1-1-2009

Optimal generative and discriminative acoustic model training for speech recognition

Neil Joshi
Ryerson University

OPTIMAL GENERATIVE AND DISCRIMINATIVE ACOUSTIC MODEL TRAINING FOR SPEECH RECOGNITION

by

Neil Joshi

MS Electrical Engineering. University of Massachusetts, 2002

BSc Electrical Engineering, University of Calgary, 1996

A dissertation

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2009

I hereby declare that I am the sole author of this dissertation.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this disseration by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholary research.

Abstract

Neil Joshi
OPTIMAL GENERATIVE AND DISCRIMINATIVE ACOUSTIC MODEL TRAINING FOR SPEECH RECOGNITION
Doctor of Philosophy
Electrical and Computer Engineering
Ryerson University
Toronto, ON, Canada
2009

The focus of this dissertation is to derive and demonstrate effective stochastic models for the speech recognition problem. Acoustic modeling for speech recognition typically involves representing the speech process within stochastic models. Modeling this high frequency time series effectively is a fundamental problem.

This dissertation devises an objective function that relates the true speech distribution to its estimate. It is shown that through optimizing this function the speech process time series can be modeled without loss of information.

The thesis proposes two such models that are developed to optimize the devised objective function. The first an acoustic model formulated for the speech with noise problem. The second a discriminately trained model consisting of optimal discriminant ML estimators.

The first, a combination of recognizers that through a simple system fusion, combines multiple speech processes at the decision level. This is a stochastic modeling method devised to combine a parameterized spectral missing data, MD, theory based and a cepstral based speech process using a coupled hidden variable topology. In using a fused coupled hidden Markov model, HMM, topology, an optimal acoustic model is proposed that is inherently more robust than single pro-

cess models under noisy conditions. The theoretical capability of this model is tested under both stationary and non stationary noise conditions. Under these test conditions the fused model has greater recognition accuracies than those of single process models.

The second, formulated with a methodology that segments the acoustic space appropriately for discriminately trained models that optimize the devised objective function. This acoustic space is modeled with discriminant ML estimators formed with optimal decision boundaries using the large margin, support vector machine, SVM, learning method. These discriminately trained models maximize the entropy of the observation space and thereby are capable to model the speech process without loss. This is demonstrated experimentally with frame level classification error rates that are $\sim \leq 3\%$.

Dedication

John von Neumann

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

science is built up of facts, as a house is built of stones; but an accumulation of facts is no more a science than a heap of stones is a house. — Henri Poincare (1905)

Great advances have been made in speech recognition research over the past few decades. From its early incarnation in the 1950's, when the discovery was made using statistical classification methods for speech patterns it seemed very likely that the speech recognition problem would be solved in its entirety within a short period of time thereafter. This obviously was not to be. The high variability of speech, the different dialects, tones, and accents of the spoken language and the interfering noise from the environment have prevented this realization.

In the 1950s various researchers tried to exploit the fundamental ideals of acoustic-phonetics. The initial effort in 1952 by members of Bell-Labs[26] (Davis, Biddulph and Balashek) resulted in a system for isolated digit recognition relying on measuring spectral resonances during the vowel region of each digit. This work spurred numerous research efforts based mostly on the use of filter analyzers for the measurement of the spectral information for pattern isolation and recognition. Japanese researchers made great advances in the 1960s with hardware based filter bank spectral analyzers for phoneme and vowel recognition. This decade also saw the development of the first methods to address the nonuniformity of time scales in speech events[50][51][61]. The decade to follow, the 1970s, witnessed several advances in speech recognition research. Namely, the development and demonstration of reproducible and viable isolated word recognition techniques using pattern recognition and dynamic programming methods. This decade also

saw the rise of great research houses for speech recognition research such as those at IBM and Bell Labs.

With isolated word recognition techniques established, the 1980s saw these techniques extended to tackle the problem of continuous connected word recognition. This decade coincided with a shift of research focus from template based to stochastic models. Hidden Markov models, HMMs, were introduced to the speech community in this decade and this method was rapidly adopted[31][58] by the speech community. Of notable mention, this decade also gave rise to the reintroduction of neural networks, NNs, to the speech recognition problem[49][74]. Though NNs were first investigated in the 1950s, it was deemed too problematic to use at that time. Large continuous speech recognition systems and databases that were developed by DARPA, CMU, BBN, Lincoln Labs, SRI, MIT and Bell Labs became widely available for the research community. This availability and the advancements made in research seeded the necessary conditions for the rapid progress that was seen in speech recognition research in the following decades.

Researchers made great advances in robust speech recognition research during the 1990s. Robust, in the sense of speech recognition under noisy conditions. The foundations for noise adverse and speaker independent speech recognition were established during this period. Such works included RASTA[40], HMM decomposition[71], maximum likelihood linear regression[32], Parallel Model Combination[33] and Missing Data[19], MD, techniques. Furthermore, the availability of new standardized noise corrupt speech corpora, such as the Noisex 92 and Aurora corpuses aided in the proliferation of this research topic. Great advances were made throughout the 1980s in computing resources. Subsequently, the acceleration of this technology in the 90s together with the enormous advances in computer networking led to tremendous progress in the decade to come in speech and language research which had, at this time, become to be known as Human Language Technology, HLT. This decade saw the introduction of audio visual speech recognition as well as, due to the increasing connectedness and globalization of the world, the advancement of machine translation and multilingual speech recognition. The increased popularity and services offered on the World Wide Web, WWW, spurred interest into the research of HLT for the indexing of information and information retrieval that included part of speech tagging, noun phrase, NP, deciphering of text and speech.

Building on the advances over the past few decades with stochastic speech recognition, researchers focused on strengthening the models through the use of discriminative techniques and combining classifiers. The 2000s also saw the rise of statistical learning theory applied to speech recognition and HLT due in part to

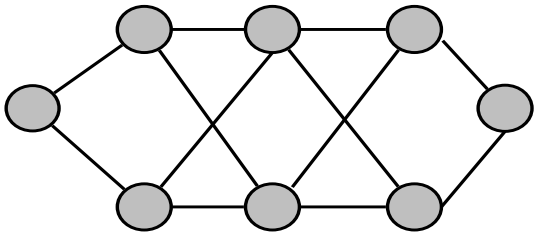


Figure 1.1: Trivial Speech Recognition Network

the commoditization of computing resources and the availability of computation power that would permit its realization in this decade.

The speech process is highly variable and non stationary in nature. Due to this, the speech phenomenon, as outlined in the previous passage, is predominately researched as a stochastic process. Under this premise the objective is to determine the best word sequence, $\mathscr{W} = \{W_1, W_2, \ldots, W_n\}$, given a set of observations, $\mathscr{O} = \{O_1, O_2, \ldots, O_n\}$,

$$\arg\max_{W} P(\mathbf{W}|\mathbf{O}) \tag{1.1}$$

where, $\mathbf{W} : w_i \in \mathscr{W}$, $\mathbf{O} : o_i \in \mathscr{O}$ and $w_i : i \in \mathscr{I}$ and o_i are indexed elements of $\mathbf{W}$ and $\mathbf{O}$ respectively. With this structure a trellis is formed containing word nodes and edges as depicted in Figure 1.1. The observations that correspond to each node represent the probabilities of the network and Equation 1.1 is satisfied by *minimizing the cost* or equivalently by maximizing the likelihood, ML, of the negative log probabilities. In this ***naive***, construct the edges represent the probabilities of a given node in relation to an observation, or rather, for a specific edge, the ***posterior*** probability of a word, W, given an observation, O, and stochastic model, θ,

$$P(W|O,\theta) \tag{1.2}$$

These posteriors can be determined using differing methodologies. The predominate, conventional, technique is through the Bayesian view for determining pos-

teriors where,

$$P(W|O,\theta) = \frac{P(O|W,\theta)P(W,\theta)}{P(O|\theta)P(\theta)} = \frac{P(O|W,\theta)P(W|\theta)}{P(O|\theta)} = \frac{P(O|W,\theta)P(W)}{P(O)} \tag{1.3}$$

bases the posterior on the *likelihood*, the leftmost numerator factor, and priors, $P(W)$ and $P(O)$. This likelihood or *generative* model determines the most probable word model combination that may have generated any given observation. The optimal discriminative model, in contrast, attempts to optimize the problem to determine an, x^*, such that,

$$x^* = \min_x f(\mathbf{x_i}, \mathbf{x_{j\neq i}}) \quad \forall i \tag{1.4}$$

In this case, $\mathbf{x} \equiv \mathbf{O}$, where W is inferred from $\mathbf{x}\,\forall W_i$, $W_i \in \mathscr{W}$. Methods that satisfy Equation 1.4 are referred to as ***discriminative techniques*** and they include information theoretic[2], least squares[43] approaches as well as non parametric solutions[49].

The stochastic models of the speech process that form the network of Figure 1.1 creates the decision boundaries that ultimately permit the determination of a word sequence from a set of observations. These models are commonly formed either through generative methods involving density estimation or by discriminative techniques. To construct these stochastic models and to evaluate Equation 1.1, the observations that make up the speech process are transformed into a *parameterized* form or feature vectors, a format suitable for this pattern recognition task. Within this process the signal is sampled at a frequency greater than the Nyquist frequency and commonly passed though a bank of filter banks so as to expose the critical signal attributes that can be used to characterize the signal. As is described in[43][77], and is depicted in Figure 1.2 to complement the discussion, the continuous time signal is discretized with a sampling rate of, f_s, together with a sliding window of duration t_w and a parameterization period of, t_r. A common feature representation is the Mel log frequency, MF, or cepstral coefficient, MFCC, feature. In this case each sample that represents t_r of the signal is passed through

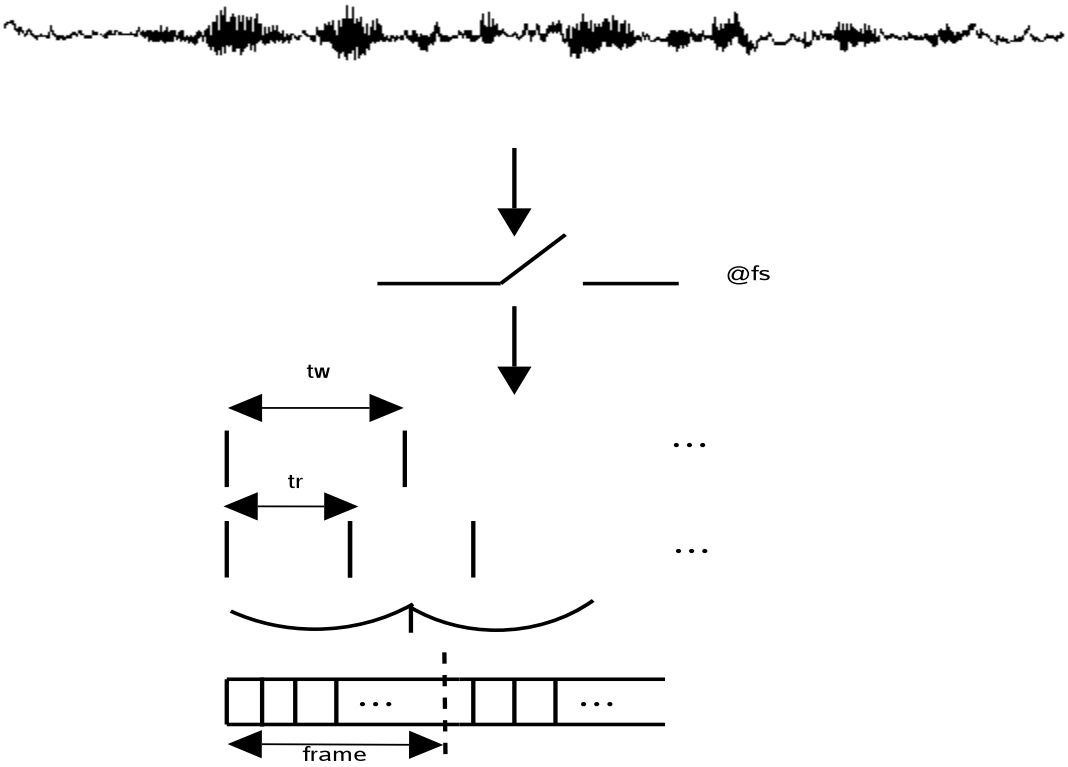


Figure 1.2: Parameterization of speech signal

filter banks that represents the spectrum. The log representation of the spectrum is then transformed back to the time domain to provide what may be referred to as the *spectrum of the log spectrum*. In other words, as the initial transformation provides the *spectrum* of the signal in the frequency domain, the inverse transformation of the log signal can be considered as the spectrum of this signal in the time domain. Each parameterized sample results in a feature vector or rather a speech frame.

As is described in the following subsection, the focus of this dissertation is in devising effective stochastic acoustic models for the speech recognition problem. These probabilistic models are devised with techniques that minimize the distance, or error, between the true speech stochastic model, θ, and its estimate. Therefore, it implies that the models presented within this thesis are optimal and are referred to as such.

1.1 Research focus

The nature of the body of work, and focus of this dissertation is in direct connotation to the derivation and the demonstration of optimal[1] stochastic models for the speech recognition problem. This entails both, the examination of the problem of speech in noise, Equation 1.1, and the use of a discriminative method for describing Equation 1.2. This closely follows the direction of current speech recognition research, as discussed previously in this chapter, with regards to addressing robust speech recognition and the investigation of non parametric, discriminative techniques for modeling the problem.

The motivation of this thesis is to effectively model the speech process for speech recognition. Here, optimal stochastic models are devised and developed. This is done first by defining, through information theoretic concepts, an expression that can represent the true speech, or observation, stochastic process. As is detailed in Chapter 2, an expression is formulated that represents the process in terms of a hidden variable stochastic model. It is shown that in maximizing this hidden variable expression, the stochastic model is capable of representing the observation process, or rather, the speech process without loss[2]. The expression, Equation 2.16,

$$H(O_n \,|\, O^{(n-1)}) \geq H(O_n \,|\, U_n)$$

expresses the entropy of the true observation speech process in terms of n random variables, ***rvs*** O, the term on the left of the inequality, and its estimate in terms of hidden variable topology ***rvs*** O and U. This describes the true speech observation at time, n, O_n given its previous *n-1* realizations, $O^{(n-1)}$, in relation to its estimate at time n, or the ML estimator. This objective function together with the manner from which it is arrived from provide the basis for this dissertation.

Specifically this thesis investigates,

1. *speech in noise:* the development of an optimal coupled stochastic model to combine two separate streams of features for robust speech recognition under adverse noise conditions. This work extends the missing data[19], MD, methodology to accommodate multiple sets or streams of observation features.

[1] optimal in the sense of minimizing the distance, or error, between the true speech stochastic model, θ, and its estimate

[2] Thoughout this thesis, in describing so-called *lossless* modeling, it is referred to in this sense

Speech recognition under noisy conditions is an open research problem. Though noise robust techniques such as cepstral mean normalization, CMN[1] and RASTA[40] have been successfully applied to the problem, the benefits of those approaches are usually realized under stationary noise conditions. In general, approaches to enhance speech recognition under noisy conditions can either attempt to remove/suppress the noise perturbations or to accommodate them within an adapted recognizer stochastic model. The later approach, as the HMM decomposition[71], the Parallel model combination, PMC[33] and similar techniques[7] have demonstrated significant results under non stationary conditions. Accommodating a noisy signal by adapting the stochastic model is promising, though these methods do require *apriori* knowledge of the noisy condition to be effective.

The missing data approach[20], in contrast, has been demonstrated to be effective for robust speech recognition under all noisy conditions. Here, speech recognition is performed using only the speech bearing, or reliable components of a noisy signal. As is presented in Section 3.4, a problem, arguably a very significant drawback, with missing data techniques is that it generally requires spectral based features[20][13][37][42][59]. Unlike past efforts to resolve this, the presented body of work devises an optimal coupled stochastic model to permit the use of cepstral based features within the missing data framework.

A novel optimal coupled model methodology is devised to combine classifiers, or separate streams of features within the missing data framework. In using information theoretic concepts to assess the dynamics or relationship between ***rvs*** in a hidden variable structure, potential coupled topologies[11][63][55] can be compared to determine the most appropriate structure to model the speech process. It is shown that in minimizing the objective function, Equation 3.32,

$$KL\left(p\left(O_{(1)},O_{(2)},\ldots,O_{(g)}\right)\parallel p\left(\hat{O}_{(1)},\hat{O}_{(2)},\ldots,\hat{O}_{(g)}\right)\right)=$$
$$-\int\ldots\int p\left(O_{(1)},O_{(2)},\ldots,O_{(g)}\right)\,ln\left(\frac{p\left(\hat{O}_{(1)},\hat{O}_{(2)},\ldots,\hat{O}_{(g)}\right)}{p\left(O_{(1)},O_{(2)},\ldots,O_{(g)}\right)}\right)\mathrm{d}O_{(1)}\mathrm{d}O_{(2)}\ldots\mathrm{d}O_{(g)}$$

an optimal stochastic model[3] can be devised to represent the speech process. Moreover, the resultant coupled probabilistic space representing missing

[3] see Equation 3.32, or Kullback-Leibler, KL, distance[48] with *random variables O*, where $O_{(i)}$ represents the observations i of g time series and $\hat{O}$ its estimate.

data and cepstral processes is capable of increasing the information content, or *capacity*[22] of the acoustic model. This is shown both theoretically and experimentally. Theoretically it is shown that the expected performance of combined coupled acoustic model should be greater than that of a spectral feature missing data model as well as a cepstral based model. Results from a series of recognition experiments under both stationary and non stationary noise conditions are empirically in agreement with the theoretical capability of the combined models.

2. *optimal ML estimators:* the formulation of speech stochastic model posteriors with discriminative learning methods. In furthering the thesis topic of devising acoustic models to model the speech process, the expression developed in Chapter 2, Equation 2.16, is refined using discriminative classification techniques. Specifically the large margin, or support vector machine[68], discriminative method. Acoustic modeling using discriminative learning methods presents a manner that may be more suitable to represent the speech process than traditional density estimation modeling methods. Such techniques adverts the ill posed problem that density estimation methods are to solve. As is shown in detail Section 4.1, the hidden variable stochastic model is capable of representing the speech observation process. Through segmenting the acoustic space in the manner that is described in that section (Section 4.1), the hidden variable construct can be represented in a way that is suitable for discriminative training methods. Here the observation process is shown to be able to be expressed in terms of maximum likelihood, ML, estimators. Presented in this work is a methodology to formulate and model optimal *discriminant* ML estimators to model the speech process.

   There have been several research efforts that have used discriminatively trained acoustic models. Most notably, neural network based methods[10][64][41]. Though these pioneering works have addressed modeling speech with NN classifiers they have been hindered by limitations that may due to the discriminative training method used. Such limitations include controlling the complexity, or generalization capability, of the model whilst maintaining a low classification error rate. Many differing discriminative learning methods can be applied to model the speech process with ML estimators. However, it will be reasoned[4], that large margin methods can overcome some of the perceived drawbacks that confront many of them.

[4] Section 4.2

This work presents a novel methodology to model the speech process using ML estimators that are discriminatively trained using the large margin technique. Unlike other support vector variants[34], that have researched support vector machine speech recognizers, this work formulates and defines a large margin method that is capable of representing the speech process without loss. Moreover, this is realized, unlike past efforts to discriminatively train acoustic models for speech, by forming models at the speech frame level. The devised optimal acoustic models are not only capable of representing the speech process without loss, but are also shown to maximize the entropy of the observation distribution. This is demonstrated experimentally with speech frame classification error rates $\sim\leq 3\%$.

1.2 Research Contributions

Problem :
To devise and develop effective stochastic models for modeling the speech process.
Dissertation Contributions :

- Development of objective function relating true observation distribution to an estimate in terms of the directly observable measurements and latent hidden variables. *Chapter 2.*
- Formulation and development of an optimal stochastic model for the speech with noise problem. Proposal of a combination of recognizers that through a simple system fusion, combines multiple speech processes at the decision level. This is a novel stochastic method devised to combine a parameterized spectral missing data, MD, theory based and a cepstral based speech process using a coupled hidden variable topology. In using a fused coupled hidden Markov model, HMM, topology, an optimal stochastic model is proposed that is inherently more robust than single process models under noisy conditions. *Chapter 3.*
- A novel analysis and comparison of the capabilities of coupled hidden variable topologies to model the speech process. *Chapter 3.*
- Through the maximization of the devised objective function, Equation 2.16, it is shown that the resultant optimal combined acoustic space contains

greater information content of the true observation distribution. Thus is capable of improved recognition accuracies. *Chapter 3.*

- Segmentation of the speech acoustic space in a manner that can represent the speech process effectively and can be modeled with discriminative learning methods. *Chapter 4.*
- Devising an optimal discriminant ML estimator to model the speech observation distribution. *Chapter 4.*

1.3 Organization of thesis

The organization of this thesis is as follows. Chapter 2 and its subsections provides substantial background in support of this dissertation. It comprises of introducing and describing speech stochastic acoustic models. Within this chapter the mapping of the speech recognition problem to that of Equation 1.2 is given as well as an in depth discussion of representing the speech problem in terms of Equation 1.3, generative, and Equation 1.4, discriminative methodologies. This encompasses both the derivation of the parameters of the models and methods to determine the minimum cost, Equation 1.1, of the speech network topology with the resultant models. Chapters 3 and 4 together describe the main methodologies of the thesis topic. Speech with noise is an open research topic.

Chapter 3 proposes a methodology based on missing data theory to perform effective noise robust speech recognition through combining classifiers. Under this premise the described methodology fuses two speech processes, two streams of features, at the pattern recognition stage. The combination of the two processes is presented as coupled time series problem and within the chapter the optimal fused model is proposed and demonstrated to be an effective method for determining the statistical properties of each stochastic model of the network. The resultant models are demonstrated through a series of experiments to achieve higher recognition accuracies than those of conventional and MD based recognizers under both stationary and non stationary noise conditions.

Chapter 4 is devoted to establishing a methodology to satisfy Equation 1.4 using support vector machines, SVMs, for speech recognition. This entails describing the problem in a manner that is appropriate for applying discriminative techniques while maintaining compatibility with well established recognition modeling techniques. Within this description, an approach is proposed to map the

acoustic space to a format that can be used to train the vector machine classifiers. With a method to train the classifiers established, the chapter proceeds to describe the derivation of speech stochastic model posterior probabilities from the constructed classifiers. The exceptional effectiveness of the method is furthermore demonstrated with experiments with a speech corpus. Chapter 5 concludes the thesis with a general discussion of the presented methodologies and offers insight into further directions that the current research could take.

1.4 Guide to the reader

This dissertation is on the topic of effectively modeling the speech process. It proposes two acoustic models that are capable of modeling this process effectively for the speech recognition problem.

Dissertation in five minutes

For the casual reader: Each chapter of this dissertation contains a section that identifies significant findings and summaries its content. These sections, read in their entirety, can provide the reader a good grasp of the proposed acoustic models. This "dissertation in five minutes" is found in the following sections: Section 2.3 (p.28), Section 3.8 (p.63), Section 4.5 (p.109).

For the eager reader

Chapter 2 is required reading. This chapter provides substantial background in support of this dissertation. Each of the subsequent chapters are self contained and may be read on its own. The proposed acoustic models are presented in this dissertation within 4 books. Chapter 3 contains the first two and Chapter 4 the final two. The first book of each chapter contains the methodologies for the proposed models. The second of the two, the supporting experiments for the formulated acoustic models.

Chapter 3 proposes an acoustic model for the speech with noise problem. A noise robust optimal acoustic model is formulated as a simple system fusion of two speech processes. The proposed model is demonstrated to be capable of higher recognition accuracies than single process models under both stationary and non stationary noise conditions.

Chapter 4 presents the proposed discriminant ML estimators. Though segmenting the acoustic space in a manner that captures the acoustic space of speech process, optimal discriminant ML estimators are formed. The resultant acoustic models are shown to be not only capable of effectively capturing the observation process, but also maximize the entropy of the observation distribution.

1.5 Commonly used notation and symbols

Throughout this dissertation the following symbols are commonly used and can be taken as such unless it is otherwise specified.

O	random variable representing an observation or input
U	random variable representing a [hidden] state
y	output variable
x	input variable
b	bias or result variable
Z	random variable representing an observation or input
$\mathbf{I}$	identity matrix
Σ	covariance matrix
$\mathscr{R}$	real number space
$\mathscr{I}$	integer number space
θ	variable representing a stochastic model
O_r	random variable representing observation reliable, speech bearing components
O_u	random variable representing observation unreliable, or noise, components
(g)	variable in brackets representing random process g, $g \in \mathscr{I}$

:	such that
$\Longleftrightarrow$	if and only if
$\Longrightarrow$	implies
$\vdash$	infers
$\perp$	statistical independence
$\perp\!\!\!\perp$	conditional independence
$\equiv$	equivalence

Furthermore, the mathematical notation used throughout this dissertation follows that with sets and or spaces defined between braces or brackets. An array of elements or a vector is defined with bold font variables. Generally vectors and matrices are presented with column vector notation unless it is otherwise specified. Therefore, a column vector of, n, real number elements, can be defined as $\mathbf{x} = [x_1, x_2, x_3, \ldots, x_n]^T$, where T is the transpose, or $x \in \mathscr{R}^n$. Similarly, a row vector of these elements within this space is expressed as $\mathbf{x}^T$.

Commonly found in this dissertation are lowercase and uppercase bold font variables that represent vectors and matrices respectively. Though for the most part each, x_i, of $\mathbf{x}$ is a scalar, it may *also* represent, at times, a multivariate in an effort to preserve a common form for clarity. Another vector notation used in this thesis is, $X^{(n)}$, that is equivalent to, $\mathbf{x}$, with elements, $x_i : i \in 1 \ldots n$. Such a notation permits clarity in the derivation and assessment of optimal[5] stochastic acoustic models to model the speech process.

The focus of this dissertation is to devise and develop effective stochastic probabilistic models to represent speech. As such, a majority of the variables used in this thesis are random variables, or ***rvs***. Speech itself is often considered as a random process composed of random variables O. The notation used to denote distributions is generally a tilde preceding a variable representing the distribution. One such example is the Gaussian or normal distribution, $\sim N(x\,|\,\mu,\sigma) = \frac{1}{\sqrt{2}\sigma} e^{\frac{(x-\mu)^2}{\sigma^2}}$, where x is a ***rv*** and μ, σ^2 the distribution mean and variance respectively.

Subscripts in this thesis are generally used to specify distinct incarnations of a random process or variable. As such, for a set of, $g, g \in \mathscr{I}$, random processes, $O_{(i)}$ represents the i^{th} random process. For a vector, $\mathbf{x}$, the i^{th} element of the vector is represented as x_i. Similarly for a column vector of n ***rvs***, $O^{(n)}$, O_n is the n^{th} such element. Probability distributions, or densities, may be represented in terms of parameters that define the distribution. Subscripts may be used to denote specific components of the distribution and to specify the ***rv*** its distribution represents. For a Gaussian mixture distribution containing, k, components that represents the distribution of a multivariate ***rv*** O_r for model, θ_i, $i \in \{1 \ldots h\}$, this is expressed as, $\sum_{l=1}^{k} \pi_{rl\theta_i} N(O_r\,|\,\mu_{rl\theta_i}, \Sigma_{rl\theta_i})$. Where $\pi_{rl\theta_i}$ is the l^{th} mixture weight, and $\mu_{rl\theta_i}, \Sigma_{rl\theta_i}$ are the mean and covariance respectively for the l^{th} mixture.

