in ways which are hard to describe in a parametric manner.

6.2 Contributions

We may sum up the major contributions of this dissertation as follows:

- We have proposed using the topology of the distribution of the feature vectors as an invariant under the distortion model transformation and developed a framework which incorporates the preservation of toplogy in a convenient manner.

- We have demonstrated how the topology preservation can be included as a constraint in distortion minimization in VQ and how the minimization can be carried out with a stochastic approximation algorithm.

- We have introduced an information geometry framework for topology preservation where the class posteriors of the testing environment are constrained

to be within ε I-divergence proximity of the class posteriors of the reference environment.

- In the information geometry framework, we have developed a constrained alternating minimization algorithm which preserves topology. We have proved its local convergence and commented on its global convergence.

- We have empirically demonstrated the effectiveness of the approach in three different tasks which incorporate different acoustical conditions. The most notable was how the algorithm performed with cellular telephone speech which presented the most ardous challenge.

6.3 Future Work

One thing that is missing from the topology preservation discussion is a true measure of how global order is preserved. The mathematical frameworks we have introduced describe the topology preservation as constrained minimization or class probability distribution invariance, however they stop short of quantifying how good the preservation has been accomplished. This seems like a difficult problem right now, as the notion of order in high dimensional spaces is hard to quantify mathematically. It is conceivable that a more sophisticated measure then I-divergence of class posteriors, proposed in this thesis, can be derived.

In this thesis, we have presented the TPAVQ algorithm as capable of using both a priori information about likely environments and adaptation to fit new environments as more data become available during testing. There are many issues about adaptation in a real-world application such as the minimum duration of the speech transaction which whould make adaptation useful, intialization of

the neighborhood function before adaptation which may be made more peaked than in the off-line case due to the existence of a good starting point in terms of available environment codebooks, and overcoming problems such as extended periods of silences which may bias the codebook towards a better representation of the noise components of the environment. These problems are also left for future study both experimentally for real-world applications, and in theory for studying the viability of adaptation not just for environment compensation but also for applications such as speaker adaptation.

REFERENCES

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Ph.D. Thesis, Dept. of Electrical and Computer Engineering, CMU, Sept. 1990

[2] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.

[3] K. Aikawa, H. Singer, H. Kahawara, Y. Tohkura, "A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition", *ICASSP-93*, May 1993.

[4] L. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes", *Inequalities* 3:1-8, 1972.

[5] S. Amari, *Differential Geometrical Method in Statistics,* Springer Lecture Note in Statistics, 28, Springer, 1985.

[6] W. Byrne, "Alternating minimization and Boltzmann machine learning," *IEEE Trans. Neural Networks*, vol. 3, no. 4, pp. 612-620, 1992.

[7] S. Amari, "Information Geometry of the EM and em Algorithms for Neural Networks," *Neural Networks*, vol. 8, No. 9, pp. 1379-1408, 1995.

[8] S. Amari, "The EM Algorithms and Information Geometry in Neural Network Learning," *Neural Computation*, vol. 7, pp. 13-18, 1994.

[9] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," in *Statistics and decisions, Supplementary issue,*

No. 1, (E. Dedewicz *et. al.*,eds.), pp. 205-237, Munich, Oldenburg Verlag, 1984.

[10] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Annals of Probability*, vol. 3, no. 1, pp. 146-157, 1975.

[11] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *Jour. of the Royal Statistical Society B*, vol. 39, pp.1-38, 1977.

[12] J. Flanagan, J. Johston, R. Zahn, G. Elko, "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms",*JASA*, vol. 78, pp.1508-1518, Nov. 1985.

[13] A. Gersho, R. Gray, "Vector Quantization and Signal Compression", Kluwer, 1992.

[14] M.F. Gales, "Model-Based Techniques for Noise Robust Speech Recognition", Ph.D. Thesis, Engineering Department, Cambridge University, Sept. 1995.

[15] O. Ghitza, "Auditory Nerve Representaion as a Front-end for Speech Recognition in a Noisy Environment", *Computer, Speech and language* vol. 1, pp.109-130, 1986.

[16] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication,* **16**, 1995 pp. 261-291.

[17] H. Hermansky, N. Morgan, A. Bayya, P. Kohn, "COmpensation for the Effect of teh Communication Channel in Auditory-Like Analysis of Speech",

Proc. of the Second European Conf. on Speech Comm. and Tech., Sept. 1991.

[18] H. Hermansky, N. Morgan, H. Hirsch, "Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing", *ICASSP-93*, pp. II-83-86, April 1993.

[19] F. Jelinek, "Continuous Speech Recognition by Statistical Methods", *Proceedings of the IEEE*, 64(4): pp. 532-556, April 1976.

[20] B. Juang, "Speech Recognition in Adverse Environments," *Computer Speech and Language*, Vol. 5, pp. 275-294, 1991

[21] T. Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences, Springer-Verlag, Berlin 1995

[22] C.J. Leggeter, P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer, Speech and Language*, vol.9, pp.171-185.

[23] Y. Linde, A. Buzo, R.M. Gray,"An Algorithm for Vector Quantizer Design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84-95, January 1980.