Some of the devised stochastic models in this thesis are illustrated in diagrams for clarity. Such illustrations represent probabilistic spaces and describe the statistical relationship between random variables, ***rvs***. The following convention is used throughout this thesis. Illustrated in Figure 1.3 is a probabilistic space $P(U_1,U_2) = P(U_1)P(U_2|U_1)$ for two ***rvs***, U_1 and U_2. Similarly, Figure 1.4 describes the probabilistic space $P(U_1,O_1)$. The relationship between the ***rvs*** is described within the connections (arrows) between them. This is described in full in Chapter 2. Unless it is otherwise noted, in illustrations such as these, a circle, or node, that proceeds another node connected with a red arrow indicates the rela-

[5] optimal in the sense of minimizing the information loss of a model; in other words, minimizing the distance between the true probabilistic distribution and its estimate

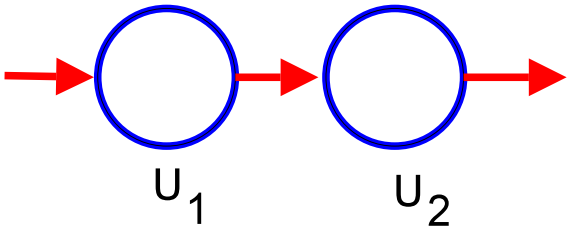


Figure 1.3: Illustration of probabilistic space containing ***rvs*** U

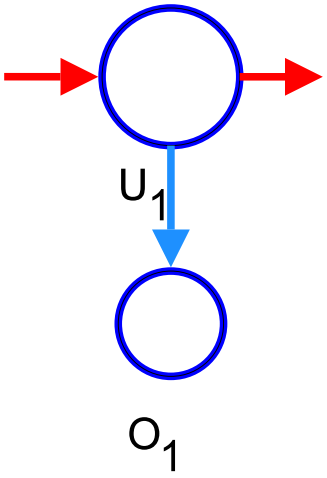


Figure 1.4: Illustration of probabilistic space containing ***rvs*** U and O

tionship between ***rvs*** U_i, $i \in \mathscr{I}$ and U_{i+1}. Similarly, a node that proceeds another node connected with a blue arrow indicates the relationship between ***rvs*** U and O. In this case the node at the arrowhead of the connection is a ***rv*** O and U the node at the tail.

Chapter 2

Acoustic Modeling

This chapter provides substantial background in support of this dissertation. Stochastic modeling of the speech process is fundamental to the speech recognition problem. This chapter describes stochastic modeling of the speech process in terms of both generative, Equation 1.3, and discriminative, Equation 1.4 methodologies. Through information theoretic concepts, expressions are devised that can be used to analyze the effectiveness of stochastic models to model the speech process.

Hidden variable stochastic models are introduced, namely the hidden Markov model, that consist of modeling speech in terms of directly observable variables, O, and latent, hidden, variables, U. Using the concepts and expressions presented near the beginning of this chapter, the capability of the hidden variable topology to represent the speech process is evaluated. In doing so, an objective function is developed that serves as a main motivation for this thesis.

Parameter training of these hidden variable models with the common expectation maximization, EM, is also detailed prior to introducing the stochastic modeling problem as a discriminative learning problem. Here, several discriminative based methods are described including Bayesian based techniques and methods that attempt to determine optimal decision boundaries that distinguish between distinct patterns, or classes of speech.

Together, the fundamental models presented, the concepts introduced and the objective function formulated provide the background and insight for the models formulated in this thesis.

The speech process is characteristically highly non stationary in nature. In order to effectively model this signal, the signal is transformed to a piecewise

short term spectral representation, the parameterization of speech as described in Figure 1.2, such that each sample, i, of the n observations, $O^{(n)}$, can be classified as stationary.

Subsequently, each observation sample, i, forms the observation vector,

$$O^{(n)} = [O_1, O_2, O_3, \ldots, O_n]^T \tag{2.1}$$

using column vector notation. Each sample, O_i, as well, is a multivariate so it consists of m coefficients accumulated while parameterizing the speech signal[1]. As such each likewise component, j, of each sample, i, can be grouped together to form vectors,

$$\begin{aligned} &Z_j^{(n)} = [O_{1j}, O_{2j}, O_{ij}, \ldots, O_{nj}]^T \\ &j \in \{1 \ldots m\} \end{aligned} \tag{2.2}$$

Insight into the relationship between the signals' observation measurements can be gained from expressing the parameterized signal in this form. Here, each $Z_j^{(n)}$ is a vector of random variables that represents successive measurements for a single parameter, or dimension, of the signal.

When evaluating the relationship between distributions, elements of information theoretic[22] concepts can provide a useful framework to assess the stochastic traits and interconnections. Just as the inner product of two vectors portrays the projection of one to another, or in other words, determines the minimum distance, the information theoretic concept of *mutual information*, the Kullback-Leibler, *KL*, distance between joint and independent distributions, *I()*, measures the similarity between probabilistic distributions. The former satisfies the Cauchy-Schwartz inequality, the later does not. Thus,

$$\begin{aligned} &KL(p(a)p(b) \parallel p(a,b)) = \\ &I(a,b) = -\iint p(a,b)\, ln\left(\frac{p(a)p(b)}{p(a,b)}\right) \mathrm{d}a\mathrm{d}b \end{aligned} \tag{2.3}$$

is a measure of similarity between **rv**s a and b with distributions *P(a)* and *P(b)* respectively and a joint space of *P(a,b)*.

As such, in examining each measurement, of $Z_j^{(n)}$, as a ***rv***, and using the concept of mutual information to analyze the relationship between each and every ***rv***,

[1] In other words, $O_i \in \mathscr{R}^m$.

i, of $Z_j^{(n)}$, Equation 2.3 can subsequently be rewritten as[2],

$$I(Z_{ji}, Z_{lk}) = -\iint p(Z_{ji}, Z_{lk})\, ln\left(\frac{p(Z_{ji})p(Z_{lk})}{p(Z_{ji}, Z_{lk})}\right) \mathrm{d}Z_{ji}\mathrm{d}Z_{lk} \tag{2.4}$$
$$\forall j$$
$$i,k \in \{1 \ldots n\}$$
$$l \in \{1 \ldots m\}$$

In assessing the measure of similarity between distributions, the resultant *KL* distances derived from the above equation are,

$$I(Z_{ji}, Z_{lk}) = \begin{cases} 0 & \text{, if } p\left(Z_{ji}, Z_{lk}\right) = p(Z_{ji})p(Z_{lk}) \\ > 0 & \text{, if } p\left(Z_{ji}, Z_{lk}\right) \neq p(Z_{ji})p(Z_{lk}) \\ H(Z_{ji}) & \text{, if } i == k \; and \; j == l \end{cases} \tag{2.5}$$

where, $H(Z_{ji})$ is the entropy of Z_{ji}. Evident from the relational results of Equation 2.5 is the degree of correlation between two distributions. A non zero result represents the degree of correlation within the two. Implied from Equation 2.5 is the self similar information, indicative of degree of similarity of a distribution when it is compared with itself, that serves as the upper bound. The lower bound of this expression represents the independence of two distributions that results from the orthogonality of the distance measure when $KL = 0$. Equation 2.5 may be rewritten in an alternative form such as,

$$I(\cdot,\cdot) = \begin{pmatrix} H_{1,1} & I_{1,2} & \cdots & I_{1,n} \\ I_{1,2} & H_{2,2} & \cdots & I_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ I_{1,n} & I_{2,n} & \cdots & H_{n,n} \end{pmatrix} \tag{2.6}$$
$$I(\cdot,\cdot) = I(Z_{ji}, Z_{lk}),$$
$$j == l,$$
$$\forall i, \forall j, \forall k$$

[2]Given $Z_j^{(n)}$, the ith element of the column vector is Z_{ji}, thus the mutual information between each i over all j is $I(Z_{ji}, Z_{lk}),\ j,l \in \{1 \ldots m\},\ i,k \in \{1 \ldots n\}$.

and,

$$
\begin{aligned}
& I(\cdot,\cdot) \simeq 0 \\
& I(\cdot,\cdot) = I(Z_{ji}, Z_{lk}), \\
& j \neq l, \\
& \quad \iff P(Z_{ji}) \perp P(Z_{lk}), \\
& \forall i, \forall j, \forall k
\end{aligned}
\tag{2.7}
$$

that lends itself to further interpretation. The above *nxn* matrix indicates that the maximum measure occurs on the diagonal and subsequently the off-diagonal elements decrease the measure the further from the diagonal. The zero case, for this non-negative measure occurs when the distributions contain no interconnected information and are thus statistically independent.

Extending the basic relation of this distance measure to the context of acoustic modeling, one can state that given the observation vectors, Equation 2.2, each component represents a measurement of the speech signal taken at successive instances in time and that the correlation between these measurements can be interpreted with the *KL* divergence. Furthermore, with respect to this *time series*, any given model, in order to represent the true characteristics of the signal, must take into account Equation 2.6 and Equation 2.7 to accurately model the signal. This infers that each self similar measurement maximizes the distance, $Z_{\cdot,\cdot} = H(Z_{\cdot,\cdot})$, and the correlation of each successive measurement thereafter is proportional to the mutual information, and hence decreases over time. Moreover, in order for the relation of Equation 2.7 to hold, the parameterization of the speech signal should be such that each realization is independent within each speech frame. If each parameter within a measurement is not independent an effective acoustic model should encode this mutual information to prevent information loss.

2.1 Hidden variable acoustic models

The inference of words or sub word units such as phonemes[53] from the speech signal can be modeled as a sequential process that, in the discrete case, is geometrically distributed. Given this characteristic, the inference process possesses a *Markovian* property that in turn implies that the future state of the system is only dependent on the immediate past. In other words, the state, U, of the system, at time, $t+1$, is dependent on what is currently transpiring, t, and is *conditionally independent* from all past events for all time instances $T < t$. Formally, this

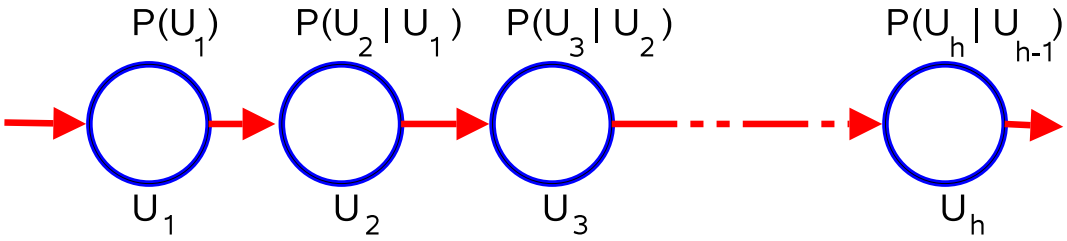


Figure 2.1: 1st order Markov chain stochastic graph

construct forms a *Markov chain* consisting of h states,

$$U_1 \longrightarrow U_2 \longrightarrow U_3 \ldots \longrightarrow U_h \tag{2.8}$$

The above equation describes a first order left to right topology that is depicted in Figure 2.1. Inherent within this structure is the conditional independence that exists between nonconnected states. This can be visualized graphically[3] with nodes of the graph representing each state of the Markov chain and edges reflecting the stochastic conditional relationship that flows from left to right. In essence, a node that succeeds another node that shares no common edge is conditionally independent to that other node. This visual relationship can be expressed in terms of the relation, $U_{(.)} \perp\!\!\!\perp U_{(.)} \,|\, U_{(.)}$, as in the case of U_3[4],

$$\begin{aligned} & P(U_1,U_2,U_3) = P(U_3|U_2)P(U_2|U_1)P(U_1) \\ & \iff P(U_3 \perp\!\!\!\perp U_1 \,|\, U_2) \end{aligned} \tag{2.9}$$

Under this premise, the inference of words, sub word units, W, is modeled as a sequential process, more specifically a Markov chain. This inference with respect to the observations process, $O^{(n)}$, forms a *hidden Markov model*, HMM[58][8]. In this manner, W, is a *hidden variable* inferred from the directly observable, $O^{(n)}$. Hence, the term hidden variable model. As such, Equation 1.2, $P(W \,|\, O,\theta)$, is satisfied by its generative equivalent, the likelihood of Equation 1.3, $\approx P(O \,|\, W,\theta)$, and it is represented by Figure 2.2, where $W \in \{U_1,U_2,\ldots,U_h\}$. The observation distribution, $P(O)$, is, under this construct, now factorized over multi-

[3]note: Stochastic topologies may be depicted graphically in this dissertation; Its purpose is to describe, visually, the relationships inherent between ***rvs*** within a given topology. Each stochastic relationship between ***rvs*** can equivalently be expressed in terms of compound probabilities[30]

[4]recall that $P(c|ab) = \frac{P(a,b,c)}{P(a,b)} = \frac{P(a|b,c)P(c|b)P(b)}{P(a|b)P(b)} = P(c|b)\ iff\ P(c \perp\!\!\!\perp a \,|\, b)$, this is likewise the case with ***rvs*** U_i.

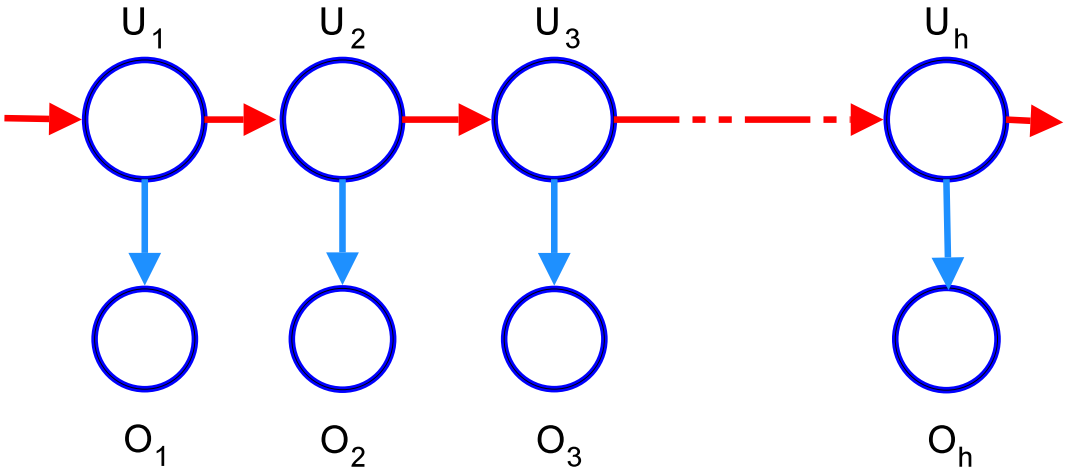


Figure 2.2: Hidden Markov model with hidden variable U and observations O

ple stages, $U^{(h)}$; $U_i \in \{U_1, U_2, \ldots, U_h\}$, or states, furthermore this implies that $P(O_i \perp\!\!\!\perp O_j | U_i)$, $j \neq i$ over all time steps or stages i and j where $i, j \in \mathscr{I}$, and $P(O_i \perp\!\!\!\perp U_j | U_i)$, $j \neq i$, $\forall i\, j$. The resultant general form for the joint distribution, given $O^{(n)}$ and $U^{(n)}$ is,

$$
\begin{aligned}
P(U^{(n)}, O^{(n)}) &= P(O_n O^{(n-1)} U^{(n)}) \qquad (2.10)\\
&= \psi(n)\\
&= P(O_n | O^{(n-1)} U^{(n)})\, P(O^{(n-1)} | U^{(n)})\, P(U^{(n)})\\
&= P(O_n | O^{(n-1)}\, U_n U^{(n-1)})\, P(O^{(n-1)} | U_n U^{(n-1)})\, P(U_n | U^{(n-1)})\, P(U^{(n-1)})\\
&= P(O_n | U_n)\, P(O^{(n-1)} | U^{(n-1)})\, P(U_n | U^{(n-1)})\, P(U^{(n-1)})\\
&= P(O_n | U_n)\, P(U_n | U_{n-1})\, P(O_{n-1} O^{(n-2)}\, U^{(n-1)})\\
&= P(O_n | U_n)\, P(U_n | U_{n-1})\, \psi(n-1)\\
&= \pi_0 \prod_{n=2}^{n} P(U_n | U_{n-1}) \prod_{n=1}^{n} P(O_n | U_n)
\end{aligned}
$$

The left most product factor of the final expression, represents the *transitional probabilities*, the probability of the state of the system in U_n given that it is in U_{n-1}. Similarly π_0 represents the steady-state initial state probabilities and the final product factor, is indicative of the *emission probabilities*, or rather the prob-

ability of a given state U_n *generating* observation O_n.

Assessing hidden variable model capabilities

The techniques described and developed earlier in the chapter are useful in assessing the "goodness of fit", or appropriateness, of HMM acoustic models applied to the problem of speech recognition. Information loss may result from an acoustic model that cannot properly encode the sequential, or transient, aspect of the speech signal. The sequential element, in this instance, arrives from the successive measurements taken with respect to time. The correlation, or KL divergence between these measurements, $O^{(n)}$, and further defined as, $Z_j^{(n)}$, that is represented by Equation 2.6, is a measure that can be used to assess the capability of the model to capture transient information within the signal. The HMM is inherently sequential due to the underlying Markov chain that represents the inference of words through a succession of states. This structure can be represented by mutual information for multiple ***rvs*** through the chain rule[22] as is evident from the following relation. As in Figure 2.1, consider three *rvs* that form a Markov chain, U_i, U_{i+1} and U_{i+2} respectively, $U_i \longrightarrow U_{i+1} \longrightarrow U_{i+2}$,

$$
\begin{aligned}
& I(U_i U_{i+1} U_{i+2}) \\
& = I(U_i U_{i+2}) + I(U_i U_{i+1} | U_{i+2}) \\
& = I(U_i U_{i+1}) + I(U_i U_{i+2} | U_{i+1})
\end{aligned}
\tag{2.11}
$$

$$
\begin{aligned}
& \Longleftrightarrow U_{i+2} \perp\!\!\!\perp U_i \,|\, U_{i+1} \\
& \Longrightarrow I(U_i U_{i+1}) \geq I(U_i U_{i+2})
\end{aligned}
$$

The inequality of Equation 2.11 describing the mutual information for all ***rvs*** that form a Markov chain satisfies the relation derived for the *KL* divergence for successive measurements of a time series. This relation is inferred from the mutual information of the chain which states that the measure decreases with each successive measurement. This is true of the upper triangle off-diagonal elements of Equation 2.6. Thus the hidden variable chain underlying the HMM is capable of encoding the transient relationship within the speech signal.

The HMM model factorizes the true observation distribution over multiple stages. Each factorized observation distribution is linked through the hidden variable process. The ability of this model to represent the true observation distribution, $P(O^{(n)}) \simeq \psi(n)$, of Equation 2.10 can be analyzed through relationships

derived from the mutual information between the observations, O, and hidden states, U, over n. Using the relationship demonstrated in Equation 2.11, it can be reasoned that for O_m, O_n, and U_n,

$$\begin{aligned} & I(O_m, U_n, O_n), \quad m < n \\ & = I(O_m, O_n) + I(O_m, U_n | O_n) \\ & = I(O_m, U_n) + I(O_m, O_n | U_n) \end{aligned} \tag{2.12}$$

$$\implies I(O_m, U_n) \geq I(O_m, O_n)$$

This infers that the hidden states of the HMM can capture and represent the information contained in O_m, $m < n$, $\forall m$ and so $\psi(n-m)$ is capable of representing the observation distribution, or factorized content for the successive stage $n-m+1$. This becomes further evident when expressing Equation 2.12, with $m = n-1$, as a vector of ***rvs***,

$$I(O^{(n-1)}, U^{(n)}) \geq I(O_n, O^{(n-1)}) \tag{2.13}$$

Similarly,

$$I(O_n, U^{(n)}) \geq I(O_n, O^{(n-1)}) \tag{2.14}$$

Furthermore, the equivalence of the HMM process, $\psi(n)$ to $P(O^{(n)})$ can be expressed in terms of the entropy of the system through the following reasoning:

Since, the entropy of a vector of, n, independent and identically distributed, *iid*, ***rvs*** is defined as[5],

$$H(X^{(n)}) = \sum_{i=1}^{n} H(X_i \,|\, X_{i-1}, \ldots, X_1)$$

where, i, is the ith element of the column vector. This may be expressed in terms

[5] [22]

of a vector of ***rvs*** that represent the observation process such as,

$$
\begin{aligned}
H(O^{(n)}) &= \sum_{i=1}^{n} H(O_i \,|\, O_{i-1}, \ldots, O_1) \qquad (2.15)\\
&= \sum_{i=1}^{n-1} H(O_i \,|\, O_{i-1}, \ldots, O_1) + H(O_n \,|\, O^{(n-1)})\\
&= H(O^{(n-1)}) + H(O_n \,|\, O^{(n-1)})
\end{aligned}
$$

By definition, the mutual information between two ***rvs***, X and Y is[6],

$$I(X,Y) = H(X) - H(X \,|\, Y)$$

As such, using the above definition of mutual information expressed in terms of entropy, together with Equation 2.15, Equation 2.14 may be written as,

$$
\begin{aligned}
& I(O_n, U^{(n)}) \geq I(O_n, O^{(n-1)}) \qquad (2.16)\\
& \implies H(O_n) - H(O_n|U^{(n)}) \geq H(O^{(n-1)}) - H(O^{(n-1)} \,|\, O_n)\\
& \implies H(O^{(n)}) \geq H(O^{(n-1)}) + H(O_n \,|\, U^{(n)})\\
& \vdash H(O_n \,|\, O^{(n-1)}) \geq H(O_n \,|\, U_n)
\end{aligned}
$$

The significance of the relation of Equation 2.16 is in that it demonstrates that the HMM topology can represent the true observation distribution *given* sufficient hidden states and accurate generative, emission distributions. Moreover, the expression on the right of the final inequality, $H(O_n \,|\, U_n)$[7], represents the *expected value of the log likelihood* of the emission probability. Thus the ***maximum likelihood*** of $\psi(n)$ can potentially fully represent the inference of words from the speech process without loss.

Parameter training

The HMM model is established, with relations Equation 2.16 and Equation 2.11, to be capable of modeling the speech process effectively given that the topology consists of an adequate number of states and accurate emission probabilities. The parameters for this generative model are determined efficiently through an iter-

[6] [22]

[7] In other words, the expected value of the log of the ML estimator.

ative process that is the weighted *ML*, or expectation maximization[28], *EM*, of $\psi(n)$ of Equation 2.10 with h states,

$$EM(\theta) = argmax \sum_h ln(\pi)\xi + \sum_n \sum_n \sum_h ln(\mathbf{A})\xi + \sum_n \sum_n \sum_h ln(\mathbf{B})\gamma \tag{2.17}$$

where, **A** and **B** are matrix expressions of the transitional and emission probabilities respectively and γ and ξ are the posteriors of $P(U_n \,|\, O^{(n)})$ and $P(U_n U_{n-1} \,|\, O^{(n)})$ respectively. The latter two parameters can be expressed, through Bayesian inference, as,

$$\begin{aligned} \gamma &= \frac{P(O^{(n)} U_n)}{P(O^{(n)})} = \frac{P(O^{(m)} U_n)\, P(O^{(n-m)} \,|\, U_n)}{P(O^{(n)})} \\ \xi &= \frac{P(O^{(n)} U_n U_{n-1})}{P(O^{(n)})} = \frac{P(O^{(m)} O^{(n-m)} U_n U_{n-1})}{P(O^{(n)})}, \\ & n, m \in \mathscr{I}, m < n \end{aligned} \tag{2.18}$$

where, the denominators for both variables may be taken as a normalization factor. The numerator term in the right most expression for γ may further be expressed as, $\gamma = \frac{\alpha(U_n)\beta(U_n)}{P(O^{(n)})}$, in terms of two recursive elements, $\alpha(U_n)$, $\beta(U_n)$, where,

$$\begin{aligned} \alpha(U_n) &= P(O_m O^{(m-1)} U_n) \\ &= P(O^{(m-1)} \,|\, O_m U_n)\, P(O_m \,|\, U_n)\, P(U_n) \\ &= P(O^{(m-1)} U_n)\, P(O_m \,|\, U_n) \\ &= P(O_m \,|\, U_n) \sum_{n-1} P(O^{(m-1)} U_{n-1} U_n) \\ &= P(O_m \,|\, U_n) \sum_{n-1} P(U_n \,|\, O^{(m-1)} U_{n-1})\, P(O_{m-1} O^{(m-2)} U_{n-1}) \\ &= P(O_m \,|\, U_n) \sum_{n-1} P(U_n \,|\, U_{n-1})\, \alpha(U_{n-1}) \end{aligned} \tag{2.19}$$

and,

$$\begin{aligned}\beta(U_n) &= P(O^{(n-m)}\,U_n) \\ &= P(O_{n-m}\,O^{(n-m+1)}\,|\,U_n) \\ &= P(O_{n-m}\,|\,U_n)\sum_{n+1} P(O^{(n-m+1)}\,|\,U_n U_{n+1})\,P(U_{n+1}\,|\,U_n) \\ &= P(O_{n-m}\,|\,U_n)\sum_{n+1} P(U_{n+1}\,|\,U_n)\beta(U_{n+1})\end{aligned} \tag{2.20}$$

Furthermore, the second of the two parameters of the weighted ML, ξ, can be expressed in terms of α and β to complete the models' parameter learning process as in,

$$\begin{aligned}\xi &= P(U_n\,O^{(m)}\,O^{(n-m)}\,U_{n-1}) \\ &= P(O_m\,O^{(m-1)}\,O^{(n-m)}\,U_n U_{n-1}) \\ &= P(O_m|U_n)\,P(O^{(m-1)}\,|\,U_{n-1})\,P(O^{(n-m)}\,|\,U_n)\,P(U_n\,|\,U_{n-1})\,P(U_{n-1}) \\ &= P(O_m\,|\,U_n)\,P(U_n\,|\,U_{n-1})\,P(O^{(m-1)}\,U_{n-1})\,P(O^{(n-m)}\,|\,U_n) \\ &= P(O_m\,|\,U_n)\,P(U_n\,|\,U_{n-1})\,\alpha(U_{n-1})\,\beta(U_n)\end{aligned} \tag{2.21}$$

2.2 Discriminative acoustic models

Discriminative techniques can be applied to the acoustic model problem in many differing manners. As described in Equation 1.4, $x^* = \min_x f(\mathbf{x_i}, \mathbf{x_{j\neq i}})$, discriminative techniques can be used to determine the posterior, Equation 1.2, $P(W|O,\theta)$, by optimizing, or differentiating between all possible classes, $\boldsymbol{x}$. In other words, given a set of models, $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$, find the model,$\theta_i$, that the word is associated with in relation to all other models, $\theta_j,\ j \neq i,\ j \in \{1 \ldots n\}$.

The optimal differentiation between classes of data can be conducted using both Bayesian inference techniques and through optimizing the distance between classes. One such example of the Bayesian approach includes expressing the posterior as

$$P(\theta\,|\,O) = \frac{P(O\,|\,\theta_i)}{\sum_j P(O\,|\,\theta_j)P(\theta_j)} \tag{2.22}$$

Which can easily be shown to be equivalent to Bayes' rule[35] applied to the

term on the left with the term on the right written in terms of marginals. Here, the posterior is expressed in a generative form. Specifically this is in terms of a given model, θ_i, *generating* the observation. Another Bayesian method that can be used to optimally discriminate between classes is to determine the minimum KL divergence between a given model generating an observation and all other classes, $KL\left(P(O\,|\,\theta_i) \parallel \sum_j P(O\,|\,\theta_j)\,P(\theta_j)\right)$.

Optimizing the distance, or minimizing the distance between classes to determine the posterior can be performed using a variety of techniques. These models, in contrast to the Bayesian methods, are non generative models. A basic quadratic form of this method is determining the cross sectional plane at the minimum of a surface. As in,

$$x^* = \min_x \frac{1}{2}\mathbf{x^T A x} + \mathbf{x^T b} \tag{2.23}$$

The solution to this form of *quadratic programming* problem is a linear solution of the form $\mathbf{A\,x} + \mathbf{b}$, where $\mathbf{x}$, $\mathbf{b} \in \mathscr{R}^m$ and $\mathbf{A}$ is a matrix with, m, columns $\mathbf{a} \in \mathscr{R}^m$. Logit regression is another common form for minimizing the distance between classes. With a set of observations, $\mathscr{O} = \{O_1, O_2, \ldots, O_n\}$, and vectors $\mathbf{x} : \mathbf{x} \in \mathscr{O}$ representing input samples belonging to a given class. Here, the ever familiar *regression* expression, together with coefficient matrices $\mathbf{A}$ with m columns, $\mathbf{a} \in \mathscr{R}^n$, weights, $\mathbf{W} : \mathbf{w} \in \mathscr{R}^m$,

$$x^* = \left(\mathbf{A^T W^T W A}\right)^{-1} \mathbf{A^T W^T W b} \tag{2.24}$$

and outputs $\mathbf{b} : \mathbf{b} \in \mathscr{R}^m$, can be used to determine the optimal distance.

Typically, these techniques model the speech process with acoustic models that discriminate between classes of data. In other words, they classify words from the speech signal. The primary intent of these approaches is to reduce or minimize the classification error rate in distinguishing one word from another. Though minimizing the error results in effective classifiers, the resultant models may not fully describe a time varying signal. Earlier in this chapter it was shown, Section 2.1, that a hidden variable topology is effective for acoustic modeling. Through capturing the transient behavior of the speech signal within its hidden states and expressing the observation distribution with ML estimators, this stochastic representation is capable of modeling the speech process without loss. As is described in Chapter 4, this dissertation presents a methodology that describes the speech process with discriminatively trained acoustic models. More

formally, it poses the speech modeling problem as that of modeling ML estimators with discriminative learning methods. It is shown that in segmenting the acoustic space to one that lends itself to these discriminative methods and by modeling ML estimators with large margin techniques, the resultant models maximize the entropy of the observation distribution.

2.3 Findings and Summary

Problem :
To devise and develop effective stochastic models for modeling the speech process.
Dissertation Contributions :

- Development of objective function relating true observation distribution to an estimate in terms of the directly observable measurements and latent hidden variables.
- Formulation and development of an optimal stochastic model for the speech with noise problem. Proposal of a combination of recognizers that through a simple system fusion, combines multiple speech processes at the decision level. This is a novel stochastic method devised to combine a parameterized spectral missing data, MD, theory based and a cepstral based speech process using a coupled hidden variable topology. In using a fused coupled hidden Markov model, HMM, topology, an optimal stochastic model is proposed that is inherently more robust than single process models under noisy conditions.
- A novel analysis and comparison of the capabilities of coupled hidden variable topologies to model the speech process.
- Through the maximization of the devised objective function, Equation 2.16, it is shown that the resultant optimal combined acoustic space contains greater information content of the true observation distribution. Thus is capable of improved recognition accuracies.

- Segmentation of the speech acoustic space in a manner that can represent the speech process effectively and can be modeled with discriminative learning methods.

- Devising an optimal discriminant ML estimator to model the speech observation distribution.

The main focus of this thesis is to devise and develop effective stochastic probabilistic models for speech recognition. This chapter described stochastic modeling of the speech process. Within it the hidden variable topology was described and fundamental discriminative learning concepts were introduced. The capability of a hidden variable topology to model the speech process was analyzed. A significant objective function, Equation 2.16, was devised as a result of this analysis.

This expression, Equation 2.16, describes the observation distribution of the speech process, $P(O)$, in terms of directly observable observations and latent hidden variables, O and U respectively. Its expresses the entropy of the true observation speech process in terms of n random variables, ***rvs*** O, the term on the left of this inequality,

$$H(O_n \,|\, O^{(n-1)}) \geq H(O_n \,|\, U_n)$$

and its estimate in terms of hidden variable topology ***rvs*** O and U. This describes the true speech observation at time, n, O_n given its previous *n-1* realizations, $O^{(n-1)}$, in relation to its estimate at time, n, or the ML estimator. In maximizing the objective function on the right of this inequality, the observation distribution can be represented by the ML estimator without loss. Put another way, just as the entropy of a ***rv***, that takes on a specific number of states each with a given probability, can be defined to be the minimum number of bits necessary to recover a message. Here the term on the right of the inequality can represent the capability of the ML estimator to represent the observation distribution. As the entropy of this term increases, its *information content* that represents the observation distribution increases. As it approaches its upper limit, the amount of information loss decreases. Thus in maximizing the objective function, the hidden variable topology is capable of encoding the observation distribution of the speech process without loss.

The fundamental models presented, the concepts introduced and the objective function formulated provide the basis for the models formulated in this thesis. Specifically, acoustic models devised to effectively model the speech process. Ef-

fective models are devised for both, the speech with noise problem through combining classifiers and for discriminatively trained acoustic models. The models formulated and developed in this thesis apply the objective function of Equation 2.16 to increase the information content of the speech process in the resultant models. Therefore, the models experience increased recognition accuracy performance for speech recognition.

1. *speech in noise:* Using the concepts introduced in this chapter, Chapter 3 proposes an effective stochastic acoustic model for the speech with noise problem. Here, through increasing the information content of acoustic models by combining classifiers and exploiting complementarity[15] information, effective stochastic models can be formulated. Using an optimal[8] coupled hidden variable topology, two streams of parameterized speech signals are fused at the decision level. This approach together with missing data, MD, techniques can provide acoustic models that have improved robustness under both stationary and non stationary noise conditions without any *apriori* knowledge of the noise disturbance. It is shown that the fusion of classifiers strengthens the structure of the acoustic model by satisfying the objective function devised in this chapter, and that it enhances the inference of words under noisy conditions.

2. *optimal ML estimators:* Chapter 4 proposes a methodology for discriminatively trained acoustic models. Whereas the proposed model of Chapter 3 maximizes the devised objective function, Equation 2.16, and thereby increases the acoustic content with a coupled topology. Here, an estimator is devised using large margin discriminative classification techniques that optimize this objective function. It is shown that the resultant models are not only capable of minimizing the information loss, but also maximize the entropy of the observation distribution.

[8] optimal in the sense of minimizing the error between the true observation distribution of the speech process and its estimate.

Chapter 3

Speech with Noise: Combination of Recognizers

Book I

Combination of Recognizers

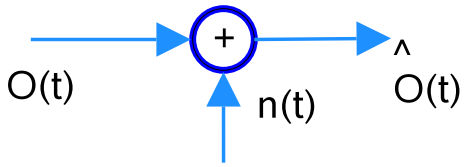


Figure 3.1: Speech with additive noise

Speech with noise is an open research problem. This chapter develops and demonstrates a methodology proposal to enhance speech recognition under adverse conditions. As was described in the previous chapters, this entails mapping the speech recognition problem to that of a stochastic time series problem with an input signal, $\hat{O}$. In this case, $\hat{O}$ is a combination of a clean signal, O, with additive noise, $\mathfrak{n}$. Through the use of missing data, MD, techniques[20] the time series problem becomes that of deciphering an incomplete input signal. These techniques exploit the inherent redundancy of speech[60] to achieve robust speech recognition even under *non stationary* noise conditions. Combining classifiers can provide a method to improve the probabilistic acoustic content accuracy of acoustic models. This approach together with MD techniques can provide acoustic models that have improved robustness under both stationary and non stationary noise conditions. Under this premise, the methodology proposes a combination of recognizers that fuses two streams of parameterized speech signals at the decision level to both, strengthen the structure of the acoustic model, as in Equation 2.16, and to enhance the inference of words from, $\hat{O}$, Equation 1.1. Specifically, a fused coupled time series model that forms an optimal acoustic model to model the speech process. Furthermore, this combination of classifiers method addresses a known problem common to typical MD methods[20][13][37][42][59] as it provides an effective method to incorporate cepstral features in the MD process.

3.1 Speech with Noise

The speech with noise problem, can be described as, $\hat{O}(t) = O(t) + \mathfrak{n}(t)$, Figure 3.1. As was implied in the introduction, the past few decades have seen great advances in speech recognition under adverse conditions. Techniques such as cepstral mean normalization[1] and RASTA[40] have been successfully applied to improve the robustness of speech recognition under some noise conditions. Cepstral mean normalization, CMN, for instance, subtracts the mean of the signal so as to remove the *glottal* effect[53] on the input signal, thus, with an input, $O^{(n)}$, where, n, are

the number of samples.

$$\hat{O}^{(n)} = O^{(n)} - \frac{1}{n}\sum_{i=1}^{n} O_i, \hat{O}^{(n)} \equiv \mathbf{o} : \mathbf{o} \in \mathscr{O} = \{\hat{O}_1, \hat{O}_2, \ldots \hat{O}_n\} \tag{3.1}$$

The demonstrated improvements, though, are usually limited to predictable and controlled environments. Moreover, the benefits of such methods are generally witnessed when the underlying noise condition can be classified as stationary. These accommodating conditions, unfortunately, are often not present. To enhance speech recognition under these *non stationary* noise conditions, approaches have been developed that may be viewed to be either signal processing front end based, or model based approaches. The former generally consists of processing the input signal, $\hat{O}$, to suppress, $\mathfrak{n}$, or to heighten the signal attributes of O. Such examples include noise masking[46] and blind deconvolution[66] approaches. The latter of the two, the model based approaches, adapt the clean speech, O, stochastic model of the recognizer, Equation 2.10, $P\left(U^{(n)}, O^{(n)}\right)$, to be a model of $\hat{O}$.

One such example of the model based approaches is the HMM decomposition[71] method. In this method an adapted model is forged by combining the clean speech, O, model with a stochastic model of, $\mathfrak{n}$. In terms of ψ of Equation 2.10, where, $\psi = P\left(U^{(n)} O^{(n)}\right)$ the resultant joint space can be defined to be ψ_d. Let the joint space of these two combined models, (1) and (2), that are combined using an operator, $\otimes$, be $\psi_d(n-m, q-r)$. The following relation expresses this in terms of model the clean speech model, (O), at time step, or stage, $n-m$ and the noise model, (n) in some stage $q-r$. Its resultant joint space in this case is,

$$\psi_d(n-m, q-r) = \psi_{(O)}(n-m) \otimes \psi_{(\mathfrak{n})}(q-r) \tag{3.2}$$

$$\because log(\hat{O}) = log(O + \mathfrak{n}) \approx max(O, \mathfrak{n}),\ O \perp \mathfrak{n},$$

$$\psi_O(n-m) \otimes \psi_{\mathfrak{n}}(q-r) =$$
$$P(U_{n-m}|U_{n-m-1})\,P(U_{q-r}|U_{q-r-1})\,\left[\psi_{n-m-1}\,P(O_{q-r}|U_{q-r}) + \psi_{q-r-1}\,P(O_{m-n}|U_{m-n})\right]$$

where, ψ_{n-m-1} and ψ_{q-r-1} are the marginals. The expression for the combined, $\hat{O}$ probabilistic space, Equation 3.2 is based on the $log(\hat{O})$ max operator on the joint expression function, $f(O, \mathfrak{n})$, the second line of the expression. The loga-

rithmic representation describes the *separation* of $f(.)$ and hence the model based decomposition of $\hat{O}$. Due to the combined joint distribution space of both speech and noise, this technique is effective for robust speech recognition under both stationary and nonstationary noise conditions. It is hindered, though, by its assumption of log based, or spectral parameterization representation of the speech signal. On the other hand, another model based compensation method, the Parallel Model Combination[33], PMC, method, does not require this assumption. This model adapts the *parameters* of the acoustic models to represent the distribution of $\hat{O} = O + \mathfrak{n}$, or in other words, the joint distribution of $f(O, \mathfrak{n})$. Should the parameters of the model be based on non spectral parameterizations of the signal, each model is transformed to the log spectral domain. Like the HMM decomposition method, PMC models the speech acoustic space with hidden variable stochastic models. The emission densities, or generative ML estimators, $P(O^{(n)} \mid U^{(n)}, \theta_i), \forall i, i \in \mathscr{I}, i \in \{1 \ldots h\}$, in terms of the model, θ, of state i of h can be expressed as a normal (Gaussian) distribution. Such a distribution is defined by its first two moments, the mean μ and the variance σ^2. The multivariate form of these emission densities can take the form of,

$$P(O^{(n)} \mid U^{(n)}, \theta_i) = N(O^{(n)} \mid \mu_{\theta_i}^{(n)}; \Sigma_{\theta_i}^{(n)}) \tag{3.3}$$

where Σ is the covariance. The PMC model adaptation method determines the speech noise distribution parameters for each, i of h. In this case the composite signal is composed of clean speech, O, and noise, $\mathfrak{n}$, thus PMC models a combination of models, (1), and (2), for clean speech and noise respectively. More specifically, for each acoustic model that represents the signal,$\hat{O}$, namely, $N(O_{(1)}^{(n)} \mid \mu_{\theta_i}^{(n)}; \Sigma_{\theta_i}^{(n)})$ and $N(O_{(2)}^{(n)} \mid \mu_{\theta_j}^{(n)}; \Sigma_{\theta_j}^{(n)}), \forall i,\ j$, the PMC method determines from the statistics of the underlying model distribution the parameters for the combined joint space. The emission ML estimators, typically, due to the nature of speech and its inherent variability, as well as due to the multimodal distribution requirement of modeling $\mathfrak{n}$, are represented by a sum of Gaussians, or rather a k mixture distribution. The sum of Gaussians in this sense is with regard to the frequency domain and not the time domain. As such each observation distribution, Equation 3.3, for

each corresponding model can be expressed as,

$$P(O^{(n)} \,|\, U^{(n)}, \theta_i) = \sum_{l=1}^{k} \pi_{l\,\theta_i} N(O^{(n)} \,|\, \mu_{l\,\theta_i}^{(n)}; \Sigma_{l\,\theta_i}^{(n)}) \tag{3.4}$$

$$\sum_{l=1}^{k} \pi_l = 1$$

where, $\pi_{l,\theta}$ is the l^{th} mixture weight for model, θ. Thus the PMC method devises a method to infer the model parameters for the joint space of $f(O,n)$, $P(\hat{O}^{(n)} \,|\, U^{(n)}, \theta_i) = \sum_{l=1}^{k} \pi_{l\,\theta_i} N(\hat{O}^{(n)} \,|\, \mu_{l\,\theta_i}^{(n)}; \Sigma_{l\,\theta_i}^{(n)})$ from those of $P(O^{(n)} \,|\, U^{(n)}, \theta_i)$ and $P(\mathfrak{n}^{(n)} \,|\, U^{(n)}, \theta_j)$. Though both of the model based compensation methods are effective for noise robust speech recognition for both stationary and nonstationary adverse conditions, they require, due to the determination of the joint space of $f(O, \mathfrak{n})$ in the log spectral domain, *apriori* knowledge of, $\mathfrak{n}$, the potential interfering noise source.

In contrast to the described model based methods, the MD approach to robust speech recognition is not reliant on apriori knowledge of the underlying acoustic conditions. Thus it is a promising method for achieving noise resilient speech recognition. The foundations of MD theory Automatic Speech Recognition, ASR, are based on the premise that recognition should only be conducted with the speech bearing components of the signal, O_r, where $O_r \in \hat{\mathscr{O}} - \mathscr{N}$ and $\hat{\mathscr{O}} = \{\hat{O}_1, \hat{O}_2, \ldots, \hat{O}_n\}$, $\mathscr{N} = \{\mathfrak{n}_1, \mathfrak{n}_2, \ldots, \mathfrak{n}_n\}$. As such, in this speech recognition process, a speech signal that is composed of both speech and noise, $\hat{O} = f(O, \mathfrak{n})$, is segregated prior to the pattern recognition and inference of words from the signal stage. The potential of MD ASR can be inferred from the near perfect recognition accuracies that have been reported when the signal is properly segregated[4].

The subsequent subsections of this chapter introduce the promising MD theory ASR methods. Each base realization of MD theory is presented and described from the proposed segregation techniques to the manner that each realization decodes the speech speech signal. The case for the use of cepstral based features with the MD framework is subsequently made. This leads to the proposed combination of classifiers, or combination of recognizers methodology to achieve noise robust speech recognition with MD theory.

3.2 Missing Data Theory

The missing data approach to speech recognition advocates that robust ASR can be achieved by using only the speech bearing components of a composite signal that consists of both speech and noise. Referring to Figure 3.1 this segregation of the signal may be represented in terms of, n, parameterized samples as in,

$$\begin{aligned}
&\hat{O}^{(n)} = f(O,\mathfrak{n})^{(n)} : O \in \mathscr{O}, \mathfrak{n} \in \mathscr{N}, \\
&O_r^{(n)} : O_r \in \hat{\mathscr{O}} - \mathscr{N}, \\
&O_u^{(n)} : O_u \in \hat{\mathscr{O}} \cap \mathscr{N}, \\
&\\
&\hat{\mathscr{O}} = \{\hat{O}_1, \hat{O}_2, \ldots, \hat{O}_n\}, \mathscr{O} = \{O_1, O_2, \ldots, O_n\}, \mathscr{N} = \{\mathfrak{n}_1, \mathfrak{n}_2, \ldots, \mathfrak{n}_n\}
\end{aligned} \tag{3.5}$$

where $\hat{O} = f(O,n)$ represents the composite signal with O_r and O_u the speech and noise bearing components of the signal respectively. How is it possible to perform pattern matching and infer words from the resultant, O_r, incomplete signal representation? This has been explained through the understanding that a speech signal is highly redundant[60]. This very redundancy forms the main idea behind MD ASR. The theory itself is derived from how humans are believed to perceive and process speech[12]. The human auditory system, under this premise, segments or groups auditory signals based on the originating auditory source to form distinct auditory objects. Thus, an auditory signal that is a mixture of speech and noise, $f(O,n)$, is segregated into separate auditory streams with speech perception and understanding conducted using solely the speech, O_r, auditory object. The resultant segregation of the signal leaves gaps in the speech signal where, $\hat{\mathscr{O}} \cap \mathscr{N}$, and thus the noise has completely occluded the original speech signal, implying that the clean speech signal component cannot be recovered, or $O_i \neq \hat{O}_i$ and that $\hat{O}_i \in \mathscr{N}$. This incomplete speech signal is accommodated by the human auditory system by utilizing the inherent redundancy of the speech signal in a process known as the *continuity illusion*. This very process forms the basis of MD theory and from this premise the theory constructs a framework to mimic the human auditory system.

The segregation of an auditory signal into speech, or rather *reliable components*, O_r, and noise *unreliable components*, O_u, for the purpose of MD ASR can be conducted in many different manners[14][27]. The more sophisticated manner, and the one that holds the most promise, is based on computational auditory scene analysis, CASA, techniques[73][62]. Here, the process follows that of

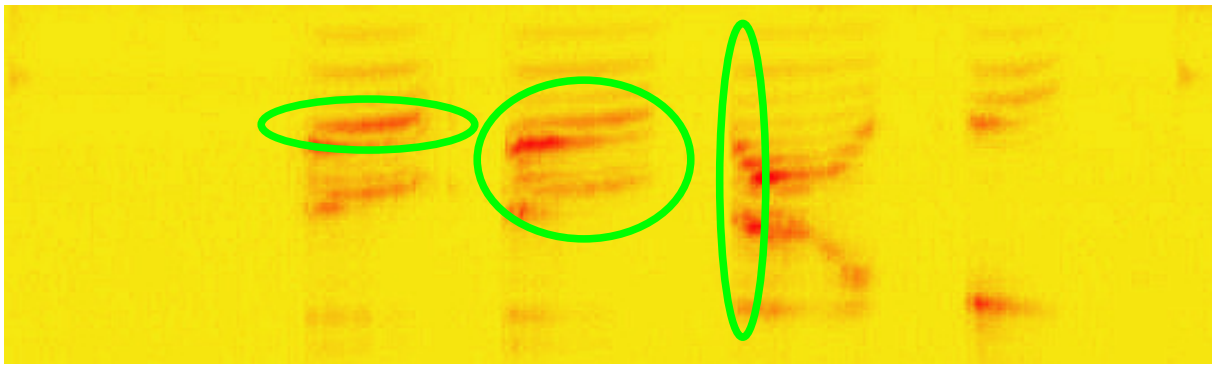
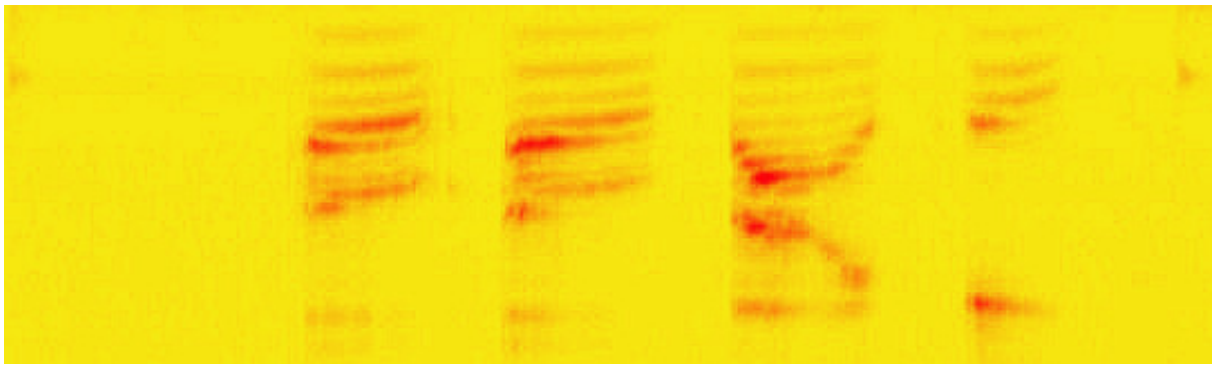


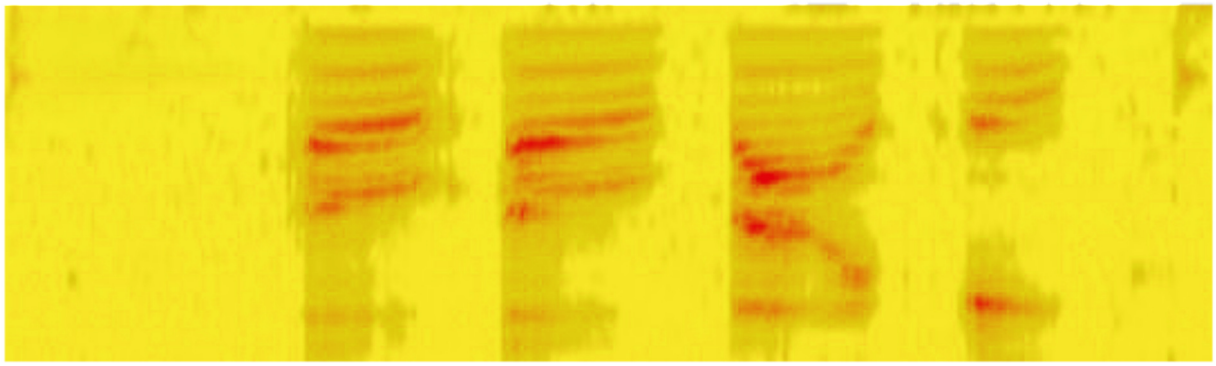
Figure 3.2: CASA groupings with spectral representation of the utterance "oh oh 5 9"; regions marked from left to right: frequency proximity, harmonicity, and onset times

what some consider analogous to image scene analysis techniques[52] whereby primitive groupings are formed with the aid of auditory cues such as frequency proximity, spatial location and harmonics. Such a grouping may be formed by examining the frequency location and the continuity of the signal from a time frequency representation of the auditory signal as depicted in Figure 3.2. Another method commonly employed in MD ASR is the segregation of auditory signals through signal processing techniques. Such methods decompose the signal into O_r and O_u through estimating the noise content in the signal. One such method is the decomposition based on a noise floor estimate derived from the signal. With the use of such an estimate, all signal components that possess energy content greater than those of the estimate, $\left|\hat{O}(f)\right|^2 > |\mathfrak{n}(f)|^2$ can be deemed to be reliable while those below are classified as unreliable.

The log spectral representation is commonly used in MD theory ASR due to its effectiveness for segregating the auditory signal. Its suitability is especially evident for CASA techniques as implied from Figure 3.2. The resultant segregation of an auditory signal takes form in an MD mask[14]. In the log spectral domain, this mask typically takes a binary form, as in Figure 3.3c, with frequency bands that are isolated as noise assigned one value and bands that are deemed to be reliable the other. This mask in relation to the log spectrum of an utterance is depicted in 3.3a and 3.3b. Of particular note is the application of so called soft MD masks to the MD ASR problem that has sometimes led to greater recognition accuracies than those of their binary counterparts[4]. In this case each frequency coefficient within a spectrum representation is assigned a probabilistic value corresponding



(a) Spectral representation of the utterance “oh oh 5 9”



(b) Spectrum overlayed with binary missing data mask (shadowed)

(c) Binary missing data mask

Figure 3.3: Spectral utterance and binary MD mask

to the degree of confidence that each band is speech.

3.3 Missing Data Theory: Pattern Recognition

MD theory predominately presents pattern recognition, or the inference of words, W, from a speech signal, O, as in $P(W|O,\theta)$ with hidden variable stochastic models, θ. More specifically, each emission distribution of the recognizers' acoustic models is a Gaussian density that generally takes the form of Equation 3.4 and thus constitutes *continuous density*, or CDHMMs.

MD pattern recognition through *data imputation* applies MD techniques for speech with noise decomposition though, in this case, infers speech from a reconstructed signal such that as in Equation 1.1, $\arg\max_W P(\mathbf{W} \mid f(\mathbf{O_r}, \hat{\mathbf{O}}_\mathbf{u}))$. Here, $\hat{O}_u$ is an estimate of the speech signal where $\hat{O}_i \in \mathscr{N}$. This reconstructed signal can then be used with any standard ASR. This implies compatibility with cepstral based speech recognizers and well established speech enhancement techniques that can be used with that standard feature. A common realization of data imputation can be forged in the following manner. Given the joint density for the model, θ of the hidden state, i,

$$\begin{aligned} p(\hat{O}^{(n)} \mid U^{(n)}\, \theta_i) &= p(O_u^{(n)}\, O_r^{(n)} \mid U^{(n)}\, \theta_i) \\ &= p(O_u^{(n)} \mid U^{(n)}\, O_r^{(n)}\, \theta_i)\, p(O_r^{(n)} \mid U^{(n)} \theta_i) \end{aligned} \tag{3.6}$$

since, this emission density is typically modeled as a Gaussian mixture, in case this with, l, such mixtures and weights, π,

$$p(O^{(n)} \mid U^{(n)}\, \theta_i) = \sum_{l=1}^{k} \pi_{l\,\theta_i} N(O^{(n)} \mid \mu_{l\,\theta_i}^{(n)}; \Sigma_{l\,\theta_i}^{(n)}) \tag{3.7}$$

then, Equation 3.6, may be expressed as,

$$p(O_u^{(n)} \mid U^{(n)}\, O_r^{(n)}\, \theta_i) = \frac{p(O_u^{(n)}\, O_r^{(n)} \mid U^{(n)}\, \theta_i)}{\sum_{l=1}^{k} \pi_{l\,\theta_i} N(O_r^{(n)} \mid \mu_{r\,l\,\theta_i}^{(n)}; \Sigma_{r\,l\,\theta_i}^{(n)})} \tag{3.8}$$

where the numerator of Equation 3.8 can be expressed as,

$$p(O_u^{(n)} O_r^{(n)} \,|\, U^{(n)}\, \theta_i) = p(O_u^{(n)} \,|\, U^{(n)}\, \theta_i)\, p(O_r^{(n)} \,|\, U^{(n)}\, \theta_i) \tag{3.9}$$
$$\iff p(O_u^{(n)} O_r^{(n)} \,|\, U^{(n)}\, \theta_i) = N(\mathbf{O} \,|\, \mu, \Sigma);\ \Sigma = \mathbf{I}\sigma^{\mathbf{2}}$$

so the expected value of O_u, or rather its estimate, $\hat{O}_u$, is

$$\hat{O}_u^{(n)} = \frac{\sum_{l=1}^{k} \pi_{l\,\theta_i} N(O_r^{(n)} \,|\, \mu_{r\,l\,\theta_i}^{(n)}; \Sigma_{r\,l\,\theta_i}^{(n)})\, \mu_{u\,l\,\theta_i}}{\sum_{l=1}^{k} \pi_{l\,\theta_i} N(O_r^{(n)} \,|\, \mu_{r\,l\,\theta_i}^{(n)}; \Sigma_{r\,l\,\theta_i}^{(n)})} \tag{3.10}$$

which renders use of the mean, $\mu_{u\,l\,\theta_i}$ as the estimate of every component of $\hat{O} \in \mathscr{N}$. This relation is true given that $\Sigma = \mathbf{I}\sigma^{\mathbf{2}}$ can hold as a parameter of the emission distribution.