[24] F.H. Liu, A. Acero, R.H. Stern, "Efficient Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering," *ICASSP-92*, pp. I-257-260, March 1992

[25] F.H. Liu, R.H. Stern, A. Acero, P.J. Moreno, "Environment Normalization for Robust Speech Recognition using Direct Cepstral Comparison," *ICASSP-94*, pp. 61-64, April 1994

[26] F.H. Liu, "Environmental Adaptation for Robust Speech Recognition," Ph. D. Thesis, ECE Department, CMU, July 1994

[27] X.L. Meng, D.B. Rubin, "Recent extensions of the EM algorithm (with discussion)", inJ.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (editors), *Bayesian Statistics 4*, Oxford, Clarendon Press, 1992

[28] P.J. Moreno, B. Raj, E. Gouvêa, R.M. Stern, "Multivariate Gaussian Based Cepstral Normalization", *ICASSP-95*

[29] P.J. Moreno,"Speech Recognition in Noisy Environments", Ph.D. Thesis, Dept. of Electrical and Computer Engineering, CMU, April 1996

[30] L. Neumeyer, M. Weintraub, "Probabilitic Optimum Filtering for Robust Speech Recognition", *ICASSP-94*, pp. I-417-420, May 1994.

[31] R. M. Neal and G. E. Hinton, "A new view of the EM algorithm that justifies incremental and other variants", submitted to *Biometrika*, 1993.

[32] L. Rabiner, B. Juang, "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine*, 3(1) pp. 4-16, Jan. 1986.

[33] J. Rissanen, *Stochastic complexity in statistical inquiry*, World Scientific, Teaneck, N.J. 1989.

[34] R. Schwartz, T. Anastakos, F. Kubala, J. Makhoul, L. Nguyen, G. Zavaliagkos, "Comparative Experiments on Large Vocabulary Speech Recognition," *Proc. ARPA Human Language Technology Workshop,* Plainsboro, New Jersey, March 1993.

[35] S. Seneff, "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing", *Journal of Phonetics*, vol. 16, pp.55-76, January 1988.

[36] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, 1986.

[37] T. Sullivan, R. Stern, "Multi-Microphone Correlation-Based Processing for Robust Speech Recogntion", *ICASSP-93*, April 1993.

[38] A. Viterbi, "Error Bounds for Convolutional codes and an Asymptotically Optimum Decoding Algorithm", *IEEE Trans. on Info. Theory*, vol. IT-13, pp. 260-269, 1967.

[39] P.C. Woodland, M.J.F. Gales, D. Dye, V. Valtchev, "The HTK Large Vocabulary Recognition System for the 1995 ARPA H3 Task", *Proc. of the 1996 ARPA Speech Recognition Workshop*, Feb. 1996.

[40] T. Adalı, X. Liu, and M. K. Sönmez, "Conditional distribution learning with neural networks and its application to channel equalization," *IEEE Trans. Signal Processing*, vol. 45, no. 4, pp. 1051-1064, Apr. 1997.

[41] D. M. Titterington, "Comments on 'application of the conditional population-mixture model to image segmentation'," *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 6, No. 5, pp. 656-658, September 1984.

[42] J. Rissanen, "Minimax entropy estimation of models for vector processes," *System Identification*, pp. 97-119, 1987.

[43] D. M. Titterington, A. F. M. Smith, and U. E. Markov, *Statistical analysis of finite mixture distributions.* New York: John Wiley, 1985.

[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc. 1991.

[45] S. Haykin, *Neural Networks: A Comprehensive Foundation.* New York: Macmillan College Publishing Company, 1994.

[46] J. L. Marroquin and F. Girosi, "Some extensions of the K-means algorithm for image segmentation and pattern classification," Technical Report, MIT Artificial Intelligence Laboratory, Jan. 1993.

[47] T. Adalı, M. K. Sönmez, and K. Patel, "On the dynamics of the LRE Algorithm: A distribution learning approach to adaptive equalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Detroit, MI, 1995, pp. 929-932.

[48] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, December 1974.

[49] J. Rissanen, "A Universal Prior for Integers and Estimation by Minimum Description Length," *The Annals of Statistics,* Vol. 11, No. 2, 1983.

[50] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, 4, pp. 1-52, 1992.

[51] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, Vol. 108, No. 2, pp. 620-630/171-190, May 1957.

[52] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlay, 1988.

[53] L. Xu and M. I. Jordan, "On convergence properties of the EM algorithm for Gaussian mixture," Technical Report, MIT Artificial Intelligence Laboratory, Jan. 1995.

[54] E. Weinstein, M. Feder, and A. V. Oppenheim, "Sequential algorithms for parameter estimation based on the Kullback-Leibler information measure," *IEEE Trans. Acou. Speech, and Signal Processing*, Vol. 38, No. 9, pp. 1652-1654, 1990.

[55] R. A. Redner and N. M. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.* , Vol. 26, pp.195-239, 1984.

[56] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Processing*, pp. 1361-1364, 1990.

[57] J. L. Marroquin, "Measure fields for function approximation," *IEEE Trans. Neural Nets.*, Vol. 6, No. 5, pp. 1081-1090, 1995.

[58] M. Wax and T. Kailath, "Detection of Signals by Information Theoretic Criteria," *IEEE Trans. Acoust. Speech, Signal Processing,* Vol. 33, No. 2, April 1985.

[59] L. Kullback, and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics 22*, pp. 79-86, 1951.

[60] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," *NATO ASI Series, vol. F68, Neurocomputing*, pp. 227-236.