MD pattern recognition through, *marginalization*, is another method that can be used to infer words from speech. This approach, in contrast to data imputation, does not attempt to recover a clean signal from $\hat{O}$, but rather conducts inference on an incomplete representation, $\hat{O} - \mathfrak{n}$. Such a method is in line with how humans are believed to perceive and process speech; recognition based upon an incomplete signal representation. Like data imputation, the MD process by marginalization, segregates the signal prior to pattern recognition, though unlike the imputation process, it modifies the acoustic models to be representative of only the speech bearing components of the signal. This can formulated in the case of acoustic models formed with Gaussian mixture densities as follows. Given,

$$p(\hat{O}^{(n)} \,|\, U^{(n)}\, \theta_i) = \sum_{l=1}^{k} \pi_{l\,\theta_i} N(O^{(n)} \,|\, \mu_{l\,\theta_i}^{(n)}, \Sigma_{l\,\theta_i}^{(n)}) \tag{3.11}$$

where, $p(\hat{O}^{(n)} \,|\, U^{(n)}\, \theta_i)$ may be expressed as,

$$p(O_u^{(n)} O_r^{(n)} \,|\, U^{(n)}\, \theta_i) = p(O_u^{(n)} \,|\, U^{(n)}\, \theta_i)\, p(O_r^{(n)} \,|\, U^{(n)}\, \theta_i) \tag{3.12}$$
$$\iff p(O_u^{(n)} O_r^{(n)} \,|\, U^{(n)}\, \theta_i) = N(\mathbf{O} \,|\, \mu, \Sigma);\ \Sigma = \mathbf{I}\sigma^{\mathbf{2}}$$

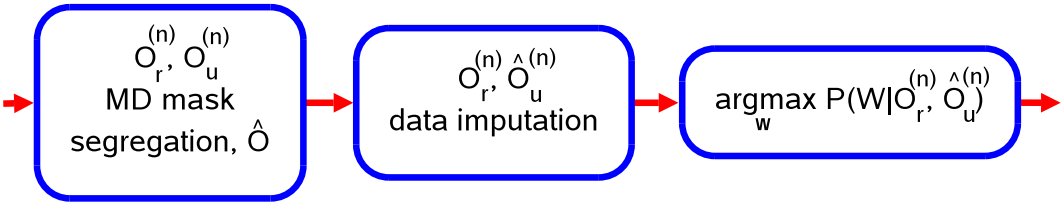


Figure 3.4: MD data imputation pattern recognition

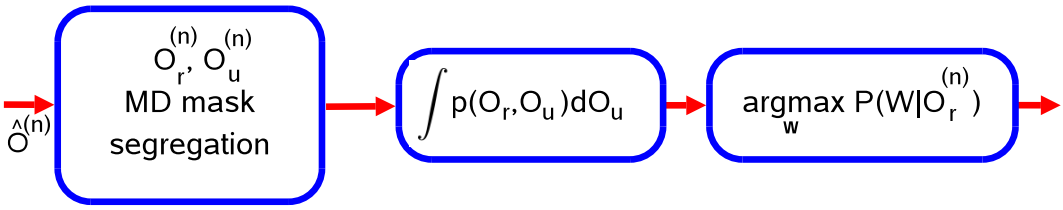


Figure 3.5: MD marginalization pattern recognition

Under this premise the acoustic density may be expressed as,

$$\int p(\hat{O}^{(n)}|U^{(n)}\theta_i)\,\mathrm{d}O_u = p(O_r^{(n)}|U^{(n)}\theta_i)\int p(O_u^{(n)}|U^{(n)}\theta_i)\,\mathrm{d}O_u \qquad (3.13)$$
$$= \sum_{l=1}^{k} \pi_{rl\theta_i} N(O_r^{(n)}|\mu_{rl\theta_i}^{(n)}, \Sigma_{rl\theta_i}^{(n)})$$

to account for only the reliable components of the signal. The expression of Equation 3.13 formulates the marginal pattern recognition problem to that of a marginalized Gaussian density CDHMM problem.

Pattern recognition with MD theory, in summary, primarily advocates the decomposition or segregation of the speech noise mixture prior to inferring words from the signal. In the case of data imputation, Figure 3.4, this involves determining estimates for elements of the signal that are deemed to be corrupt. The marginalization approach, on the other hand, as is depicted in Figure 3.5, conducts inference on elements of a segregated signal that are deemed to have originated from the speech auditory source. It is with this approach that the benefit of robust ASR with MD techniques is advocated. This pattern recognition method is associated with greater recognition accuracies[60][20] than those of some data imputation methods. This implies that a promising inference method may be pursued by applying the method that humans are believed to perceive and process

speech.

3.4 Missing Data Theory: The Case for Cepstral Features

Parameterization of speech, as depicted in Figure 1.2, for MD theory based ASR commonly takes the form of a spectral representation through processing the signal through a bank of filters. These resultant parameterized speech feature vectors permit the identification and isolation of reliable speech components, O_r, from the signal speech noise mixture, $\hat{O}$. Though the parameterized bandpass representation is suitable for speech noise decomposition, it does not lend well to typical HMM based ASR. This can be attributed to the ML estimators of Equation 2.16, $P(O|U)$, that generally take the form of Gaussian densities, $\sim \sum_{l=1}^{k} \pi_{l\,\theta_i} N(O^{(n)} \,|\, \mu_{l\,\theta_i}^{(n)}; \Sigma_{l\,\theta_i}^{(n)})$, $\Sigma = \mathbf{I}\sigma^2$.

A common approach to the density estimation problem is to maximize the likelihood of $P(O^{(n)} \,|\, \theta_i)$ over all, n, observations, O to determine the parameters of the underlying model, $\theta_i : i \in \{1 \ldots h\}$. Furthermore, given that the estimated density is $\sim N(\mu; \Sigma), \Sigma = \mathbf{I}\sigma^2$ implies that for,

$$O^{(n)} = [O_1, O_2, O_3, \ldots, O_n]^T$$

a vector of M multivariates due to parameterization, rewritten in terms of the M column vectors, j, each consisting of a dimension, n,

$$Z_j^{(n)} = [O_{1j}, O_{2j}, O_{3j}, \ldots, O_{nj}]^T$$

as in Equation 2.2, the KL divergence, or,

$$I(Z_{j}iZ_{lk}) = 0$$

the mutual information between parameterized coefficient in each speech frame is zero and thus independent in accordance with the first condition of Equation 2.5, $I\left(Z_{j}iZ_{lk}\right)$. This condition is frequently assumed to hold for much of pattern recognition with MD. Unfortunately, the information content captured in each frequency band tends to not be independent across frames. This is true of most parameterized time varying signals and therefor the condition seldom truly holds. Whereas acoustic models forged from spectral parameterized representations may

incur loss of information due to such an assumption, cepstral based features have the potential to provide statistically independent dimensions. In this case, since each frame is derived from, $\sim DCT^{-1}\left(log\,DCT(O^{(n)})\right)$, the spectrum of the log spectrum, it ensures that each coefficient within the frame is independent with respect to each other, thus satisfying the relation of Equation 2.7, $I(\cdot,\cdot)$. This attribute of cepstral based features has prompted wide use of the feature for HMM based ASR. Though cepstral features are well suited for ASR, they do not lend themselves well for segregating the signal due to its feature vector representation. During the cepstral transformation process, corrupted components of the signal, O_u, are smeared globally as a result of DCT^{-1}. Direct application of MD techniques in this case, where regions of uncertainty have been isolated and masked, results in pattern recognition with an estimation of O_r that is severely degraded. The recognition performance from this resultant signal is inferior to that of recognition with $\hat{O}$ itself. The identification of localized uncertainties within a speech frame is crucial in MD theory and so the very characteristic that allows MFCC features to be well suited for HMM based ASR hinders their adoption in MD techniques.

The attractiveness of cepstral based features has spurred many efforts to see their use within the MD framework. The majority of these efforts have addressed the use of such features with acoustic models consisting of Gaussian mixture emission densities. This can be expressed in terms of, k mixtures for a given hidden state model, $\theta_i : i \in \{1,\ldots,h\}$ as,

$$\sum_{l=1}^{k} \pi_{l\,\theta_i} N(O^{(n)} \,|\, \mu_{l\,\theta_i}^{(n)}; \Sigma_{l\,\theta_i}^{(N)})$$

which is equivalent to the quadratic[29],

$$\left(O^{(n)} - \mu_{l\,\theta_i}^{(n)}\right)^T \mathbf{W}^T\mathbf{W}\left(O^{(n)} - \mu_{l\,\theta_i}^{(n)}\right),\ \mathbf{W}^T\mathbf{W} = \Sigma^{-1}, \qquad (3.14)$$
$$O^{(n)} : O_i \in \mathscr{R}^m, \mathbf{W} : \mathbf{w} \in \mathscr{R}^m$$

for each mixture, l, of the density. The cepstral distance weight method[42], for instance, formulates the use of cepstral features with MD problem as an optimization problem in line with that of the Equation 2.23, $\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{b}$. In this case, the distance to be minimized is between that of O and $\hat{O}$, and not between classes as discussed in Section 2.2. Here this is equivalent an approximation of the maxi-

mum likelihood of Equation 2.16, $H(O_n|U_n)$. As such given,

$$\hat{O}(t) = O(t) + \mathfrak{n}(t),$$

the MD pattern recognition marginalized acoustic model with mask weights, $\mathbf{M}$, expressed in terms of Equation 3.14,

$$\left(\hat{O}^{(n)} - \mu_{l\,\theta_i}^{(n)}\right)^T \mathbf{M}^T \mathbf{W}^T \mathbf{W} \mathbf{M} \left(\hat{O}^{(n)} - \mu_{l\,\theta_i}^{(n)}\right), \mathbf{W}^T\mathbf{W} = \Sigma^{-1}, \tag{3.15}$$
$$\mathbf{M} : \mathbf{m} \in \mathscr{R}^m, \Sigma^{-1} : \sigma \in \mathscr{R}^m$$

the linear transformation of $\hat{\mathbf{O}}$ to the cepstral domain with a the cepstral transformation matrix, $\mathbf{C}$,

$$\mathbf{y} = \mathbf{C}\hat{\mathbf{O}}, \tag{3.16}$$

and Equation 3.15 expressed in terms of the cepstral domain,

$$\left(\hat{O}^{(n)} - \mu_{l\,\theta_i}^{(n)}\right)^T \mathbf{C}\mathbf{M}^T \mathbf{C}^T \mathbf{W}^T \mathbf{W} \mathbf{C}\mathbf{M}\mathbf{C}^T \left(\hat{O}^{(n)} - \mu_{l\,\theta_i}^{(n)}\right), \mathbf{W}^T\mathbf{W} = \Sigma^{-1}, \tag{3.17}$$
$$\mathbf{C} : \mathbf{c} \in \mathscr{R}^n \mathbf{M} : \mathbf{m} \in \mathscr{R}^m, \Sigma^{-1} : \sigma \in \mathscr{R}^m$$

the maximum likelihood of O can be approximated as,

$$O^{(n)} \approx \min_{\hat{O}} \frac{1}{2} \left(\hat{O}^{(n)} - \mu_{l\,\theta_i}^{(n)}\right)^T \mathbf{C}\mathbf{M}^T \mathbf{C}^T \mathbf{W}^T \mathbf{W} \mathbf{C}\mathbf{M}\mathbf{C}^T \left(\hat{O}^{(n)} - \mu_{l\,\theta_i}^{(n)}\right) + \mathbf{b}, \tag{3.18}$$
$$\mathbf{b} \in \mathscr{R}^m$$

With this method, as expressed in Equation 3.18, the marginalization MD pattern recognition process is realized with a cepstral parameterization representation by applying the MD mask in the log spectral domain prior to inference in the cepstral domain. The application of the mask and the marginalization process in the spectral domain is, in this method, considered to form *cepstral distance weights* in the cepstral domain. These weights are expected to decompose the signal, in other words, determine O from the speech mixture of $\hat{O}$. Though the ML approximation of Equation 3.18 was determined to be successful for ASR with localized noise adverse conditions, the performance of the method for other cases was reported as disappointing.

There have also been proposals to see cepstral features used within the MD data imputation pattern recognition process. Of particular note is the *state* imputation method[36] that infers clean speech from corrupted regions within the cepstral parameterized representation by establishing bounds on the energy of the signal. More specifically, given the emission density of Equation 3.14, with observations, $\hat{O}$, and the cepstral transformation, Equation 3.16, together with,

$$\begin{aligned} &log(\hat{O}) = log(O + \mathfrak{n}) \approx max(O, \mathfrak{n}),\ O \perp \mathfrak{n}, \\ &\mathfrak{n} \leq \hat{O}, \end{aligned} \tag{3.19}$$

the emission density rewritten in terms of the individual signal components, O_r and O_u,

$$\left(\mathbf{C}\,O_r^{(n)} + \mathbf{C}\,O_u^{(n)} - \mu_{l\,\theta_i}^{(n)}\right)^T \mathbf{W}^T\mathbf{W}\left(\mathbf{C}\,O_r^{(n)} + \mathbf{C}\,O_u^{(n)} - \mu_{l\,\theta_i}^{(n)}\right), \tag{3.20}$$

and taken into account, $\hat{O} - O = O_u$, then

$$O^{(N)} \approx \left(\mathbf{C}\,O_u^{(N)}\right)^T \mathbf{W}^T\mathbf{W}\left(\mathbf{C}\,\hat{O}^{(N)} - \mu_{l\,\theta_i}^{(N)}\right) \tag{3.21}$$

formulates a least squares problem to infer O from $\hat{O}$. The approximation of Equation 3.19 is stated to improve recognition accuracies over both those of MD marginalization and data imputation pattern recognition processes. Unfortunately, the method was found to be feasible only for static features and was ill suited for dynamic cepstral features such as velocity and acceleration. Feature reconstruction methods as [59] first segregate the speech signal into reliable and unreliable components in the spectral domain. Aposterior distributions are then used to determine clean speech estimates for the corrupted coefficients in the signal $\hat{O}$. The use of MD techniques with standard recognizers is stated to be amongst the benefits of these *maximum aposterior*, MAP, methods. These techniques, though, generally did not achieve the same level of recognition accuracies as that of the MD marginalization method.

Posing the MD cepstral domain pattern recognition problem as that of modeling interacting stochastic processes has many advantages. Amongst those is an acoustic model that provides a richer characterization of the speech process. This implies an acoustic model that lessens the loss of the speech signal's information content as is discussed in Chapter 2. In other words, modeling the problem as that of a combining spectral and cepstral parameterized representations of the signal

at the pattern recognition, or *decision level*, stage, results in a joint space that increases the speech content, or rather, the entropy of the system in accordance with Equation 2.16, $H\left(O_n|O^{(n-1)}\right)$.

3.5 Modeling Interacting Processes

Acoustic models that model interacting processes can use any one of a number of topologies. As a first attempt, a model of these, g, interrelated processes may be forged by considering the Cartesian product of each process. Such a model is formed with a parameterized representation, $X^{(n)} = [X_1, X_2, X_3, ..., X_n]^T$. Each element of this column vector, X_i, is a multivariate of degree $m = \sum_{i=1}^{g} M_i$, where M_i are the number of parameterized components of each process i of g. The nature of the correlation between each M_i from each observation measurement should be considered with this form since any variation between the multiple processes could be reflected in the resultant acoustic model as noise[11]. Moreover, the resulting Cartesian product feature space may suffer from the curse of dimensionality[9]. These problems may be avoided by properly smoothing the feature space. One such approach that can be considered is to reduce the correlations found between the interacting processes to a set of linear combinations that approximate the relationship between each process. These combinations can be realized by finding the projection, P, of the product space that minimizes the distance between the relationship between processes and those of the measurements. The solution to this problem may be formulated as that of Equation 2.24 such that[1],

$$\begin{aligned} \mathbf{x}^* &= \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b} \\ &= \mathbf{P}\mathbf{b}, \\ &\mathbf{A} : \mathbf{a} \in \mathscr{R}^n, \mathbf{b} \in \mathscr{R}^n, \mathbf{P} : \mathbf{p} \in \mathscr{R}^m \end{aligned} \tag{3.22}$$

[1] Recall that the coefficient matrix, $\mathbf{A}$, represents the coefficients of $X^{(n)} = [X_1, X_2, X_3, \ldots, X_n]^T$

where, P may be obtained through the spectral theorem. Therefore, this projection can be rewritten in terms of orthogonal matrices, $\mathbf{U}$ and $\mathbf{V}$ so,

$$\begin{aligned} &\mathbf{P} = \mathbf{A}^{-1} \\ &\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T, \\ &\mathbf{U}^T\mathbf{U} = \mathbf{I},\ \mathbf{V}^T\mathbf{V} = \mathbf{I}, \\ &\mathbf{U} : \mathbf{u} \in \mathscr{R}^n, \Sigma : \sigma \in \mathscr{R}^m, \mathbf{U} : \mathbf{u} \in \mathscr{R}^m \end{aligned} \tag{3.23}$$

where, $\Sigma = \mathbf{I}\sigma$ and $\sigma = \sqrt{\lambda}$. Though this method can be effective to reduce the feature space, its effectiveness to determine independent dimensions relies upon sufficient samples to make available the underlying correlations in the processes. Furthermore, removing the variations that potentially exist within each parameterized sample can incur loss of information that may prevent effective modeling the of speech process. This may become evident with resultant acoustic models that are difficult to train, suffer from overfitting and have reduced recognition accuracies.

Through decision level modeling the interacting processes are combined within the model. As opposed to altering the feature space to approximate the relationship between multiple processes, decision level modeling infers the interconnected attributes of all process within the modeling stage. Given this technique, models prescribing to this methodology may not be subject to feature dimensional issues and can be demonstrated to effectively capture the speech process within the resultant acoustic models. The potential effectiveness of such models can be inferred from the development of acoustic models as described in Chapter 2 from which it can be reasoned that stochastic time series can capture the speech process without loss. Furthermore, the use of decision level modeling in the form of stochastic coupled time series is capable of supporting both the MD marginalization pattern recognition method of Section 3.3 and standard cepstral HMM ASR. Such coupled time series may be realized in any number of topologies based on the connections made between each process. The subject of the proceeding section is to examine the capability of each such topology to model the combined process. This novel analysis, using the statistical analysis methods established in Chapter 2, may reveal an appropriate coupled acoustic model to capture both the MD and cepstral ASR processes.

3.6 Coupled Stochastic Time Series

The modeling of multiple stochastic processes can be realized in a coupled time series topology. The resultant model infers the statistical dependencies between multiple time series with interconnections that exhibit these dependencies and are captured within the marginal probabilities of the model. In other words, given the standard HMM topology of Chapter 2 with observations, O, and hidden states, U and initial state probabilities, π_0,

$$\begin{aligned} P(U^{(n)}, O^{(n)}) &= P(O_n O^{(n-1)} U^{(n)}) \\ &= \psi(n) \\ &= \pi_0 \prod_{n=2}^{n} P(U_n|U_{n-1}) \prod_{n=1}^{n} P(O_n|U_n), \end{aligned}$$

depicted in Figure 3.6, the statistical dependencies should be realized through each factorized stage, n, and captured in the resultant joint space, $\psi(n)$. This can be achieved through careful connections made between each of the, g, processes, or time series in Figure 3.7. The *coupled* models[11][6][63][47], that are formed through connections either between hidden states, observations or both, attempt to describe the combined processes. The potential for such a combined model to represent the statistical attributes for all processes can be examined through analyzing the mutual information relationship between observations, O, and hidden states, U for a given factor stage, n.

One such combined coupled model is the coupled HMM model[11]. This time series topology, Figure 3.8, models the dependencies between multiple, g, HMM chains with *cross transitional* connections in an effort to capture the temporal, or transient, qualities between the multiple stochastic processes. Thus connections are made between the hidden states of one model with hidden states from the previous stage of the other time series. In other words, for a given, U, node of the model representing stage, m, the following relationship holds for time series (1), in terms of the (g) time series,

$$P(U_{(1)}^{(n-m)} \,|\, U_{(1)}^{(n-m-1)} U_{(2)}^{(n-m-1)}, U_{(3)}^{(n-m-1)}, \ldots U_{(g)}^{(n-m-1)}) \tag{3.24}$$

or in general for any given time series, h, and node, m, within the combined model

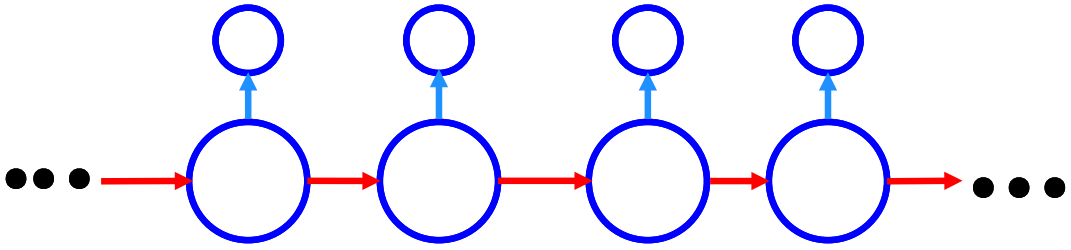


Figure 3.6: HMM topology

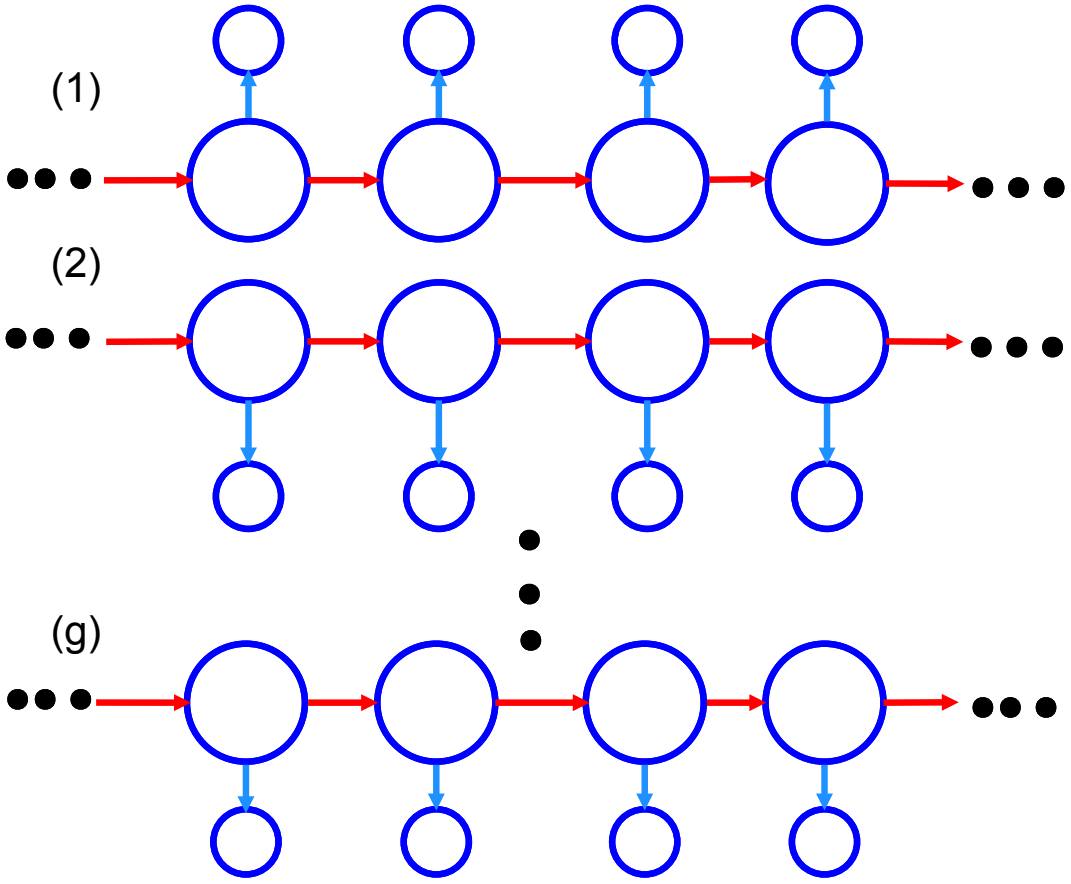


Figure 3.7: Multiple, g, stochastic time series

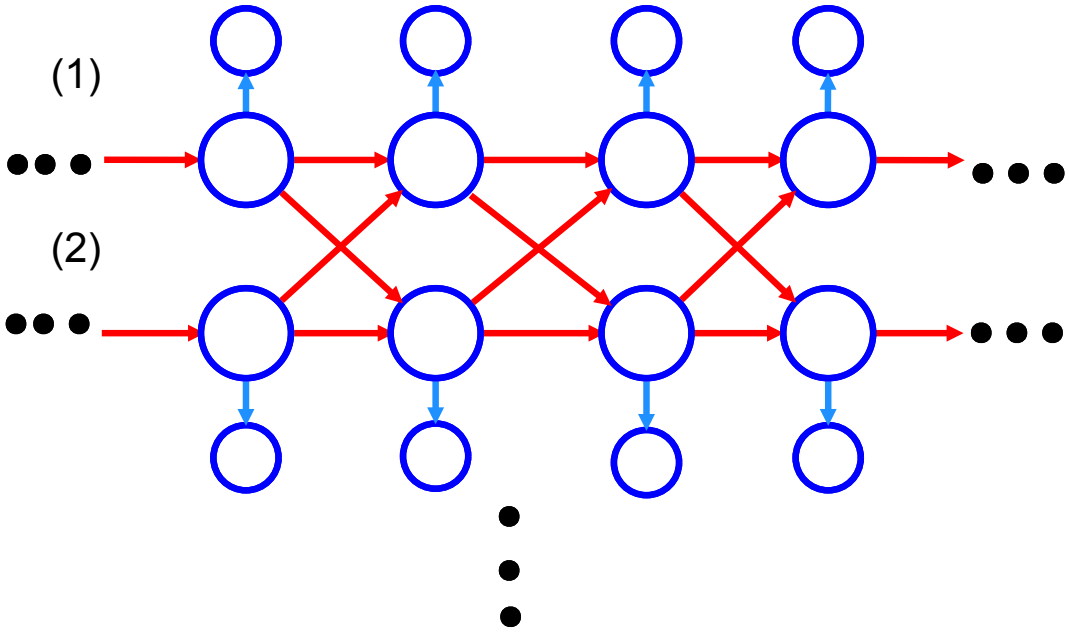


Figure 3.8: Coupled HMM

representing all (g) time series[2],

$$P(U_{m_{(h)}} \,|\, U_{m-o_{(f)}}), \tag{3.25}$$
$$\forall h, f, h, f \in \mathscr{I}, 1 \ldots g, \forall o, o \in \mathscr{I}, 1 \ldots m-1$$
$$U_{m-o_{(.)}} \in \{U_{m-o_{(1)}}, \ldots U_{m-o_{(g)}}\}$$

The capability of this model to capture the temporal characteristics of the combined stochastic process can be analyzed by, for simplicity, considering the coupling of two time series. Through the use of the information theoretic concept of mutual information, the relationship between the nodes of the model for a given

[2] $\cdot_{\cdot}$ for example as in Equation 3.24, $P(U_{(1)}^{(n-m-1)}) = P(U_{n-m-1_{(1)}} U_{(1)}^{(n-m-2)})$
$= P(U_{n-m-1_{(1)}} U_{n-m-2_{(1)}}, U_{(1)}^{(n-m-3)})$, etc.

factor stage m may be represented in the form of the chain[3],

$$U_{m_{(1)}} \longrightarrow U_{m-1_{(2)}} \longrightarrow U_{m_{(2)}} \tag{3.26}$$

The hidden states of model (1) and (2) in this stage can be written in terms of Equation 2.11 as,

$$\begin{aligned} & I\left(U_{m_{(1)}} U_{m-1_{(2)}} U_{m_{(2)}}\right) \\ & = I\left(U_{m_{(1)}} U_{m-1_{(2)}}\right) + I\left(U_{m_{(1)}} U_{m_{(2)}} \,|\, U_{m-1_{(2)}}\right) \\ & = I\left(U_{m_{(1)}} U_{m_{(2)}}\right) + I\left(U_{m_{(1)}} U_{m-1_{(2)}} \,|\, U_{m_{(2)}}\right) \\ & \\ & \implies I\left(U_{m_{(1)}} U_{m-1_{(2)}}\right) \geq I\left(U_{m_{(1)}} U_{m_{(2)}}\right) \end{aligned} \tag{3.27}$$

Furthermore,

$$I\left(U_{m-1_{(1)}} U_{m_{(2)}}\right) \geq I\left(U_{m_{(1)}} U_{m_{(2)}}\right) \tag{3.28}$$

Equation 3.27 together with Equation 3.28 imply that the cross connections, $U_{m-1_{(2)}} \longrightarrow U_{m_{(1)}}$ and $U_{m-1_{(1)}} \longrightarrow U_{m_{(2)}}$ respectively contain greater or at least an equal amount of the temporal information than that of the hidden variables $U_{m_{(1)}}$ and $U_{m_{(2)}}$ combined. Therefore, should the expression of Equation 3.26 hold, the transient behavior of all stochastic processes can be captured by the resultant joint space of this topology.

More elaborate HMM topologies, such as that of the *mixed memory*[63] HMM, Figure 3.9, and its variants, strengthen the coupling between the multiple chains by linking together not only the hidden variables, but those of the observations as well. More specifically, to further enhance the model, at each given stage of the model, connections are made between the hidden variables of one time series to the observations of the others. In other words, for a given, O, node of the model

[3] $\because P(U_{m_{(2)}} \perp\!\!\!\perp U_{m_{(1)}} \,|\, U_{m-1_{(2)}})$ or,
$P\left(U_{m_{(1)}} U_{m-1_{(2)}} U_{m_{(2)}}\right) =$
$P\left(U_{m-1_{(2)}}\right) P\left(U_{m_{(1)}} \,|\, U_{m-1_{(2)}}\right) P\left(U_{m_{(2)}} \,|\, U_{m_{(1)}}\right) \equiv$
$P\left(U_{m_{(1)}}\right) P\left(U_{m-1_{(2)}} \,|\, U_{m_{(1)}}\right) P\left(U_{m_{(2)}} \,|\, U_{m-1_{(2)}}\right)$,
which is by definition, $U_{m_{(1)}} \longrightarrow U_{m-1_{(2)}} \longrightarrow U_{m_{(2)}}$

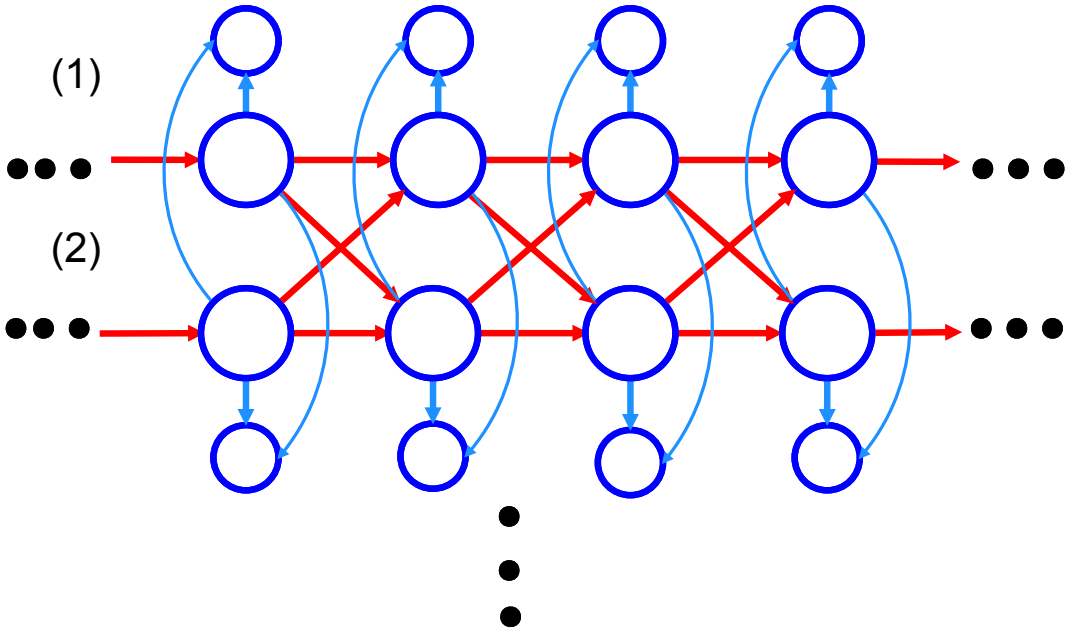


Figure 3.9: Mixed memory HMM

representing stage, m, each node is represented as,

$$P(O_{m_{(h)}} | U_{o_{(f)}}), \tag{3.29}$$
$$\forall h, f,\ h, f \in \mathscr{I}, 1 \ldots g, \forall o, o \in \mathscr{I}, 1 \ldots m$$
$$U_{m_{(.)}} \in [U_{m_{(1)}}, \ldots U_{m_{(g)}}]$$

in addition to satisfying Equation 3.25 for each node U. Modeling interacting stochastic processes in this manner is capable of describing complex interdependent systems. The interconnections between hidden variables and observations

can be described in terms of the KL divergence as,

$$
\begin{aligned}
& I\left(O_{m_{(1)}} U_{m_{(2)}} O_{m_{(2)}}\right) \\
& = I\left(O_{m_{(1)}} O_{m_{(2)}}\right) + I\left(O_{m_{(1)}} U_{m_{(2)}} \,|\, O_{m_{(2)}}\right) \\
& = I\left(O_{m_{(1)}} U_{m_{(2)}}\right) + I\left(O_{m_{(1)}} O_{m_{(2)}} \,|\, U_{m_{(2)}}\right)
\end{aligned} \tag{3.30}
$$

$$
\implies I\left(O_{m_{(1)}} U_{m_{(2)}}\right) \geq I\left(O_{m_{(1)}} O_{m_{(2)}}\right)
$$

Similarly,

$$
I\left(O_{m_{(2)}} U_{m_{(1)}}\right) \geq I\left(O_{m_{(1)}} O_{m_{(2)}}\right) \tag{3.31}
$$

Which implies, that the joint space of $\Omega = \left\{O_{(1)}, O_{(2)}, \ldots, O_{(g)}\right\}$ can be captured within the connections between the hidden variables and observations for each stage of the model. Thus Equation 3.27, Equation 3.28, Equation 3.30 and Equation 3.31 indicate that a model such as the mixed memory HMM, can within its topology, encode both the transient and the factorized observation space of all stochastic processes within it without loss of information. Direct inference of this model, though, unfortunately cannot be realized due to the number of required parameters which in most cases makes it computationally intractable.

The fused HMM model[55] on the other hand attempts to describe a coupled HMM topology that possesses the characteristics of the prior two models, but with optimized connections between each time series. As proposed in [55], the model forms connections in accordance with minimizing the KL divergence between the combined joint space of all processes, (g), $P\left(O_{(1)}, O_{(2)}, \ldots, O_{(g)}\right)$ and its estimate, $P\left(\hat{O}_{(1)}, \hat{O}_{(2)}, \ldots, \hat{O}_{(g)}\right)$. Therefore,

$$
\begin{aligned}
& KL\left(p\left(O_{(1)}, O_{(2)}, \ldots, O_{(g)}\right) \parallel p\left(\hat{O}_{(1)}, \hat{O}_{(2)}, \ldots, \hat{O}_{(g)}\right)\right) = \\
& -\int \ldots \int p\left(O_{(1)}, O_{(2)}, \ldots, O_{(g)}\right) ln\left(\frac{p\left(\hat{O}_{(1)}, \hat{O}_{(2)}, \ldots, \hat{O}_{(g)}\right)}{p\left(O_{(1)}, O_{(2)}, \ldots, O_{(g)}\right)}\right) \mathrm{d}O_{(1)} \mathrm{d}O_{(2)} \ldots \mathrm{d}O_{(g)}
\end{aligned} \tag{3.32}
$$

Rewritten in terms of $I\left(O_{(1)}, O_{(2)}, \ldots, O_{(g)}\right)$ and $I\left(\hat{O}_{(1)}, \hat{O}_{(2)}, \ldots, \hat{O}_{(g)}\right)$ it can be

shown that this distance is equivalent to,

$$KL\left(p\left(O_{(1)}, O_{(2)}, \ldots, O_{(g)}\right) \parallel p\left(\hat{O}_{(1)}, \hat{O}_{(2)}, \ldots, \hat{O}_{(g)}\right)\right) = \\ I\left(O_{(1)}, O_{(2)}, \ldots, O_{(g)}\right) - I\left(\hat{O}_{(1)}, \hat{O}_{(2)}, \ldots, \hat{O}_{(g)}\right) \tag{3.33}$$

Thus, so as to minimize this distance, the second term of Equation 3.33 must be maximized.

The optimal connections between time series of the topology Figure 3.7 can be readily seen, from Equation 3.30 and Equation 3.31, for two such time series, and can be further extended to the factorization of the observation and hidden state space over n as,

$$I\left(O_{(1)}^{(n)} U_{n_{(2)}}\right) \geq I\left(O_{(1)}^{(n)} O_{(2)}^{(n)}\right) \tag{3.34}$$

and,

$$I\left(O_{(2)}^{(n)} U_{n_{(1)}}\right) \geq I\left(O_{(1)}^{(n)} O_{(2)}^{(n)}\right) \tag{3.35}$$

Therefore, the optimal coupled HMM for the observation space $P(O)$ is found through minimizing the KL distance and takes the form of Figure 3.10. Furthermore, the transient, or temporal, behavior of all stochastic models are captured within this construct. Just as the coupled HMM model, Equation 3.25, is forged with cross connections between U, and is shown to capture this aspect of the time series within Equation 3.27 and Equation 3.28, the fused HMM can also be reasoned to possess this capability. In this case, connections between stochastic processes in the form of time series are made, for a given stage, m, between the observable variable of one and hidden state of all other time series. This relationship can be

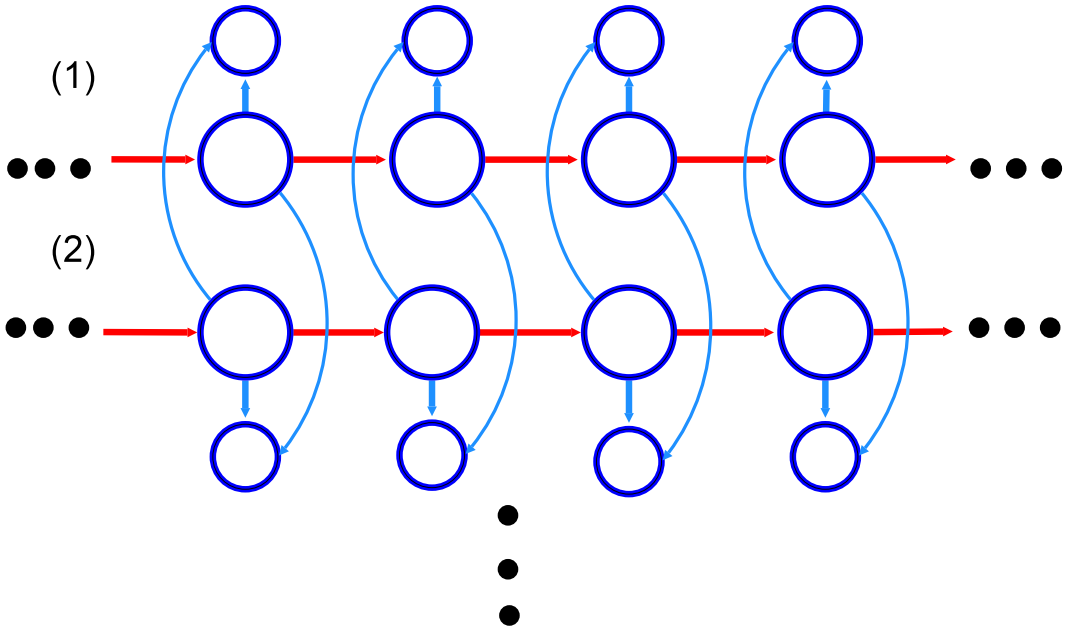


Figure 3.10: Fused HMM

described as,

$$
\begin{aligned}
& I\left(f(O)_{m-1_{(1)}} U_{m-1_{(2)}} U_{m_{(2)}}\right) \qquad (3.36)\\
& = I\left(U_{m-1_{(1)}} U_{m-1_{(2)}} U_{m_{(2)}}\right)\\
& = I\left(U_{m-1_{(1)}} U_{m-1_{(2)}}\right) + I\left(U_{m-1_{(1)}} U_{m_{(2)}} | U_{m-1_{(2)}}\right)\\
& = I\left(U_{m\_1(1)} U_{m_{(2)}}\right) + I\left(U_{m-1_{(1)}} U_{m-1_{(2)}} | U_{m_{(2)}}\right)\\
& \implies I\left(U_{m-1_{(1)}} U_{m-1_{(2)}}\right) \geq I\left(U_{m-1_{(1)}} U_{m_{(2)}}\right)\\
& I\left(O_{m-1_{(1)}}, U_{m-1_{(2)}}\right) \geq I\left(f(O)_{m-1_{(1)}}, U_{m-1_{(2)}}\right)\\
& so,\\
& \implies I\left(O_{m-1_{(1)}} U_{m-1_{(2)}}\right) \geq I\left(U_{m-1_{(1)}} U_{m_{(2)}}\right)
\end{aligned}
$$

Equation 3.36 implies that the temporal information of the stochastic processes that is propagated from one time series to another in, $U_{m-1_{(1)}} \longrightarrow U_{m_{(2)}}$, is, with

connections between the observation and hidden variables, capable of containing at least this within, $U_{m-1_{(1)}} \longrightarrow U_{m-1_{(2)}}$. This can also be true of the cross connection, $U_{m-1_{(2)}} \longrightarrow U_{m_{(1)}}$. Moreover, a coupled HMM topology with connections in each stage of the model, m, between those of the observations of one with the hidden states of all others, renders the expression Equation 3.26 not necessarily true. Therefore, in this case, the final expressions of Equation 3.27 and Equation 3.28 may not hold. Thus the cross connection links within this coupled topology may add unnecessary complexities to the resultant joint model. Both in terms of forming the joint space and in inferring the temporal statistical dependencies within the stochastic processes. Within the fused topology, as implied in Equation 3.36, the temporal dependencies may be encoded within the model without loss.

The fused HMM model is capable of representing the factorized observation space of multiple stochastic processes. In addition to accurately representing the joint space of the processes, the topology also accommodates encoding the transient properties of the time series making it appropriate, as it satisfies Equation 2.11 and subsequently Equation 2.5, to model the speech process. This model lends itself to pose the MD cepstral domain pattern recognition problem as that of modeling interacting stochastic processes. Within this combination of spectral based MD and cepstral based stochastic processes the proceeding subsection describes the resultant model as one that is capable of providing an effective robust ASR pattern recognition method.

3.7 Combination of Recognizers

Presented within this section is an approach to strengthen, or rather enhance, robust speech recognition through combining classifiers. This encompasses the use of both, MD techniques with marginalization, for the separation of the speech noise signal,$\hat{O}$ and thereafter pattern recognition with O_r, and cepstral based ASR techniques. Existing methods attempt to incorporate cepstral based features into the MD theory by first transforming the cepstral stochastic process into the spectral domain for speech noise decomposition and then subsequently perform pattern recognition in the cepstral domain. As opposed to those conventional methods, the cepstral MD ASR problem is proposed as that of combining, or fusing, interacting processes. Under this pretense, separate auditory stochastic processes are combined to enhance the representation of the originating auditory signal.

Decision level modeling provides the opportunity to combine both spectral and cepstral parameterized signals at the pattern recognition stage. Furthermore, the

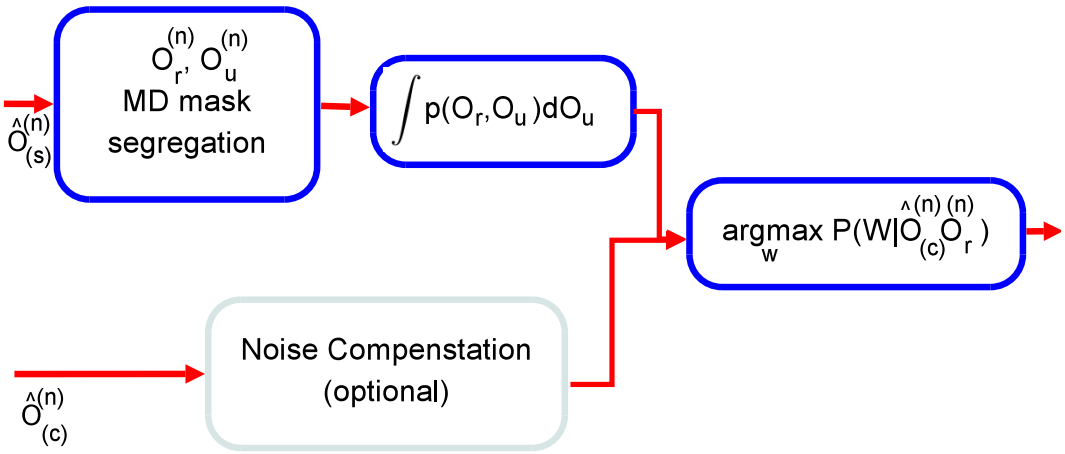


Figure 3.11: Combination of recognizers pattern recognition

decision level modeling of coupled stochastic time series lends itself to combining MD marginalization pattern recognition with cepstral HMM classification. This *combination of recognizers* serves to combine ASR with MD theory with cepstral based parameterizations. Implied within its construct is a combined model that utilizes both speech noise separation, pattern recognition with O_r and the benefits of cepstral based features. Central to the premise of the combined model is noise robust ASR without *apriori* knowledge of n under both stationary and non stationary noise conditions.

The general architecture of the combined model is that of Figure 3.11. This combination of recognizers model performs speech recognition with two separate auditory streams of the same originating auditory source. The first, $\hat{O}^{(n)}_{(s)}$, is the spectral, (s), parameterized signal that forms the top stream of Figure 3.11. The second, or bottom stream of the figure, is, $\hat{O}^{(n)}_{(c)}$, a cepstral, (c), based representation of the signal $\hat{O}$. The spectral stream, $\hat{O}^{(n)}_{(s)}$, is segregated or rather the signal is decomposed to satisfy Equation 3.5 with the resultant $O^{(n)}_r$ and $O^{(n)}_u$ formulated through the use of the MD mask of Section 3.2. The second of the two auditory streams, $\hat{O}^{(n)}_{(c)}$, can be further processed with any standard cepstral based noise compensation scheme such as Equation 3.1 or used as is during inference with the combined model.

The pattern recognition stage of this combination of recognizers, uses both $O^{(n)}_r$ and $O^{(n)}_{(c)}$ in inferring words from the combined joint space, $\Omega = \{O_{(s)}\, O_{(c)}\}$.

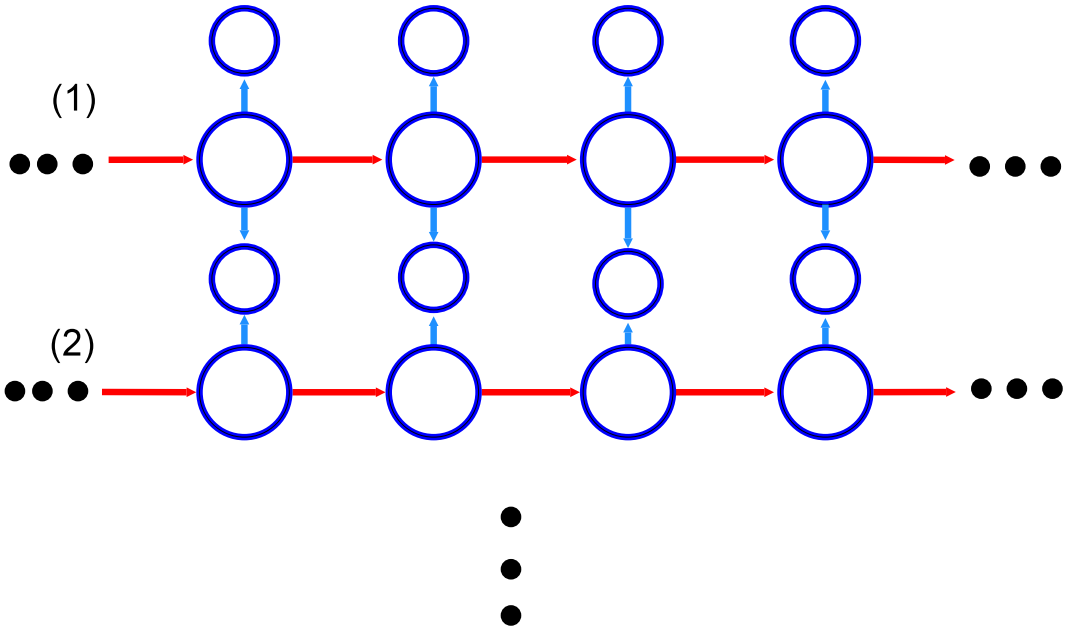


Figure 3.12: Combination of recognizers acoustic model

This fused decision level model is in the form of a coupled stochastic time series fused HMM that is depicted in Figure 3.12. The fused model not only serves to provide simple decision level combination, or rather fusion, but also optimally represents the joint sample space of, $\Omega = \{O_{(s)} O_{(c)}\}$. Optimal in the sense of minimizing the distance, or error, between the true probabilistic space and its estimate.

The effectiveness of the combined acoustic model to capture the speech process can be implied from the relations derived for the fused HMM model, Equation 3.34 and Equation 3.35. From these two expressions it can be easily shown that,

$$H\left(O_{(1)}^{(n)} | O_{(2)}^{(n)}\right) \geq H\left(O_{(1)}^{(n)} | U_{(2)}^{(n)}\right) \tag{3.37}$$

similarly,

$$H\left(O_{(1)}^{(n)} | O_{(2)}^{(n)}\right) \geq H\left(O_{(2)}^{(n)} | U_{(1)}^{(n)}\right) \tag{3.38}$$

that follows that of Equation 2.16, a significant expression of Chapter 2. Thus the

expectation of the log likelihood of the observations within this acoustic model, or rather, the ML, can capture and represent the speech process, O. Furthermore, the statistical dependencies within the stochastic processes are inferred directly within the topology of this model.

The training of this combined model follows that of standard hidden variable models as described in Section 2.1. The inference of words from the combined model can be described in terms of its joint space. The combined model is a hidden variable stochastic model that within it models two time varying time series. As such, as is described in Section 2.1, the joint space of the hidden variable topology, $P(U,O)$, can be expressed in terms of the product of the probabilities of each node, or stage, of the model. With a model of one stage, this joint space would be $P(U,O) = P(U)P(O|U)$. In general for a given stage, m, the joint expression, as is described in Equation 2.10 of Section 2.1, can be expressed in terms of $\psi(m)$. Similarly, the joint space of two series, (1) and (2), that is represented by the fused topology of Figure 3.10, can be described as $P\left(O_{(1)},O_{(2)},U_{(1)},U_{(2)}\right)$[4]. With the joint space of two stochastic time series,

$$P\left(O_{(1)}^{(n)}O_{(2)}^{(n)}U_{(1)}^{(n)}U_{(2)}^{(n)}\right) = \psi(n) \tag{3.39}$$

Define υ as the marginal [30] for a given stage, m of n so that,

$$\upsilon(m) = \sum_{\psi_{(i)\,i\neq\ m}} \psi(m) \tag{3.40}$$

The probability of a given stage, m, can then be expressed in terms of this υ as in,

$$P\left(O_{m_{(1)}}O_{m_{(2)}}U_{m_{(1)}}U_{m_{(2)}}\right) = \prod_{i=2}^{m}\upsilon(i)\prod_{i=m+1}^{n}\upsilon(i) \tag{3.41}$$

Similar to how the marginals for each stage were defined in terms of υ, Equation 3.40, let ϕ be defined as the maximum value of U and O for a given stage, m, as in,

$$\phi(m) = \max_{OU,\,\forall\psi(i)\,i\neq m} \psi(m) \tag{3.42}$$

[4]Using the calculus of statistical inference. Of course, as skeptics we may adopt a frequencist view. Though, as in [23], it can be shown that even with this skeptical view the axioms of probability still hold. As such, the supporting background and models proposed in this dissertation are presented in terms of statistical inference.

As such, the maximum O and U over the joint space for all n stages can be written as,

$$ln\left((P\left(O_{(1)}^{(n)}O_{(2)}^{(n)}U_{(1)}^{(n)}U_{(2)}^{(n)}\right)\right) = \sum_{i=2}^{n} ln\left(\phi(i)\right) \tag{3.43}$$

The inference of words from these models comes directly from determining the *best path* through a network of fused HMMs. On an individual HMM basis, this is forged from determining the best path, or most likely nodes that have generated the data. Using Equation 3.42, the maximum value or probability of a given node to have generated the data is inferred. Similarly, the maximum over the entire joint space for a given fused HMM model is expressed in Equation 3.43. Together, these two expressions permit determining the most probable path with each most probable node expressed in terms of the maximum values obtained for joint space, Equation 3.43, and the corresponding $n-1$, stages, that led to the result of n.

Specifically, each i of n, can be within any of the h states of the model. The intermediate,i, maximizations, $\max_{OU}$, between any two consecutive $i-1$ and i that satisfy $\max_{OU} P\left(O_{i_{(1)}}O_{i_{(2)}}U_{i_{(1)}}U_{i_{(2)}}\right)$, dictate the state and best (*maximal*) path. Thus ensuring a global maximized path[72] for the joint space.

The resultant combination of recognizers, COR, is capable of describing the joint space of both spectral and cepstral based time series and in doing so, conducts the inference of words from the combined joint space of cepstral and spectral based acoustic models. The benefits of this inference process is twofold. The first, is that the complementarity[15] information that may exist between the two process is inherent within the model. The statistical dependencies between the two time series are directly inferred within the pattern recognition stage due to the topology. The second is the theoretical recognition potential of this joint space. As is displayed in Table 3.1, the joint space of the model may provide robust speech recognition that exceeds recognition performance over those of cepstral based and MD theory ASR alone. This may be evident from considering that the signal, under this pretense, is that of Figure 3.1, $\hat{O}(t) = O(t) + \mathfrak{n}(t)$, and the strength of the inference, or $P\left(W\,|\,\hat{O}\,\theta\right)$ which is $\approx P\left(\hat{O}\,|\,W\,\theta\right)$ in the Bayesian sense, depends on how close $\hat{O}$ is to that of which the model, θ, represents, O. Under clean conditions, the cepstral signal, $\hat{O}$, is equivalent to the reliable components, O_r of the signal as defined in Equation 3.5. In Section 3.4 it was implied that typical HMM based acoustic models, θ, forged from cepstral based features, $O_{(c)}$, are capable of

Table 3.1: COR theoretical recognition capacities

ASR Configuration	Capacity
Conventional, Clean, $\hat{O}$	
Spectral process, $\hat{O}_{(s)}$, MD	$H\left(\hat{O}_{(s)}\right) = H\left(O_{r_{(s)}}\right)$
Cepstral process, $\hat{O}_{(c)}$	$H\left(\hat{O}_{(c)}\right) = H\left(O_{r_{(c)}}\right)$
	$H\left(\hat{O}_{(c)}\right) \geq H\left(\hat{O}_{(s)}\right)$
Fusion,COR, $\hat{O} = O + n$	
Cepstral+MD	$H\left(\hat{O}_{(c)} O_{r_{(s)}}\right) \geq H\left(O_{r_{(c)}}\right)$

modeling the speech process more effectively than those from spectral parameterizations, $O_{(s)}$. Thus should the cepstral and spectral based models be derived in this manner, then the cepstral based recognition performance, or the correct inference from that of Equation 1.3, should be greater than that of the spectral model. In other words, the second column of Table 3.1, for clean speech conditions, should hold. Furthermore, under adverse noise conditions, the joint *capacity* of the fused model to model the speech process should exceed that of the cepstral based model alone. Capacity in the sense of a given model contains increased information or true speech acoustic content. In other words, for speech recognition, it is more capable to decipher the signal. In this configuration, the acoustic space of the model includes both cepstral and spectral based models. The spectral process, of the two processes, performs pattern recognition with only O_r of $\hat{O}$ as depicted in the top portion of Figure 3.11. The cepstral process on the other hand, is more apt to capture the speech process and implies $H\left(O_{(c)}\right) \geq H\left(O_{(s)}\right)$. With the parameters of the models, θ determined from clean speech, or O_r, it can be reasoned that the joint space, more specifically, the acoustic space of the COR, is more capable of correct inference from $\hat{O}$ than the cepstral model alone. This behavior attributed to the exploitation of complementarity information between the two processes that is inherent within the topology of the COR acoustic space and is encompassed in the final relation of the second column of Table 3.1.

3.8 Findings and Summary

Problem :
To devise and develop effective stochastic models for modeling the speech process.
Dissertation Contributions :

- Development of objective function relating true observation distribution to an estimate in terms of the directly observable measurements and latent hidden variables.
- Formulation and development of an optimal stochastic model for the speech with noise problem. Proposal of a combination of recognizers that through a simple system fusion, combines multiple speech processes at the decision level. This is a novel stochastic method devised to combine a parameterized spectral missing data, MD, theory based and a cepstral based speech process using a coupled hidden variable topology. In using a fused coupled hidden Markov model, HMM, topology, an optimal stochastic model is proposed that is inherently more robust than single process models under noisy conditions.
- A novel analysis and comparison of the capabilities of coupled hidden variable topologies to model the speech process.
- Through the maximization of the devised objective function, Equation 2.16, it is shown that the resultant optimal combined acoustic space contains greater information content of the true observation distribution. Thus is capable of improved recognition accuracies.
- Segmentation of the speech acoustic space in a manner that can represent the speech process effectively and can be modeled with discriminative learning methods.
- Devising an optimal discriminant ML estimator to model the speech observation distribution.

Combining classifiers can provide a method to improve the probabilistic acoustic content of acoustic models. As was detailed in Chapter 2, hidden variable stochastic models are capable of modeling the speech process without loss. These

concepts motivate this chapters' proposed methodology on devising an effective acoustic model for the speech with noise problem.

In modeling a combined parameterized spectral MD theory based and a cepstral based speech process at the decision level, the resultant acoustic model is capable of robust speech recognition under noisy conditions. MD theory[20] based automatic speech recognition, ASR, is suitable for noise robust speech recognition since it has been demonstrated to be successful under both stationary and non stationary noise conditions without any apriori knowledge of the noise disturbance[4]. A known drawback to this method is the type of features that is commonly used to represent the parameterized speech signal. Spectral based features are typically used due to its ease of use with this method. For acoustic modeling though, spectral based features are not as suitable as cepstral based features. Pattern recognition with MD techniques is commonly done with multivariate Gaussian mixture models, $\sim \sum_{l=1}^{k} \pi_{l\,\theta_i} N(O^{(n)} \mid \mu_{l\,\theta_i}^{(n)}; \Sigma_{l\,\theta_i}^{(n)})$, where $\pi_{l\,\theta_i}$ is the mixture weight of the k mixtures for model θ, $O^{(n)}$ a vector of n ***rvs*** that represent the speech process, and Σ the covariance of the normal distribution. As such, they are commonly modeled with the assumption that, $\Sigma = \mathbf{I}\sigma^2$. In other words, that the components within a parameterized sample are independent to each other such that the mutual information,

$$I(Z_{ji} Z_{lk}) = 0,\ j \neq l,$$

where Z_{ji} and Z_{lk} are ***rvs*** that represent the ith and kth component in the jth and lth dimensional space of the speech samples. The information content captured in each frequency band of the spectral representation, though, tends to not be independent across speech frames. Thus to effectively model the speech process, alternate features such as cepstral based features have been proposed for noise robust MD ASR.

As is detailed in Section 3.4 there have been many research efforts to use cepstral based features[42][59][36] with MD theory. Unlike these previous efforts. the proposed methodology of Chapter 3 introduces these features into MD ASR through a simple system fusion. This is a novel stochastic method that is devised to combine a parameterized spectral MD and cepstral based speech process at the decision level. Through modeling the speech processes with a coupled hidden variable topology, an optimal stochastic model is formulated to increase the true speech information content of the resultant model.

- **The proposed acoustic model is optimal**,

As was described in Section 3.7 the proposed combination of recognizers fuses multiple speech observation processes at the decision level. This is realized with a fused coupled HMM model. Such a model may be expressed in terms of the Kullback-Leibler, KL, divergence (Equation 2.3) between the true joint observation distribution and its estimate. Rewritten from, Equation 3.32 of Section 3.6, this is,

$$KL\left(p\left(O_{(1)}, O_{(2)}, \ldots, O_{(g)}\right) \parallel p\left(\hat{O}_{(1)}, \hat{O}_{(2)}, \ldots, \hat{O}_{(g)}\right)\right), g \in \mathscr{J}$$

where, $O_{(i)}$ is a ***rv*** that describes the observations of speech process *(i)* and similarly, $\hat{O}_{(i)}$ is a ***rv*** that describes an estimate of the processes. Since this divergence distance measure satisfies Jensen's inequality, it is a convex function. Its minimum is a global minimum and it is therefore fully defined. As is detailed in Section 3.6, the proposed acoustic model minimizes the KL divergence between the true observation distribution and its estimate. Thus it is the optimal model in this sense[5].

- **The proposed model is capable of effectively modeling multiple stochastic processes**,
  The capability of hidden variable topologies to model multiple interacting processes is analysed in Section 3.6. From this analysis it is shown that both the transient or temporal characteristics of multiple observation processes and its joint distribution can be encoded within the coupled fused hidden variable topology. A comparison with other coupled hidden variable topologies, namely the coupled HMM[11] and the mixed memory[63] models is made. It is shown that the fused HMM model is capable of encoding multiple stochastic processes with fewer connections between the distinct time series processes than the others.

  The capability of the proposed acoustic model to encode the spectral and cepstral parameterized speech processes is described in the following inequalities of Section 3.6,

$$I\left(O_{(1)}^{(n)} U_{n_{(2)}}\right) \geq I\left(O_{(1)}^{(n)} O_{(2)}^{(n)}\right)$$

[5] proof:[22]

and,

$$I\left(O_{(2)}^{(n)}U_{n_{(1)}}\right) \geq I\left(O_{(1)}^{(n)}O_{(2)}^{(n)}\right)$$

where, $O_{(i)}^{(n)}$ is a vector of n ***rvs*** that represent the observations of the stochastic time series (i), $U_{(i)}^{(n)}$ the respective latent variables of the generative topology and $U_{n_{(i)}}$ the latent variable at time, or stage n. Here, i, can be considered to be (1) or (2) representing a spectral and a cepstral process respectively or vice versa. The expressions on the left of the two inequalities state that the mutual information, $I()$, between the observations of one process and the latent variable at n of the other is greater than that between the observations of both processes. As is described in Section 3.6, this implies that the temporal aspect of both processes is encoded within the proposed combined acoustic model.

Furthermore, it is shown that the statistical dependencies between the two time series, (1) and (2), the spectral and cepstral processes respectively or vice versa, are captured within the fused acoustic model. The two entropic, $H()$, inequalities below,

$$H\left(O_{(1)}^{(n)}|O_{(2)}^{(n)}\right) \geq H\left(O_{(1)}^{(n)}|U_{(2)}^{(n)}\right)$$

and,

$$H\left(O_{(1)}^{(n)}|O_{(2)}^{(n)}\right) \geq H\left(O_{(2)}^{(n)}|U_{(1)}^{(n)}\right)$$

are devised from the objective function of Equation 2.16, and thus imply that the maximum of the expected value of the log likelihood of the observations within this fused model, can represent the speech process. Thus together, both sets of inequalities indicate that the proposed acoustic model is capable of capturing and representing the speech process.

- **The proposed model increases the accuracy, or information content of the true speech process in the resultant models for the speech with noise problem**,
  The proposed acoustic model is a simple system fusion of a marginalized (Section 3.3) spectral MD observation process, $O_{(s)}$, and a cepstral based, $O_{(c)}$, process at the pattern recognition level. This combined model can be

reasoned to an effective speech with noise acoustic model since it contains a greater amount of true speech information content than single process models.

The speech and noise observation parameterized process, $\hat{O}$, can be considered as composed of both clean speech, O_r, and noise, O_u. Thus under clean speech, no noise, recognition conditions, $\hat{O} = O_r$. The spectral based MD process performs pattern recognition, Section 3.3, using solely O_r of $\hat{O}$. Inference from $\hat{O}$ can therefore be expressed as inference from $\hat{O} = O_{r_{(s)}}$. Inference from the cepstral based process, on the other hand, uses the entire observation, $\hat{O}$, or $\hat{O}_{(c)}$.

As is described in, Section 3.4, spectral based speech representations are not as effective as cepstral based parameterizations for acoustic modeling. Thus the amount of true speech observation content contained in spectral based models is less than cepstral based acoustic models. Let $H(O)$, the entropy, $H()$, of the true speech process, O, represent the minimum amount of information content necessary to have been encoded in the acoustic models for inference without loss. The *capacity*, or the amount of true speech acoustic content encoded in each of the spectral and cepstral models may be expressed as, $H(O_{(s)})$ and $H(O_{(c)})$ respectively. Under this premise it implies that the following is true, $H(O) \geq H(O_{(c)}) \geq H(O_{(s)})$. Which may be rewritten in terms of observation process, $\hat{O}$ as, $H(\hat{O}_{(c)}) \geq H(\hat{O}_{(s)})$. Thus, under clean conditions cepstral based models, theoretically, can be expected to experience greater recognition accuracies than spectral based models due to their greater capacity.

Under noisy conditions it is shown that the proposed fused acoustic model is capable of improved speech recognition accuracies. As the cepstral based process conducts inference from the $\hat{O}_{(c)}$ while the spectral on solely $\hat{O} = O_{r_{(s)}}$. The joint model then, under noisy conditions, can be reasoned to contain a greater amount of information content than that of a clean speech cepstral process, $O_{r_{(c)}}$, and this may be expressed as, $H\left(\hat{O}_{(c)}\, O_{r_{(s)}}\right) \geq H\left(O_{r_{(c)}}\right)$, Table 3.1. This relation is directly related to the combined acoustic model inherently capturing and encoding the statistical dependencies between the cepstral and spectral process.

Book II

Experiments

The appropriateness of the combination of recognizers, in this form, for robust speech recognition is determined from a series of experiments. The following three sections describe these experiments. The immediately following section describes the format of the experiments devised to validate the proposed methodology and to compare the behavior of the model to its theoretical potential. Experiments and results under stationary noise conditions precedes a concluding section containing experimental results of ASR with COR under non stationary noise conditions.

3.9 Combination of Recognizers: Experiments Setup

Experiments were conducted to evaluate the potential of the combination of recognizers approach to missing data automatic speech recognition. The Grid[18] corpus was used for all experiments. The selection of this particular corpus was due to the simple, phonetically balanced structure of its utterances and the appropriateness of the corpus for ASR under noisy conditions. This corpus contains 34000 utterances with each made up of 6 words in the order:
$verb=bin|lay|place|set
$color=blue|green|red|white
$prep=at|by|in|with
$letter=a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|x|y|z
$number=zero|one|two|three|four|five|six|seven|eight|nine
$coda=again|now|please|soon

($verb sp $color sp $prep sp $letter sp $number sp $coda)

Three sets of CDHMM Gaussian word level models were created with this vocabulary of 51 words, each of which possessed diagonal covariance representations using log spectral, cepstral and cepstral mean normalized features respectively. Each of the corresponding word level HMMs was composed of two phonemes per state in accordance to the mapping of words to phonemes defined in the CMU dictionary[6]. The number of states per word and the dictionary used for training purposes is listed in Table 3.2. In addition, each state of the word models consisted of 32 mixtures per state in an effort to compensate for modeling the speech pro-

[6]www. speech. cs. cmu. edu/cgi-bin/cmudict

Table 3.2: HMM States per Word in Experiments Recognizer Vocabulary

Number of States per Word	Grid Dictionary Word
4	at, by, in a-v, x-z, one, two, three, eight
6	bin, lay, place, set, blue, green, read, white, with, four, five, six, nine, now, please, soon
8	again, zero
10	seven

cess with spectral based diagonal covariance Gaussian HMMs[60], Section 3.4. The parameterized log spectral features, *ratemaps*, used in the experiments were obtained from the signal processed through a bank of gammatone filters linearly spaced from 50 to 3850Hz in ERB-rate[53]. The envelope of the output of each of these gammatone filters was smoothed with an 8ms time constant and was sampled at a frame rate of 10ms. The cepstral parameterization of the signal consisted of 39 dimensions including energy, delta and acceleration coefficients. Both the stationary noise experiments Section 3.10, and non stationary noise experiments Section 3.11 used the configurations described above, though differed in the manner the corpus was split for training and testing and is described in Table 3.3 and Table 3.4. In the case of training the stationary noise experiment models, half of the grid corpus was used, consisting of 500 unique sentence utterances generated from 34 different speakers. As for the non stationary ASR experiments, again, half of the corpus was used for training each of the three acoustic models, but in this instance, the training set consisted of 1000 unique utterances from 17 different speakers. All of these aforementioned acoustic models were constructed using the HTK toolkit[77].

Using the acoustic models formed from the 3 different sets of features, speech recognizers were set up to analyze the recognition performance of the spectral, cepstral and the fused, combination of recognizer models. Four configurations of the combination of recognizers models were constructed for these evaluations. The first two, a fusion of a ratemap based model with and without MD and an MFCC based model, the others the combination of ratemap with and without MD and normalized, CMN, features. The MD mask used for the spectral stream of features was a binary mask forged from a noise floor estimate extracted from the first few frames of each test utterance. The test utterances used in the experiments

Table 3.3: Experiments Corpus Training Sets

Stationary Noise Experiments		Non Stationary Noise Experiments	
Utterances per spkr	No. spkrs	Utterances per spkr	No. spkrs
500	34	1000	17

Table 3.4: Experiments Corpus Test Sets

Stationary Noise Experiments		Non Stationary Noise Experiments	
No. Utterances	No. spkrs	No. Utterances	No. spkrs
300	34	560	17

came from a subset of the corpus that *was not* used in the training sets. Specifically, for the stationary noise experiments, 300 utterances were used that consisted of sentences generated from 34 different speakers. For the non stationary experiments, the size of the test set was 560. This set was composed of 40 utterances generated from 14 different speakers. In this case the utterances from speakers in the test set differed from those used in the training set. A total of 8 different configurations of the recognizers, see below, were used in the experiments. All of the aforementioned ASRs were constructed using a custom implementation of the CTK toolkit[3].

i. spectral, ratemap features, rate32

ii. cepstral features, MFCC

iii. cepstral features with normalization, CMN

iv. ratemap using missing data techniques, MD

v. COR, MFCC+rate32

vi. COR, CMN+rate32

vii. COR, MFCC+MD

viii. COR,CMN+MD

Several test sets were devised to ascertain the performance of the proposed fusion of features with the combination of recognizers approach to MD theory ASR. These test sets consisted of the test corpus subjected to adverse noise conditions of varying SNRs. The stationary noise tests sets consisted of the test utterances subject to stationary noise at 6dB and 0dB. These test sets were derived from the test corpus prepared for the 2006 *speech separation challenge*[21]. As for the non stationary noise experiments, two types of *non stationary* noise disturbances, from the Noisex[70] database, were added to the test corpus. The first, the Destroyer Operations Room noise source, the second noise generated from the Factory Noise I source. The latter was chosen as it was a good example of highly non-stationary noise as it contains general machine hums with spontaneous hammer blows. These noise perturbations were added to the test corpus, using the Fant software tool[7], to form three separate sets each representative of SNRs at 18dB, 12dB and 6dB.

3.10 Combination of Recognizers: Stationary Noise Experiments

Outlined in this section are the results from experiments conducted to support the methodology of the combination of recognizers, COR, for robust speech recognition under stationary noise conditions. The objective of this section is to demonstrate that the premise in using the fusion of cepstral based and spectral based stochastic processes at the decision level, or rather in modeling the joint space of both processes, a method for robust speech recognition is satisfied. Using the experimental setup described in Section 3.9, for stationary noise conditions, the baseline, *clean speech*, recognition accuracies for recognizer configurations, *i thru iii* and *v thru vi* of Section 3.9 are detailed in Table 3.5. The recognition accuracies listed in this table, and in all subsequent results, represent the percentage of correct words recognized over each utterance in the test corpus. These baseline results are indicative of each models' ability to infer words from the test set under no noise conditions. It may be inferred from these results that each model's capacity to model the speech process corresponds to its theoretical performance

[7] http://dnt.kr.hs-niederrhein.de/download.html

Table 3.5: Stationary Noise: Baseline Recognizer Results, Clean Data

ASR Configuration		Recognition Accuracy %
Conventional		
	Spectral Features, rate32	96.6
	MFCC Features	98.0
	CMN	95.6
Fusion,Combination of Recognizers		
	MFCC+rate32	97.8
	CMN+rate32	97.8

described with the relations of Table 3.1. Here, it is implied that the MFCC, or, cepstral based model follows that of what was stated in Section 3.4 and is capable of more accurately modeling the speech process than the spectral based models. Hence, satisfying $H\left(\hat{O}_{(c)}\right) \geq H\left(\hat{O}_{(s)}\right)$, thus the results of this baseline are to be expected. Furthermore, the fused, COR, baseline results indicate that, under clean speech conditions, the model capacity is approximately the same if not greater than the more capable of the two stochastic processes to model the signal. Thus the characteristics of this joint space follows that $H\left(\hat{O}_{(c)}\, O_{r_{(s)}}\right) \geq H\left(O_{r_{(c)}}\right)$ with $H\left(\hat{O}_{(c)}\right) = H\left(O_{r_{(c)}}\right)$.

For the stationary test conditions, the recognition results were generated with the recognizer in configurations *ii thru iv* and *vii thru viii* of Section 3.9 and test sets respective of SNR 6dB and 0dB conditions. These results are reflected in Table 3.6. The results of the fused model under stationary conditions indicate that the combination of spectral based MD and cepstral based stochastic processes is capable of providing an effective robust ASR pattern recognition method. Over both test conditions, 6dB and 0dB, the COR method experiences greater recognition accuracies than those of the conventional MD and cepstral based ASR alone. The joint space forged through this fusion of features at the decision layer appears to provide an acoustic space that is better apt to capture and represent O and for the inference of W from $\hat{O}$. Moreover the results of Table 3.6 are in line with the expected behavior of the joint space, Table 3.1, namely, $H\left(\hat{O}_{(c)}\, O_{r_{(s)}}\right) \geq H\left(O_{r_{(c)}}\right)$.

In an effort to further analyze the behavior of the joint space acoustic model

Table 3.6: Recognizer Results With Test Corpus + Stationary Noise

ASR Configuration	**Recognition Accuracy %**	
Conventional	SNR 6dB	SNR 0dB
Spectral Features, MD	76.4	69.8
MFCC Features	78.2	64.9
CMN	69.4	62.1
Fusion,Combination of Recognizers		
MFCC+MD	82.4	71.3
CMN+MD	82.7	72.2

Table 3.7: Stationary Noise 6dB SNR: Recognizer Configuration Rankings Over Entire Test Set, Relative To All Experimented Configurations

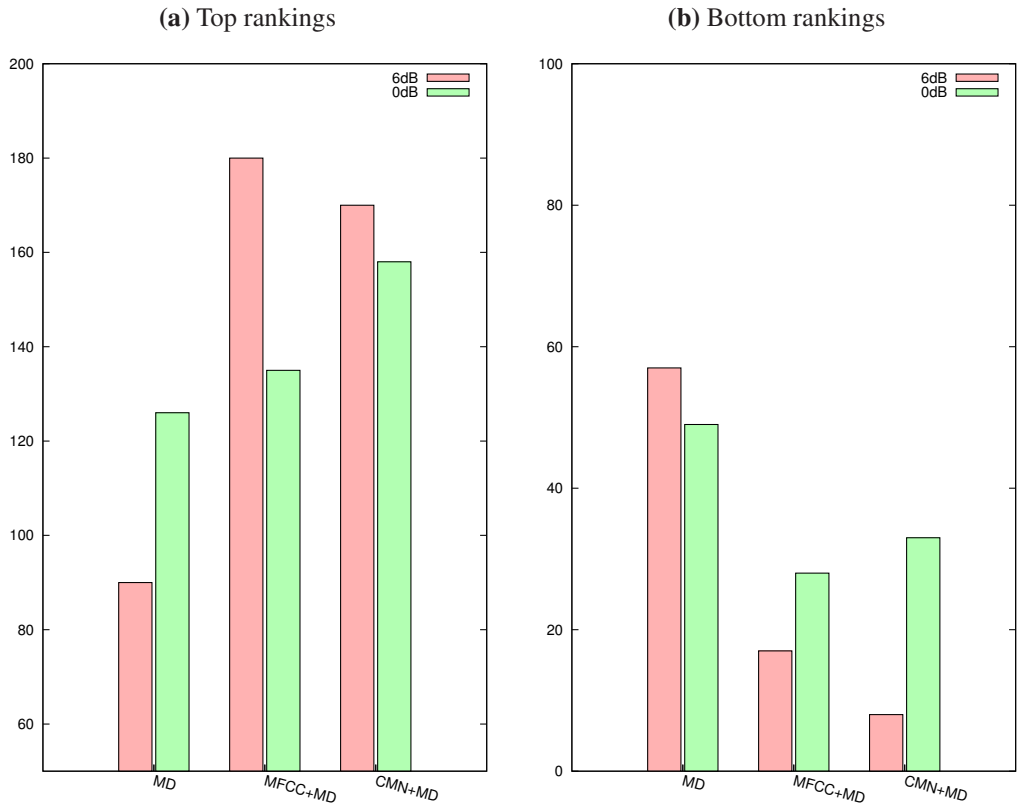
ASR Configuration	**#utterances Top Ranked**	**%Top Ranked**	**#utterances Bottom Ranked**	**%Bottom Ranked**
Missing Data	90	30.0	57	19.0
COR, MFCC+MD	180	60.0	17	5.67
COR, CMN+MD	170	56.67	8	2.67

relative to those of MD spectral based models, the recognition accuracies generated by those recognizers for each utterance were ranked and compared in Table 3.7, and Table 3.8. Implied from these ranking results is the capacity of the fused model to represent the speech process more accurately than that of typical MD theory models. Within the rankings, the top ranked category represents the total number of utterances, 300 in total, that a particular recognizer configuration achieves the highest relative accuracy. Similarly, the number of bottom ranked utterances indicates the number of utterances that a particular recognizer had the

Table 3.8: Stationary Noise 0dB SNR: Recognizer Configuration Rankings Over Entire Test Set, Relative To All Experimented Configurations

ASR Configuration	#utterances Top Ranked	%Top Ranked	#utterances Bottom Ranked	%Bottom Ranked
Missing Data	126	42.0	49	16.33
COR, MFCC+MD	135	45.0	28	9.33
COR, CMN+MD	158	52.67	33	11

Figure 3.13: Stationary noise recognizer configuration rankings



lowest recognition score relative to the other recognizer configurations. From Table 3.7 and Table 3.8, and depicted graphically in Figure 3.13, it appears that the COR model consistently achieves higher recognition accuracies than the MD model. As is expected with this combined model, the relative performance between the fused, MFCC+MD and CMN+MD models and the spectral MD model is not constant over SNRs. This is expected as the fused acoustic model is formed from a probabilistic space that represents the spectral and cepstral parameterization of the speech process. Thus, the performance of the fused model is proportional to the capability of each stream (MD spectral and cepstral) to contain and infer the true speech signal. As the SNR decreases, the true speech content of the cepstral stream deteriorates. The capability of correct speech inference from the spectral, MD, stream however, decreases less rapidly as the SNR decreases(see Table 3.6). The inference capability of true speech from the fused model will therefore not maintain a constant margin of performance gain over the single MD model for differing noise conditions. But, as a result of more accurately representing the speech process than cepstral and spectral models, it is capable of greater recognition performance. Interpreting the rankings from the bottom ranked category further implies this. The rankings depict that the joint space is more capable of the correct inference of W from $\hat{O}$, or rather in terms of MD theory Equation 3.5, inferring W from O_r than the conventional MD process. Essentially, the COR model will perform at least the same if not better than the uncoupled process.

3.11 Combination of Recognizers: Non Stationary Noise Experiments

To further investigate the potential of the proposed methodology for noise robust ASR, experiments were conducted under non stationary noise conditions as outlined in Section 3.9. As outlined in that section, from recognizer configurations *i thru iii* and *v thru vi*, recognition results were generated using the clean speech test set to form baseline results for each recognizer. The tabulated results from this test condition are listed in Table 3.9. As discussed in the previous section, Section 3.10, these clean speech recognition results tend to coincide with the expected comparative behavior of cepstral, spectral and joint models described in Table 3.1.

For the non-stationary test conditions, the recognition results were generated with the recognizer in configurations *ii thru iv* and *vii thru viii*, (Section 3.9), and

Table 3.9: Non Stationary Noise: BaseLine Recognizer Results, Clean Data

ASR Configuration	**Recognition Accuracy %**
Conventional	
Spectral Features, rate32	94.64
MFCC Features	95.15
CMN	81.49
Fusion,Combination of Recognizers	
MFCC+rate32	94.04
CMN+rate32	95.22

Table 3.10: Recognizer Results With Test Corpus + Destroyer Operations Room Noise

ASR Configuration	**Recognition Accuracy %**		
Conventional	SNR 18dB	SNR 12dB	SNR 6dB
Spectral Features, MD	76.68	73.51	67.63
MFCC Features	82.79	74.52	65.12
CMN	67.07	62.64	60.31
Fusion,Combination of Recognizers			
MFCC+MD	83.86	77.22	68.36
CMN+MD	87.96	81.88	74.17

test sets representative of SNR 18dB, 12dB and 6dB conditions. Table 3.10 and Table 3.11 reflect the recognition results over all recognizer configurations for the destroyer additive noise source and the factory conditions respectively.

Over all tested non-stationary noise conditions for both the additive *destroyer* and *factory* sources, as well as for all tested SNRs, the COR experiences greater recognition accuracies than those of all other tested recognizer configurations. These results also exhibit typical characteristics of cepstral and spectral based

Table 3.11: Recognizer Results With Test Corpus + Factory Noise

ASR Configuration	**Recognition Accuracy %**		
Conventional	SNR 18dB	SNR 12dB	SNR 6dB
Spectral Features, MD	76.5	73.3	67.4
MFCC Features	83.7	73.9	64.7
CMN	66.0	61.6	60.3
Fusion,Combination of Recognizers			
MFCC+MD	84.5	76.7	67.5
CMN+MD	88.6	81.8	73.5

models. In particular the relationship between recognition with conventional cepstral based ASR and the MD pattern recognition process with marginalization(Section 3.3. As is evident in the recognition performance of those two recognizers, ratemap + MD and MFCC, as the SNR decreases, the recognition accuracies of the MD model gradually increases relative to that of the cepstral model. This behavior may be explained by considering each respective acoustic models' ability to infer W as the SNR of $\hat{O}$ decreases. As the adverse condition worsens, $\hat{O}$ differs to a greater degree with that of which the model, θ, represents, O. The MD based model bases recognition on the decomposed signal, namely O_r whereas the cepstral model infers directly from $\hat{O}$. Thus the cepstral based ASR experiences a decline in recognition accuracies as the SNR decreases while the magnitude of the decline is less for the MD based system.

The COR configuration, MFCC+MD, results from both tested additive noise conditions exhibits greater recognition accuracies than those of the conventional ASRs. The benefit of the fused process for recognition performance is to a lesser extent under the 6dB condition though under all test conditions the relation between the capacity of the fused model to that of the typical models still holds,$H\left(\hat{O}_{(c)}\, O_{r_{(s)}}\right) \geq H\left(O_{r_{(c)}}\right)$ (Table 3.1). Though the normalized model, CMN, exhibits a degraded characteristic under clean conditions[67], when it is combined in a COR configuration, the resultant joint space is capable of effective speech recognition in all tested conditions. The joint space, as a result of coupling a normalized cepstral and a spectral process, may be effective in its inferences due to the inherent

Table 3.12: Destroyer Noise 18dB SNR: Recognizer Configuration Rankings Over Entire Test Set, Relative To All Experimented Configurations

ASR Configuration	#utterances Top Ranked	%Top Ranked	#utterances Bottom Ranked	%Bottom Ranked
Missing Data	97	17.32	169	30.8
COR, MFCC+MD	217	38.75	56	10
COR, CMN+MD	389	69.46	28	5

attributes of each of its components. The cepstral based process, $\hat{O}_{(c)}$, in this instance, is normalized, Equation 3.1, to match the acoustics of the model, while the spectral based process, $\hat{O}_{(s)}$ is segregated resulting in O_r for pattern recognition. The combination of these two features within the fused construct may reduce the discrepancy between $\hat{O}$ and that of which the model, θ, represents, O. The fused effect of combining noise features has also been observed in other works such as[40].

To further analyse the effectiveness of the COR configurations over those of the other recognizers, the performance of the MD recognizer and the COR recognizers were ranked for each additive noise source and SNR condition. Out of the 560 utterance test corpus, the percentage that a particular ASR configuration has the highest recognition accuracy in relation to all others as well as the lowest for each noise source and all SNRs are given in Tables 3.12 thru 3.17. These tables that reflect the relative effectiveness of each recognizer configuration, for the each of the test conditions, further illustrate the effectiveness of the combination of recognizers for robust speech recognition. Over all tested configurations, the COR method, generally outperforms the MD recognizer as evident in the greater percentage of top ranked utterances for the COR configurations over that of the single model recognizer. Another observation to be taken from the tabulated results is that the COR method does not appear to degrade recognition performance as this is inferred from the COR possessing the lowest percentage of bottom ranked utterances. These attributes of the COR configuration is especially evident in

Table 3.13: Factory Noise 18dB SNR:Recognizer Configuration Rankings Over Entire Test Set, Relative To All Experimented Configurations

ASR Configuration	#utterances Top Ranked	%Top Ranked	#utterances Bottom Ranked	%Bottom Ranked
Missing Data	109	19.64	161	29.01
COR, MFCC+MD	230	41.44	30	5.41
COR, CMN+MD	390	70.27	16	2.88

Figure 3.14 for the destroyer noise test condition. From this figure, the spectral, MD, model consistently has a higher number of bottom ranked utterances than the COR models. This together with the top rank tallies imply that the fused model is more resilient than the spectral model alone. This discrepancy may be attributed to the strength of the fused models's capacity to model the speech process and the inherent exploitation of the inferred statistical dependencies between the cepstral and spectral stochastic processes. Another observation, from the both the tabulated and the graphical rankings, is that in comparing the two fused models, the recognition performance of the normalized model declines less rapidly than that of the non normalized combined model. This implies that the normalization of the cepstral process may add to the resiliency of the joint space to noise perturbations.

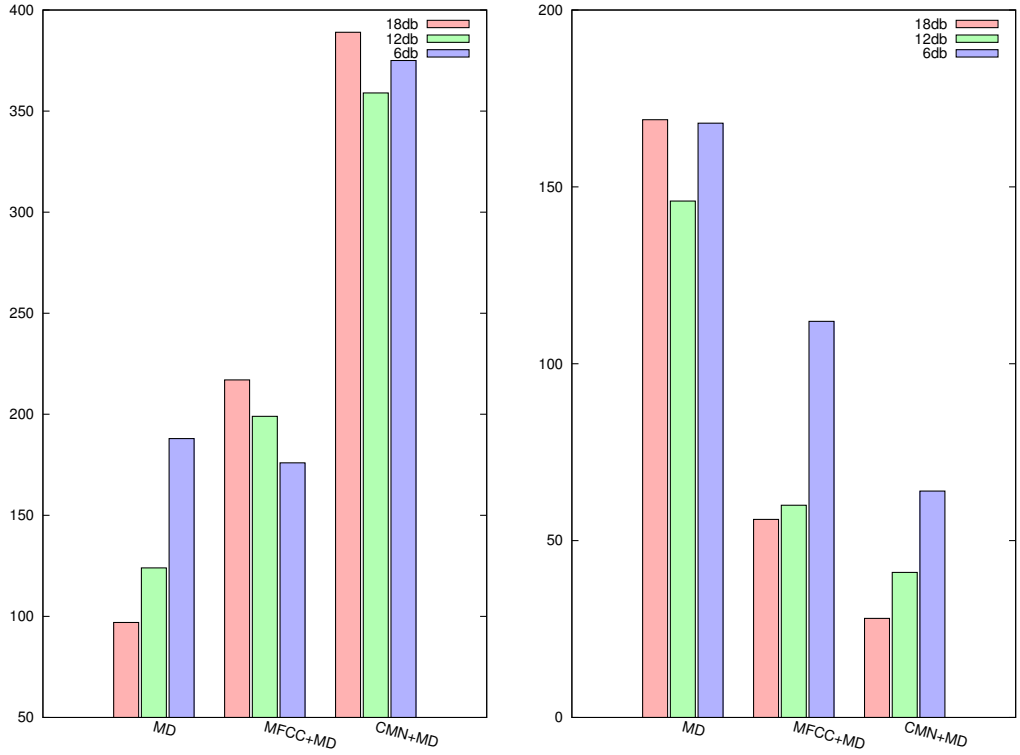
The joint space, as a result of decision level fusion of spectral and cepstral stochastic processes appears capable of achieving noise robust ASR for non stationary adverse conditions. This acoustic modeling technique, both theoretically and experimentally implies that within its strength to model and represent the speech process and its method of inferring words from a noise corrupted signal lies an effective methodology for ASR. Unlike HMM decomposition, and PMC acoustic model based noise robust techniques that model both speech and noise, the COR methodology is based on, though certainly not confined to, speech models. Here a noise robust pattern recognition method is devised from a coupled model of MD and cepstral stochastic processes. Where the statistical dependencies between the speech segregated MD process and the cepstral process are prop-

Table 3.14: Destroyer Noise 12dB SNR: Recognizer Configuration Rankings Over Entire Test Set, Relative To All Experimented Configurations

ASR Configuration	#utterances Top Ranked	%Top Ranked	#utterances Bottom Ranked	%Bottom Ranked
Missing Data	124	22.14	146	26.07
COR, MFCC+MD	199	35.54	60	10.71
COR, CMN+MD	359	64.11	41	7.32

Table 3.15: Factory Noise 12dB SNR: Recognizer Configuration Rankings Over Entire Test Set, Relative To All Experimented Configurations

ASR Configuration	#utterances Top Rated	%Top Ranked	#utterances Bottom Ranked	%Bottom Ranked
Missing Data	142	25.59	139	25.05
COR, MFCC+MD	193	34.77	60	10.81
COR, CMN+MD	383	69.01	29	5.23



(a) Top rankings **(b)** Bottom rankings

Figure 3.14: Destroyer noise recognizer configuration rankings

Table 3.16: Destroyer Noise 6dB SNR: Recognizer Configuration Rankings Over Entire Test Set, Relative To All Experimented Configurations

ASR Configuration	**#utterances Top Rated**	**%Top Ranked**	**#utterances Bottom Ranked**	**%Bottom Ranked**
Missing Data	188	33.57	168	30
COR, MFCC+MD	176	31.43	112	20
COR, CMN+MD	375	66.96	64	11.43

Table 3.17: Factory Noise 6dB SNR: Recognizer Configuration Rankings Over Entire Test Set, Relative To All Experimented Configurations

ASR Configuration	**#utterances Top Rated**	**%Top Ranked**	**#utterances Bottom Ranked**	**%Bottom Ranked**
Missing Data	189	34.05	181	32.61
COR, MFCC+MD	161	29.01	138	24.86
COR, CMN+MD	368	66.31	54	9.73

agated throughout the pattern recognition topology. That together permit noise robust ASR without apriori knowledge of the noise source. Within the MD framework, there exists another promising technique for robust recognition, the segment fragment decoder[5]. Whereas this fragment decoder finds the best hypothesis word sequence and segregation of noise from speech, the COR method finds the best word sequence from an expanded Viterbi search space within the fused decoder. Akin to the SFD that demonstrated significant increased recognition under non stationary noise conditions, the COR technique appears to signify the same traits.

Chapter 4

Optimizing the Maximum Likelihood Estimator: A Large Margin Approach

Book III

Discriminant ML Estimator

In so far as a scientific statement speaks about reality, it must be falsifiable; and in so far as it is not falsifiable, it does not speak about reality. — Sir Karl Popper

The central premise of this chapter follows that of the previous in devising optimal[1] stochastic acoustic models. Here, unlike with a coupled topology described previously, the observation distribution of the speech process is modeled using discriminative classifier methods. Through optimizing the ML estimator, the relation, first presented in Chapter 2, $H(O_n \mid U_n)$ is described. Whereas the previous chapter described maximizing this relation, or rather, increasing the speech acoustic content of the model through a coupled topology. Here an optimal ML estimator is devised using large margin discriminative classification techniques.

The parameterization of the speech process can be taken as a set of measurements of the speech signal that form a continuous time series. Modeling the stochastic characteristics of this time series can be considered to be paramount to effectively infer words from speech signal. This inference of words, can be realized from a set of models that describe the acoustic space. An acoustic space that is segmented to represent each possible word or sub word unit that can be expected to be inferred from the speech signal. The segmentation of the speech process should take a form that ensures speech process may be represented within it without loss of information content. Or at least, such as to minimize the loss. This chapter establishes a methodology to model the speech recognition problem as that of inference from stochastic models formed with discriminant functions. Encompassed within this is the mapping of the acoustic space to one that is capable of encoding the speech process and is, furthermore, appropriate to form discriminant decision rules, or boundaries.

Within this chapter, the methodology is developed through a series of sections. An appropriate segmentation of the acoustic space for discriminative classification techniques is first analyzed in terms of satisfying the necessary conditions to effectively model the speech process. Amongst the findings from the analysis is a ***likelihood estimator*** that can describe the input signal. The resultant acoustic space permits the mapping of the inference problem to one that can not only lend itself to strong classification techniques but also unhindered from loss of information content. This serves as the motivation for the presented methodology. To devise discriminatively trained acoustic models that are capable of modeling

[1] Here, optimal is in the sense of minimizing the distance between the probabilistic distribution of the speech process and its estimate

the speech process. The speech inference problem is subsequently described in terms of posteriors and as such in a form suitable for discriminatively trained stochastic models. Such models may take on one of several approaches. Several of these are described including those that form decision boundaries with linear combinations[43], density estimations[38] and neural networks[64], NNs. The large margin approach, support vector machine[68], is then reasoned to be most suitable for this speech inference problem and to optimize the likelihood estimator.

This chapter is organized as follows.

- Acoustic space: ***How to segment the acoustic space so that it can be effectively modeled using discriminative training methods?*** The speech process is presented as sequential measurements of the speech signal that form a continuous time series. Using the analysis techniques developed and described in Chapter 2, specifically Section 2.2, a segmentation of the acoustic space is determined that not only is capable of modeling the speech process without loss but also lends itself to modeling with discriminative trained acoustic models. Specifically, these acoustic models are based upon representing the speech observation process with ML estimators. It is furthermore shown, that the capability of modeling the acoustic space accurately is a function of both the number these estimators and the accuracy of the estimators to model the process. The former of the two can be expressed in terms of the latter. Thus, by formulating the ML estimators in a manner for which discriminative techniques may be applied the speech process can be modeled effectively. Moreover, in using discriminative techniques, optimal[2] discriminatively trained classifiers may be formulated to produce *optimal* ML estimators.

- Discriminative Techniques: ***There are a number of discriminative training methods that may be used to model the speech process. Which of these methods may be well suited for this problem?*** Several discriminative techniques are described in terms of its construction and its capability of modeling the speech process. Each of these methods are assessed on how they construct the decision function that separates data. How to train the model. And how to control the complexity of the model. It is shown that

[2]Optimal in the sense of minimizing the error between the true observation distribution and its estimate

the large margin discriminative method may be better suited than the others to model the speech process with the proposed ML estimators.

- Discriminative Speech Modeling: Past and Present: ***There have been several research efforts to model speech with discriminatively trained acoustic models. How does the proposed methodology differ from those?*** This proposed methodology differs from other discriminative acoustic modeling approaches as it is primarily devised to effectively model the speech process. Effective in the sense that the acoustic model is capable of modeling the speech observation distribution without loss of acoustic information content. Previously established approaches that have used discriminatively trained acoustic models have experienced problems that have hindered this realization. Some of these setbacks have been noted in this section. Setbacks such as the inability to effectively classify speech at the frame level. As well as not effectively modeling the transient behavior of the speech signal. This methodology shows that by using large margin discriminative training methods to maximize the objective function of Equation 2.16, $H\left(O^{(n)}|U^{(n)}\right)$, optimal ML estimators can be formulated. These optimal stochastic models, moreover, are shown to effectively model the speech process.

- SVM Discriminant ML Estimator: ***How to construct these so called optimal ML estimators using the large margin discriminative method? Just how effective are these models?*** This section formulates modeling the speech process with discriminatively trained ML estimators using the large margin training method. Here the acoustic space is mapped in terms of ML estimators so to maximize the speech content in the resultant model. Large margin discriminatively trained models are formulated to model these estimators. It is shown that the resultant models are not only optimal, but also maximize the entropy of the observation distribution. Thus, theoretically, they are capable of modeling the speech process without loss.

- Large Margin Discriminant ML Estimator: Experimental Results: ***Are the empirical results consistent with the theoretical expectations?*** The proposed methodology was tested experimentally. Large margin based acoustic models are trained and tested with the Grid speech corpus[18]. The results from the experiments indicate that the optimal stochastic models can effectively model speech.

4.1 Acoustic Space

A possible segmentation of the acoustic space, or rather, a segmented representation of the information content of the speech signal, that is suitable to model the speech process may be found in the following manner. Consider the signal in a parameterized format with observations, $O^{(n)}$, as in Figure 1.2 and Equation 2.1,

$$O^{(n)} = [O_1, O_2, O_3, \ldots, O_n]^T$$

where,

$$O_i \in \mathscr{R}^m$$

As such, each measurement, O_i, in this form can represent a speech frame. The relative information content between each measurement can be examined through information theoretic concepts when the parameterized representation is rewritten as in Equation 2.2,

$$Z_j^{(n)} = [O_{1j}, O_{2j}, O_{3j}, \ldots, O_{nj}]^T$$

where, $j \in \{1 \ldots m\}$. The KL divergence distance measure may be used for this purpose. Examining each of these, ***rvs***, Z_{ji}, with respect to each other in the form of $KL\left(p\left(Z_{ji}\right)\, p\left(Z_{lk}\right) \parallel p\left(Z_{lk} Z_{ji}\right)\right)$ reveals the correlation, or transient behavior, between each multivariate measurement. With this form the relative information content, $I\left(Z_{ji}, Z_{lk}\right)$, of the signal can be expressed by Equation 2.5 and equivalently, Equation 2.6 and Equation 2.7.

The temporal behavior, or transient aspect, of the signal, or time series, may be modeled as a Markovian process. As described in Chapter 2, a first order Markov chain formed with, h, ***rvs***, U, $U_1 \rightarrow U_2 \rightarrow U_3 \cdots \rightarrow U_h$, Figure 2.1, is a common incarnation of this process. It can be shown, through information theoretical concepts, that this model is capable of encoding the temporal aspect of the signal, $I\left(Z_{ji}, Z_{lk}\right)$, without loss. This is implied from Equation 2.11,

$$I\left(U_i U_{i+1}\right) \geq I\left(U_i U_{i+2}\right)$$

where the term on the left of the inequality represents the mutual information between ***rv***, U_i, and the successive $(i+1)$ ***rv*** of the chain and the term on the right the relationship between a given ***rv*** at instance, i, and one two time steps ahead. Furthermore, through factorizing the observation distribution, $P(O)$, over n stages,

as is the case with a hidden variable, HMM, stochastic model, Equation 2.10, the true observation distribution can be represented within this Markovian topology as can be reasoned from Equation 2.16,

$$H(O_n \,|\, O^{(n-1)}) \geq H(O_n \,|\, U_n) \tag{4.1}$$

The term of the left of this expression represents the entropy of the true observation distribution and the term on the right, the corresponding entropy of the distribution as modeled with the hidden variable model with observations, O, and a latent variables, U. Thus, Equation 2.11 together with Equation 2.16 imply that the hidden variable topology is capable of modeling the speech process without loss. More specifically, given the parameterized measurements, $O^{(n)}$, the true distribution of the signal, $P(O)$, can be realized from the hidden variable topology through maximizing the ML estimator, or rather, maximizing the expected value of the log likelihood. As such, a possible acoustic space segmentation that is suitable for modeling, one used within this methodology, is through factorizing the acoustic space over multiple discrete hidden variable states, U.

This acoustic space segmentation realized in the hidden variable topology effectively models the speech process as evident from its representation of the observation space and the transient qualities of the time series. Though this form of acoustic model is capable of capturing the speech process, to ensure it encodes the signal without loss, both the observation space must be well defined and there should be a sufficient number of hidden states.

The number of states, U, that may be necessary for this purpose can be expressed in the following manner. The entropy rate can be defined as the rate of growth of entropy[22],

$$H(\chi) = \lim_{n \to \infty} H(X_1 X_2 X_3 \ldots X_n) \tag{4.2}$$

with, n, ***rvs***, X. So with m_U hidden variables and n realizations, the entropy rate, is,

$$H(\varphi) = ln(m_U) \tag{4.3}$$

This expression implies that $H(U_1 U_2 \ldots U_n)$, or rather, $H\left(U^{(n)}\right)$ converges to the entropy rate and that $H\left(U^{(n)}\right) \geq H(\varphi)$. Similarly, this entropy rate can be

expressed in terms of the observation distribution, $P(O)$, such that,

$$H(\varphi) = ln(m_U) \geq KL\left(p(O_n)\, p\left(O^{(n-1)}\right) \parallel p\left(O_n O^{(n-1)}\right)\right) \tag{4.4}$$

where the term on the right of the inequality is the KL divergence between the joint distribution of an observation at instance, n, and all previous observation realizations and the product of the two. Equation 4.4 may be represented in terms of entropies and in turn with respect to the entropy rate of O, $H(\vartheta)$, as in,

$$\begin{aligned} & ln(m_U) \geq H(O_n) - H\left(O_n \,|\, O^{(n-1)}\right) \\ & \implies m_U \geq 2^{KL\left(p(O_n)\, p(O^{(n-1)}) \| p(O^{(n)})\right)} \\ & or, \\ & I\left(O_n U^{(n)}\right) \geq I\left(O_n O^{(n-1)}\right) \\ & \implies m_U \geq 2^{H(\vartheta)} \end{aligned} \tag{4.5}$$

which both can be implied from the relation described in Equation 4.1. The given expressions of the entropy rate of the hidden variables, U, in terms of the KL distance measure and the entropy rate of O, describe the number of hidden states necessary to encode the speech process without loss. Within the expression of this bound is the relationship between the number of hidden states and the observation distributions. With respect to a generative stochastic process, this distribution is evaluated as the likelihood, or in other words, the HMM emission densities, $P(O|U)$. Therefore, the capability of the stochastic model to represent the speech process depends on both the accuracy of the densities and the segmentation of the acoustic space. Essentially, the ML estimator[3] on the right hand side of Equation 4.1.

The HMM topology[58][8] is defined with ***rvs*** U and O. Moreover, its capability of encoding the signal is dependent upon both the ML estimator, $f(O|U)$, and the number of hidden states that is also dependent upon the former. Modeling this ML estimator is traditionally done using generative techniques whereby the likelihood of a hidden state, U, having generated a given observation, O_i, is estimated. In this case parametric densities are commonly used. As opposed to

[3]Recall that the objective function, Equation 2.16, represents a relation between the true observation and its estimate. This estimate, expressed as $H(O|U)$, is the expected value of the log likelihood. Thus in terms of ***rvs*** O and U, it represents the ML estimator.

estimating the likelihood from a parametric density, the inference may be done using discriminative techniques. These methods estimate decision functions, or decision boundaries that separate the data. More specifically, the separation is formed with decision boundaries that segregate the data, or observations, into distinct classes, k. The decision boundary itself defines the outcome, y, of a given observation, x, belonging to one class, k_i, or another, $k_{j\neq i}$. In other words, given a joint space, $f(x,y)$, the decision boundaries determine $f(y = k\,|\,x)$. Or in terms of probabilistic measures, $P(y = k\,|\,x)$. In this form, each outcome is formed from a decision that considers and accounts for a given observation belonging to each possible class within the joint space. This is in contrast to the density estimation method whereby the resultant decision is a confidence, or a likelihood measure of an observation event occurring given the models' sample space. Put another way, discriminative methods may be construed as estimating the *posterior*, $P(U\,|\,O)$, whilst generative inference is forged from $P(O\,|\,U)$. Furthermore, presenting the ML estimator problem as that of its posterior form opens the maximization problem, $H(O\,|\,U)$, to the use of powerful discriminative classifiers to determine it from. And so to model the speech process without loss within an HMM topology. The maximization of the emission densities, in accordance with Equation 4.1, can now be evaluated using a variety of optimization techniques that may provide a better estimate of $H(O\,|\,U)$ than having the problem limited to conventional ML methods.

Therefore, in modeling the speech process in this manner, the problem may be cast as one that lends itself to discriminative techniques and is capable, through maximizing the ML estimator, to encode the speech process without loss. As such the acoustic space is mapped to that consisting of transitional elements from time step or stage, $i-1$ to i, $f(U_i\,|\,U_{i-1})$, $i \in \mathscr{I}$, and posteriors, $f(U\,|\,O) = \frac{f(O|U)f(U)}{f(O)}$. With each state, $U_k, k \in \mathscr{I}, 1 \ldots h$, represented by a discriminative classifier. These resultant discriminative models, through optimization techniques[76][45][29], maximize the expected value of the log likelihood of the hidden variable, y, generating an observation, x. And in doing so, maximize the entropy of the system.

A number of different approaches may be considered when the problem of determining the ML estimators, or rather the modeling of the observation distribution, is posed as a discriminant analysis problem. The proceeding section provides a comparison of such methods and offers insight as to why the large margin[69] discriminant analysis approach is suitable for this task.

4.2 Discriminative Techniques

Describing or modeling the behavior of a times series, or in particular, the observation space of the speech signal can be done using discriminative techniques. Here, the problem is approached as that of separating the data into likewise components or classes sharing, or describing, some particular aspect or pattern in the signal. Likewise, inferring patterns from these resultant classes, or models, is now a pattern recognition classification problem where the inference is the outcome of which class the pattern is the closest to.

In terms of modeling the ML estimator, $f(x|y)$, in terms of ***rvs*** x, y, the expression can be written in a posterior form,

$$f(y|x) = \frac{f(x|y)\, f(y)}{f(x)} \tag{4.6}$$

where, $f(y)$, is the classifiers' priors and $f(x)$ is a normalization factor. Both are defined by the training input space.

The modeling of these posteriors can be done in many differing manners, though the underlying motivation to describe the data in terms of decision boundaries is common to all. Given the parameterized signal $\mathbf{x}$,

$$\mathbf{x} = [x_1, x_2, \ldots x_m] \tag{4.7}$$

a common approach to separate the data is to form decision boundaries with a linear combination of all components of the signal. Such a decision boundary may be expressed as,

$$\sum_{i=1}^{m} x_i a_i \tag{4.8}$$

In general over multiple input observations, $\mathbf{x} = [x_1, \ldots, x_n], x_i \in \mathscr{R}^m$, the decision boundary forms a hyperplane Figure 4.1.

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \tag{4.9}$$

The decision boundaries formed in this manner define the outcome, y, or rather the inference of an event from a given sample space. In other words, given the input space, $f(x)$ and its outcome, $f(y)$, that together form the joint space, $f(x, y)$, the decision boundaries realize the outcome of an event, $x_i, i \in \mathscr{I}$, from $f(x, y)$.

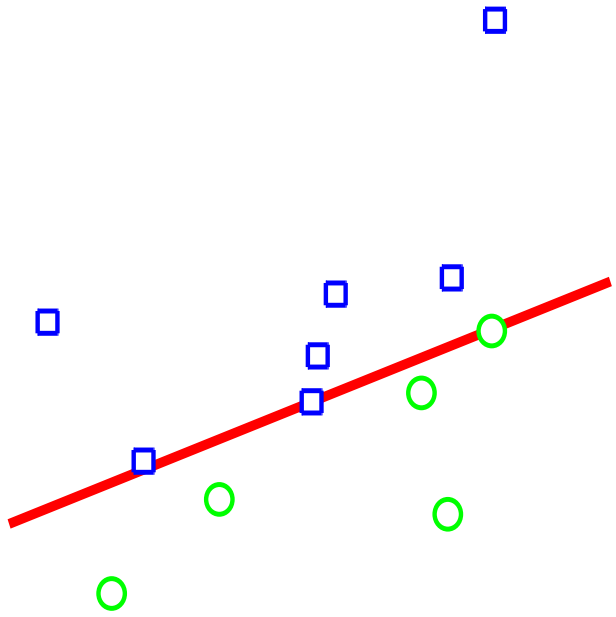


Figure 4.1: Linear combination decision boundary: Generalized hyperplane, $\mathbf{Ax} = \mathbf{b}$

In terms of a two class discriminant problem, $f(y) \in \{-1, 1\}$ and is positive when $f(y|x_i) = x_i$. The values that y take on are *indicative* of the outcome of a particular input x. In general for a k class problem, the indicator response, y, to the input x may be expressed as,

$$f(y) = \begin{cases} 1 & , \quad y = k|x \\ -1 & , \quad y \neq k|x \end{cases} \tag{4.10}$$

These indicator responses, together with the inputs form the joint space, $f(x,y)$ from which the decision functions or boundaries are determined. This process is illustrated in Figure 4.2. Therefore each input, x, is mapped to an response, y, that is indicative of which class, x, belongs to. From the resultant joint space, $f(x,y) = f(x)f(y|x)$, decision functions may be considered to be an estimate of $\hat{y} = f(y|x)$. Furthermore, in terms of specific classes, k, this may be expressed as $f(y = k|x)$.

The loss function, $L()$, or loss functional[69], expresses the relationship between the indicator response for a given, x, and the estimate formed from the decision function, $\hat{y}$. As such, the decision boundary problem, may be expressed

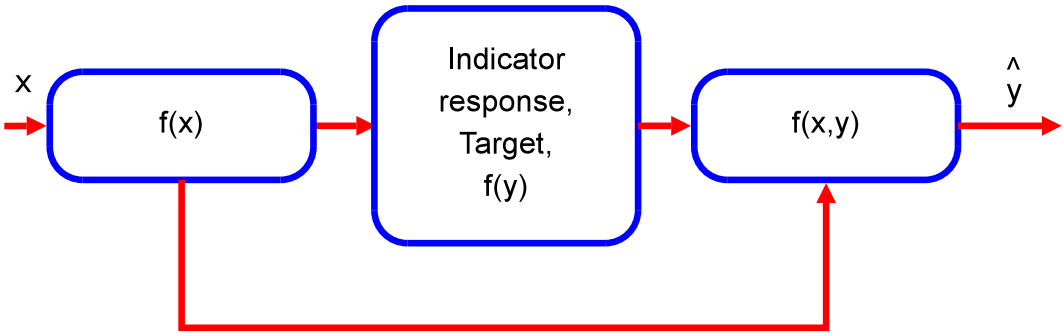


Figure 4.2: Decision functions: Given the input space, $f(x)$, and the corresponding indicator response, $f(y)$, decision functions are formed from the joint space $f(x,y)$

in the following,

$$\begin{aligned} L &= (y - \hat{y}) \\ &= (f(x) - f(\hat{x})) \end{aligned} \tag{4.11}$$

In this form, it can then be considered to be a minimizing the loss optimization problem. Through this relation the posterior form of the ML estimator can be described in terms of determining the optimal decision functions that satisfy minimizing the loss function. This interpretation of the ML estimator lends itself to numerous loss functional representations such as the L2 distance measure,

$$L = (f(x) - f(\hat{x}))^2 \tag{4.12}$$

and the L1 distance measure,

$$L = (|f(x) - f(\hat{x})|) \tag{4.13}$$

Each with its own merits and drawbacks. For instance it may be argued that the L1 distance measure is much more robust than the L2 distance measure though the L2 distance measure may be more appropriate for some tasks. Similarly, the loss functional, $L()$, lends itself to a multitude of optimization techniques. Each such method can be assessed by the trade off between computational complexity and its appropriateness for the nature of the problem.

The selection of the most appropriate discriminative method for a given prob-

lem may be assessed in terms of the risk loss functional[68], $R(\alpha)$,

$$R(\alpha) = R_{emp} + \Phi\left(\frac{l}{h_k}\right) \tag{4.14}$$

where, the empirical risk, R_{emp} can be construed as minimizing the loss and $\Phi\left(\frac{l}{h_k}\right)$ the confidence measure. Together these control the accuracy and complexity of the decision function. This relation describes the generalization capability of the decision functions. Generalization in the sense of describing, or classifying observations that were not used in the formation of the decision boundaries. The accuracy and the generalization capability of the decision functions is described, in Equation 4.14, as a trade off between minimizing the error whilst ensuring that the structure, or complexity of the decision functions, is not bound to the empirical data used to form the boundaries. Likewise, the variance bias[39] relationship may also provide insight to the relationship between the error and generalization behavior of decision functions. For decision functions formed from linear combinations, this view of the boundaries describes the complexity of the model through both the loss and the deviation between the true and estimated boundary.

Discriminant analysis, therefore, can be considered as a problem of separating data with decision boundaries formed from the joint space $f(x,y)$. Through both the indicator responses, y, and the inputs, x, decision functions are formed as an estimate of $f(y = k|x)$. This problem expressed through the loss functional can then be approached as an optimization problem to minimize the loss. With this established, a discriminant ML estimator, Equation 4.6, may be determined in a number of different manners. The following subsections describe and analyze some common techniques and describe each in terms of the speech inference, or rather, the ML estimator problem.

Logit Regression

Decision boundaries are commonly formed using regression techniques. One may consider this technique as one that distinguishes between inputs, x, through decision functions that minimize the squared error between the true boundary and the estimated boundary. This boundary is formed as a linear combination. Thus the solution is the general form of the hyperplane, Equation 4.9. As such, the loss optimization problem can be expressed in terms of the L2 loss, Equation 4.12, with

a solution that follows that of,

$$\begin{aligned} &\min L_2 \left(y - f(x)\right)^2 \\ &\equiv \frac{d}{dx}\left(\mathbf{y} - \mathbf{A}\,\mathbf{x}\right)^T \left(\mathbf{y} - \mathbf{A}\,\mathbf{x}\right) \\ &= \mathbf{A}^T\,\mathbf{y} - \mathbf{A}\,\mathbf{A}^T\,\mathbf{x} = 0 \\ &\implies \mathbf{x}^* = \left(\mathbf{A}^T\,\mathbf{A}\right)^{-1}\mathbf{A}^T\,\mathbf{y} \\ &\mathbf{A} : \mathbf{a} \in \mathscr{R}^n, \mathbf{x} \in \mathscr{R}^m, \mathbf{y} \in \mathscr{R}^n \end{aligned} \tag{4.15}$$

and,

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{A}\,\mathbf{x}^* \\ \hat{\mathbf{y}} &\in \mathscr{R}^n \end{aligned}$$

This expression describes the estimate of best fit of the model. The decision function formed in this fashion can be used to classify a given input x into one of k classes. Each component of the coefficient matrix, $\mathbf{y}$, in this case, represents $f(y)$ and thus correspond to Equation 4.10. The decision boundary formed in this fashion can be viewed, in the case of a two class problem, as one where $f\left(\hat{\mathbf{x}}_{k_1}\right) = f\left(\hat{\mathbf{x}}_{k_2}\right)$[4], or in general, $\hat{\mathbf{y}} = 0$.

Through the relation of Equation 4.15 the ML estimator can be determined as in Equation 4.6 where the conditional expression is $f(y\,|\,x) = \hat{y}$. The probabilistic form of $f(y = k\,|\,x)$ is typically determined using a *logit* transformation, where for a binary probabilistic space, $log\left(\frac{p}{1-p}\right)$,

$$p\left(y = k\,|\,x\right) = \frac{e^{f(x)}}{1 + e^{f(x)}} \tag{4.16}$$

and,

$$p\left(y \neq k\,|\,x\right) = \frac{1}{1 + e^{f(x)}}$$

Training the model parameters, θ, for the logit regression discriminative clas-

[4]inputs that belong to class k_1 are equivalent to those that belong to k_2

sifier is commonly done using the maximum likelihood technique,

$$\sum log\,(P_\theta) \tag{4.17}$$

with optimization based techniques such as steepest ascent[57], conjugate gradient[57] and iterative techniques[25]. This realization of decision boundaries does have its challenges. Finding the best fit of the model to the data may not be possible if the input space is not suitable to find a global maximum. In this case, data smoothing may be required or some sort of feature selection. Moreover, controlling the trade off between the error rate and the complexity of the model may be difficult due to the underlying L2 distance measure used to fit the model. Furthermore, the model may also require a substantial number of training inputs, x, to increase the models' generalization capability.

Linear Discriminant Analysis

Linear discriminant analysis is another discriminative technique that may be considered. It, like regression analysis, forms its decision boundaries by satisfying the relation, $f\,(\hat{\mathbf{x}}_{k_1}) = f\,(\hat{\mathbf{x}}_{k_2})$. Though, unlike regression, the decision function is forged from Gaussian distributions. As such, the density describing the class, k can be written as,

$$p\,(x) = \sim N\,(\mathbf{x}\,|\,\mu\,\Sigma)\,,\; x \in k \tag{4.18}$$

where, μ and Σ are the mean and covariance of the distribution respectively. The input space then can be represented as a mixture model with mixture weights, π, and a common covariance matrix Σ,

$$p\,(x) = \sim \sum_k \pi_k N\,(\mathbf{x}\,|\,\mu_\mathbf{k}\,\Sigma) \tag{4.19}$$

The indicator responses of this discriminative technique can serve to select a given mixture from the input distribution. Thus $f(y)$ can be a function that results in a binary value, for example $y = 1$ if $x \in \mathscr{S}_k = \{x : f(y = k|x)\}$. From the joint space formed with the marginal distribution of x and the indicator conditional distribution, estimates of the outcome,

$$p\,(y = k\,|\,x) = \frac{\pi_k N\,(\mathbf{x}\,|\,\mu_k\,\Sigma)}{p(x)} \tag{4.20}$$

can be made.

Training such discriminative models is a problem of density estimation and thus is ill posed. Commonly they are trained with the maximum likelihood technique to determine the parameters of Equation 4.17. Typically these models are formed with a common covariance, Σ, parameter, though are not limited to it. Though training the model with a shared covariance reduces its complexity, the maximization of the objective function may not lead to a global maximum. Furthermore a sufficient number of input samples, x may be necessary to accurately model the parameters.

NN

Neural networks[49], NN, form decision boundaries in a manner akin to the synapses and neurons in the brain. Its typical topology consists of multiple layers that connect the base layer, inputs, x, to the top layer, outputs, y. Each layer, within this construct, can be considered to be made up of a linear combination of the layer beneath it. As such, inputs that form the base layer feed into the layer above it. A linear combination of these inputs make up each node, z, of the upper layer. This process may be expressed as,

$$\begin{aligned} &\mathbf{z} = h(\mathbf{b_o} + \mathbf{A}\,\mathbf{x}) \\ &\mathbf{z} \in \mathscr{R}^n, \mathbf{b_o} \in \mathscr{R}^n, \mathbf{A} : \mathbf{a} \in \mathscr{R}^n, \mathbf{x} \in \mathscr{R}^m \end{aligned} \tag{4.21}$$

where, $h()$, is a basis function. Each element of $\mathbf{z}$ is commonly referred to as a hidden unit. The outputs, y, are formed, subsequently, as a linear combination of the hidden units and can be expressed as,

$$\begin{aligned} &\mathbf{z}' = \mathbf{b}'_\mathbf{o} + \mathbf{A}'\,\mathbf{z} \\ &\mathbf{y} = \sigma(\mathbf{z}') \\ &\mathbf{z}', \mathbf{b}'_\mathbf{o}, \mathbf{y} \in \mathscr{R}^K, \mathbf{z} \in \mathscr{R}^n, \mathbf{A}' : \mathbf{a}' \in \mathscr{R}^K, \end{aligned} \tag{4.22}$$

where, $\sigma()$ is the *activation* function and, K, the number of outputs of the model. Typically the activation function is a sigmoid function, though a radial basis network can be formed using a Gaussian basis function.

Therefore the NN discriminative method derives decision functions from a linear combination of basis functions that form nonlinear decision boundaries. Moreover, the topology of the model is not limited in the number of layers nor the number of hidden units within it. This *multiple layer perceptron*, MLP, model may

consist of multiple layers of hidden units that together further refine the decision functions. As a result, powerful discriminant functions can be devised with this technique.

Like the previously discussed discriminative models, the ML estimator of Equation 4.6 is determined from the estimated outcome of a given input x. In this case, the estimate comes from the expression, Equation 4.22, the output of the NN topology. As such, $f(y = k|x) = \sigma(\mathbf{z}')$.

The model parameters for NNs are inferred from the data. This inference is commonly done by using the L2 distance measure, Equation 4.12, or rather minimizing the squared error in the *backpropagation*[75] method. More specifically, the coefficient matrices, of Equation 4.21 and Equation 4.22, or weights, are determined by minimizing the error and inferred from the maximizing the likelihood. Though the NN topology is capable of describing complex decision boundaries, it may be difficult to find an appropriate trade off between minimizing the error and maintaining the models' complexity. Satisfying this trade off, the risk loss functional, Equation 4.14, may be further hindered due to the number of parameters necessary to model the decision functions and the optimization technique chosen to determine the parameters.

Large Margin

The large, or maximum, margin[68] discriminant function is formulated to determine the optimal decision surface. Unlike other discriminative methods, the hyperplane formed optimally separates the data. Similar to how the previously discussed techniques devised decision boundaries that minimized the error between the true boundary and its estimate. The maximum margin technique is derived with the same intent, but does so in a manner that ensures the resultant hyperplane **also** maximizes the distance, (the *margin*), between the set of data points on one side of the hyperplane and the other. The hyperplane that satisfies these criteria is a unique or *optimal* decision boundary.

Finding the optimal hyperplane can be expressed as the quadratic optimization problem. The solution to the minimization is a hyperplane with the following constraints,

$$\begin{aligned} \mathbf{A}\,\mathbf{x} + \mathbf{b} &\geq 1,\ y = 1 \\ \mathbf{A}\,\mathbf{x} + \mathbf{b} &\leq -1,\ y = -1 \end{aligned} \tag{4.23}$$

and $\mathbf{A}$ the normal vector. As such the problem may be cast as one of minimiz-

ing a quadratic surface subject to constraints. This representation lends itself to minimizing with Lagrangian multipliers as in,

$$L = C(x) + \alpha\,(A(x) - b) \tag{4.24}$$

where $C(x)$ is a quadratic and α is the Lagrange multiplier that is applied to the linear constraint. Rewriting Equation 4.24 in terms of the Equation 4.23 and the vector **A**, results in,

$$L = \frac{1}{2}\mathbf{A}^T\mathbf{A}\sum_i \alpha_i\,(y\,(\mathbf{A}\,\mathbf{x} + b) - 1),\ \alpha \geq 0 \tag{4.25}$$
$$\mathbf{A}, \mathbf{x} \in \mathscr{R}^m, y \in \{-1, 1\}, i \in \mathscr{I}$$

Minimizing with respect to **A** and b yields the dual form of this expression,

$$W = \sum_i \alpha - \frac{1}{2}\sum_{i,j} y_i\,y_j\,\alpha_i\,\alpha_j\,\left(x_i \cdot x_j\right),\ \alpha \geq 0, \sum_i y_i\,\alpha_i = 0 \tag{4.26}$$
$$j \in \mathscr{I}$$

that is the objective function to be maximized. The expression of Equation 4.26 may be considered as a convex optimization problem and is subject to the Kuhn Tucker conditions.

The Lagrange multipliers play a crucial role in determining the optimal hyperplane. The multipliers resulting from the maximization of the objective function determine the location, or rather, the composition of the hyperplane. Each nonzero value corresponds to the input, x, within the linear constraint (Equation 4.25) that possess the greatest value. The inputs associated with those multipliers are known as *support vectors*. The optimal hyperplane is formed from a linear combination of these support vectors and can be expressed as,

$$\hat{y} = \mathbf{A}\,\mathbf{x} + \mathbf{b},\ \mathbf{A} = \sum_i y_i\,\alpha_i\,x_i \tag{4.27}$$

A remarkable property of the optimization problem, Equation 4.26, is that it only requires the inner product between inputs. This not only reduces the complexity of the problem, but also lends itself to kernel techniques[24] in the event the inputs are transformed with basis functions, $h(x)$. In this case, the objective function is modified, replacing the inner product with a kernel function, $K(x_i \cdot x_j)$. Through this technique, nonlinear decision boundaries can be determined in high

dimensional spaces.

Classification with large margin, support vector machine, SVM, classifiers can expressed in terms of the weights, α of the model, the indicator responses, y, Equation 4.10, and the inputs, x,

$$f(\hat{x}) = \sum_i y_i \alpha_i (x_i \cdot x) \tag{4.28}$$

Rewritten as the estimate of an outcome due to an event, x from the joint space, $f(x,y)$, the conditional outcome with respect to class k may be expressed as,

$$f(y = k|x) = sign\, f(\hat{x}) \tag{4.29}$$

Through this relation the ML estimator can be determined as in Equation 4.6.

The model parameters for the support vector classifiers are inferred from the data through the relation of Equation 4.26. The objective function in this form is presented as a quadratic convex optimization problem. As such, a global maximum can be found. This discriminative method is capable of not only minimizing the error between the true decision boundary and its estimate but also of controlling the complexity of the model. This is due to the make up of the hyperplane. The derivation of the objective function is composed of a quadratic, the normal vector, **A** and a linear constraint. The solution to the problem can be construed as a saddle point. Thus, the hyperplane that results from the solution is one that has the minimum structural risk[69] with the ***largest margin***.

The support vector machine classifier is capable of determining an optimal ML estimator. The support vector classifier has many merits. The decision boundaries formed from this discriminative technique are optimal in the sense of both generalization capability and classification error rate. Furthermore, decision functions, $f(y = k|x) = sign(f(\hat{x}))$ formed from the large margin technique may be more flexible than the rigid boundaries, $f(\hat{x_k}) = f(\hat{x_l})$ found in both the regression and linear discriminant technique. Unlike the other discriminative methods discussed, the large margin technique infers its decision functions from only partial data. Whereas the other techniques discussed formulate decision functions from linear combinations of all inputs, the support vector machine assesses the value of each data point. Only those inputs that result in nonzero weights, Equation 4.26, form the resultant hyperplane. This property implies the training method is robust to outliers and may negate the need for data smoothing and feature selection. The same cannot be stated for the others discussed. Thus, the merits of the large margin technique make it an appropriate choice to form the discriminative acoustic

model ML estimator.

This section established the appropriateness of the large margin classifier for devising a discriminant ML estimator. The ML estimator, taken in its posterior form, Equation 4.6, can be modeled using discriminative training methods. These discriminant estimates pose an alternative to ML estimators inferred with generative techniques. As such, a speech recognition acoustic model such as that described in Equation 2.10 may consist of these *discriminant* ML estimators. Furthermore, an optimal ML estimator, as the large margin trained posterior is posed, is capable of modeling the observation distribution of the speech process without loss, Equation 4.1. The proceeding section describes other efforts to model speech with discriminative learning methods and discusses the novelty of the approach presented in this methodology.

4.3 Discriminative Speech Modeling: Past and Present

There have been numerous efforts to use discriminatively trained acoustic models for speech recognition. Neural networks, for instance, have been commonly used to model the speech process. Using a *connectionist* hybrid architecture[10], NN discriminant functions are used to model the observation distribution of the speech process. Radial basis networks have also been applied using this connectionist structure[64]. The discriminative model observation distribution is used within an HMM topology, thus the term hybrid, for speech recognition. The resultant models have been successfully applied to the continuous speech recognition[41][10] problem though typically perform best when classifying inputs, x, that span over multiple speech frames. More specifically, the discriminative classifiers perform best with inputs that are composed of a concatenation of several samples. In terms of the parameterized signal, this implies the information contained within a given sample may not be sufficient to form a decision boundary and to accurately classify. As is described in Chapter 2, the mutual information, or correlation, between successive speech samples is high, and thus concatenating several adjacent samples may provide a richer representation of the signal to classify from.

Recently, support vector machines have been applied to the speech modeling problem. In these works, the discriminative method has typically been used to model and classify speech patterns. It has been noted that SVMs are very powerful classifiers. In some published articles the SVM classifier was implied to be not suitable for modeling time series[16]. The SVM discriminative tech-

nique has been, though, applied in several works that model the speech process in a similar fashion to that of the connectionist hybrid structure[34][65]. Like the NN connectionist models, some of the research efforts have used windowed samples, concatenated inputs, to both form the discriminant decision boundaries and to classify. Issues generally associated with these discriminatively trained speech models include speech frame alignment problems and the need to reduce the computational complexity. As a result several different methods have been proposed to segment the acoustic space and heuristics have been introduced to reduce the computational requirements.

Like the discriminative techniques discussed in the previous section, mutual information models[44][54] are also discriminative models, as mentioned in Chapter 2, that minimize the KL divergence between the conditional and the observation distribution, $KL\left(P(O\,|\,\theta_i) \parallel \sum_j P(O\,|\,\theta_j)P(\theta_j)\right)$, though they do not minimize the classification error. Due to this they are not investigated any further in this methodology.

This methodology presents a method to model the speech process using discriminative techniques. Specifically the ML estimators within the HMM topology with SVM discriminative classifiers. Unlike other research efforts, the method presented here encodes the speech process at the speech frame level. In other words, this work models the speech process as that of an HMM topology, Equation 2.10. In this model, the correlation within the time series is encoded within the 1st order Markov chain represented by the hidden variables of the model. The relationship between hidden units, together with the ML estimators can represent the time series' properties without loss. This is unlike those works that found it necessary or beneficial to model the mutual information between successive samples within the discriminative classifiers, (use concatenated samples to train and test the discriminative models). The motivation of this work is to use discriminative techniques to model an optimal ML estimator that is capable of satisfying the relation, Equation 4.1 and thus capable to model the speech process without loss.

The following sections describe the proposed method to form an optimal ML estimator to model the speech process. Specifically, using SVM classifiers to model the ML estimator. The proceeding section describes mapping the speech acoustic space into segments suitable for discriminative models to encode the speech process. The process of deriving SVM decision boundaries from these resultant acoustic segments is then explained. This entails devising an appropriate classifier training and classification strategy for this task. The developed techniques to model the ML estimator are then tested experimentally. The description of those tests and their results are detailed in the experiments subchapter section

of this chapter.

4.4 Large Margin Discriminant ML Estimator

Modeling the speech process with discriminative techniques requires the acoustic space to be segmented into speech units that can be used to train the discriminative classifiers. These segments of speech should be capable to represent, or encode, the unique properties of the speech unit that make it distinguishable from all other units. Furthermore, these segments should be formed in a manner that can represent the speech signal without loss. One such segmentation is described in Section 4.1. As described in that section, the acoustic space is represented within the HMM topology, Equation 2.10. Through encoding the transient qualities of the time series within the hidden variables and the observation distribution within the ML estimators, this topology is capable of modeling the speech process without loss. As is discussed in Section 4.1 and implied from Equation 4.1, the capacity of the representation to encode this process is related to the strength of the ML estimator. Moreover, this representation lends itself to represent a choice of word or sub word units such as phonemes[53]. Therefore the focus of this methodology is to devise an optimal discriminant ML estimator. This estimator is formed with discriminant functions defined with the large margin discriminative learning method. Each decision boundary represents a segment of the acoustic space that is segmented by the number of states in the hidden variable topology.

With this choice of segmenting the acoustic space, the number of hidden variables, U, within the HMM topology represents the number of distinct discriminative classifiers per word or sub word unit. As such, each SVM classifier can represent a given ML estimator, $f(O|U)$ as depicted in Figure 4.3. The total number of classifiers necessary to model the speech process in this fashion would be dependent upon the number of symbols within the acoustic space. For instance, an HMM acoustic space consisting of N distinct word or subword units, each composed of M states, would need at least NM distinct classifiers. The total number of classifiers necessary to model the acoustic space is dependent upon the decision strategy employed. Consider the set, S, of distinct $Q = NM$ units to be classified. In the case of the *one versus one* training strategy, each unit, $i, i \in S$, is trained against each unit of S. This results in $Q(Q-1)$ distinct classifiers to model the acoustic space. The *one against all* strategy, though, requires only Q distinct classifiers. With this strategy, decision boundaries represent the separation of one class, i, of S from all other classes, $j \neq i$. For the purposes of this methodology,

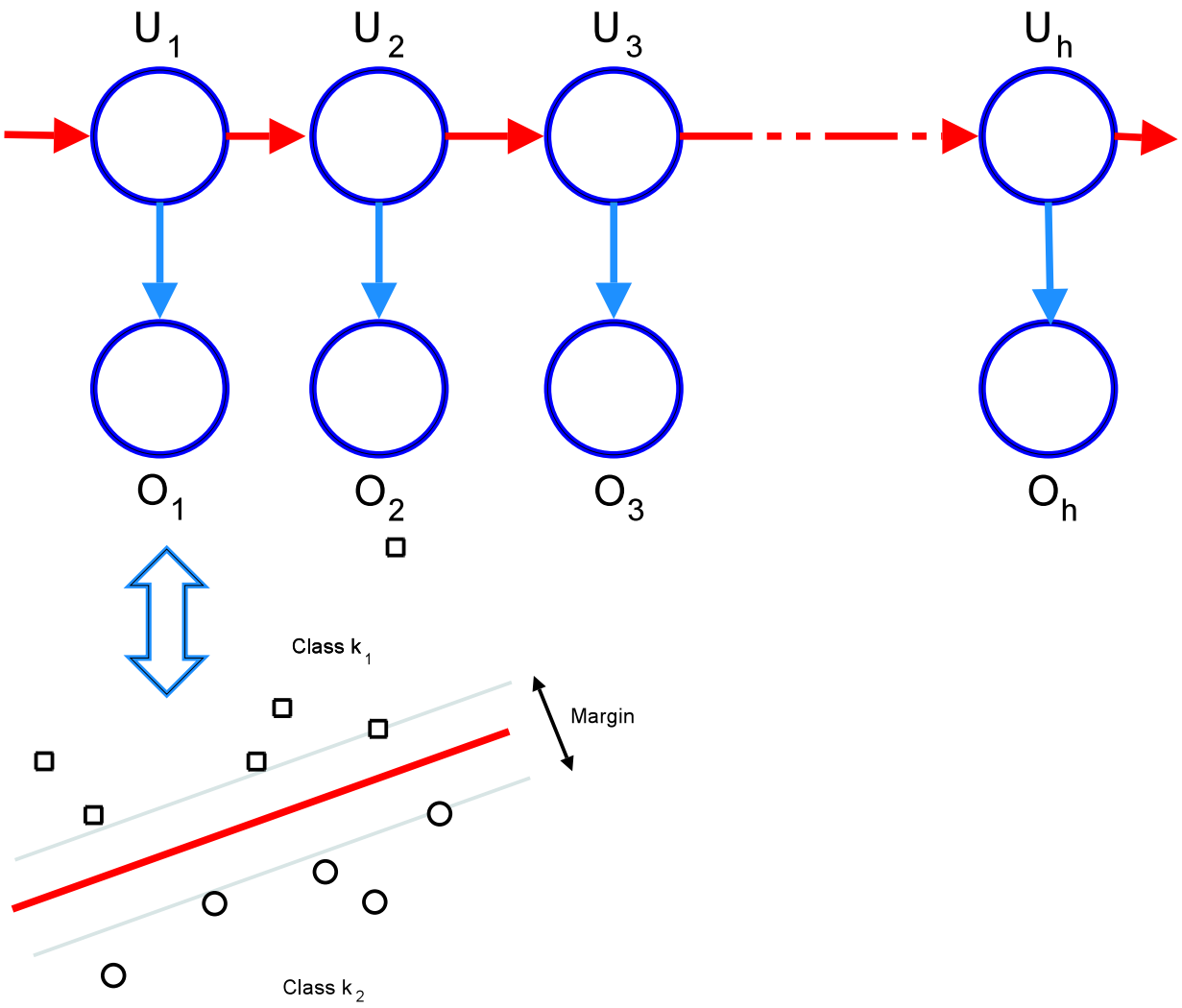


Figure 4.3: Discriminant ML Estimator: Large Margin classifiers are used to model the ML estimator, or emission densities in the HMM topology. Figure depicts the relation between the likelihood, $f(O|U)$ and its discriminative posterior estimate

it can be assumed that the number of classifiers necessary to model the speech process is Q.

The mapping of the HMM structure to large margin classifiers provides a convenient method to *bootstrap* the discriminatively trained classifiers. Existing generatively trained HMM based acoustic models can be used to form the training inputs, x, and indicator responses, y, for the discriminative models. In other words, the generatively trained models can time align and label parameterized input vectors to form the pairs (x_i, y_i). The set of input and indicator response pairs can then be used to form the joint space $f(x, y)$, Figure 4.2, from which discriminative classifiers infer their decision functions.

Support vector discriminative classifiers can be used to model the ML estimators within the HMM topology. Given the input space, $f(x)$, and the indicator responses, $f(y)$, the optimal hyperplane can be devised that separates inputs that belong to one class, k_i from those of other classes, $k_j\, j \neq i$.. The optimal decision boundaries are formed from the joint space $f(x, y)$ that is composed of n input output pairs,

$$(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$$

In the case of modeling the ML estimators with Q discriminant functions, the indicator responses within each of the l pairs are

$$f(y_j) = \begin{cases} 1 & , \quad x_j \in \mathscr{S}_i = \{x_j : f\left(y_j = k_i | x_j\right)\} \\ -1 & , \quad x_j \notin \mathscr{S}_i \end{cases} \tag{4.30}$$

for a given classifier i, $i \in S$.

From the training joint space an optimal hyperplane can be formed as a linear combination of support vectors. The decision functions described in the relations of Section 4.2 are general forms of constructing a boundary for linear separable data. Speech, though, is a highly variable signal. Parameterized forms of this speech signal, Figure 1.2, create piecewise stationary representations of the signal. Modeling this signal with discriminative methods requires classifiers with complex decision boundaries. In other words, to effectively distinguish between one set of parameterized speech patterns and another, complex decision boundaries are necessary. As described in Section 4.2, such decision boundaries are easily formed by constructing the optimal hyperplane with a linear combination of basis functions. Since the convex optimization problem of Equation 4.26 requires only the inner product of the inputs, kernel, $K()$, methods can be used in

the basis transformation. The polynomial and Gaussian kernels with a variance parameter, γ^2,

$$K\left(x_i, x_j\right) = exp\left(-\gamma^2 \left(x_i - x_j\right)^2\right) \qquad (4.31)$$
$$x \in \mathscr{R}^m$$

are commonly used to construct support vector machines to form decision boundaries for non linear separable data. Using the kernel function of Equation 4.31, the optimization problem, Equation 4.26 can be expressed as,

$$W = \sum_i \alpha - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j K\left(x_i \cdot x_j\right),\ 0 \leq \alpha_i \leq C, \sum_i y_i = 0, \qquad (4.32)$$
$$x \in \mathscr{R}^m, y \in \{-1, 1\}, i, j \in \mathscr{I}$$

where C is a penalization parameter and α, Lagrange multipliers.

From the expression of Equation 4.32, decision functions for each discriminative classifier in the set of S can be determined. The posterior expression, $f(y = k\,|\,x)$ is determined from Equation 4.29 and Equation 4.28, and is rewritten here,

$$f\left(y = k\,|\,x\right) = sign\, f\left(\hat{x}\right)$$
$$f\left(\hat{x}\right) = \sum_i y_i \alpha_i \left(x_i \cdot x\right)$$

The classification outputs, $f(y = k\,|\,x)$, though, are not probabilistic measures. To model the ML estimators with the resultant discriminative classifiers, the outputs need to be transformed into relative confidence measures so as to form probabilistic posteriors. This can be done with a post processing non linear transformation[56]

$$P(y = 1\,|\,f) = \frac{1}{1 + exp\left(1 + Af + B\right)} \qquad (4.33)$$

where, A and B are trainable parameters of a fitted sigmoid function. In using the relation of Equation 4.33, the support vector models are capable of modeling the ML estimators of the HMM topology. Moreover, these discriminative models through forming its decision functions with a hyperplane that minimizes the error and the structural risk, can be considered to be optimal. This hyperplane and resultant decision function may also be considered to be one that minimizes the expected error between the true separable boundary and its estimate. Consider the

loss function between the true hyperplane, y, and its estimate, $\hat{y}$, Equation 4.11,

$$L = (y - \hat{y})$$

Its expected value is,

$$\begin{aligned} E[L] &= E(y - \hat{y}) \\ &= \sum E[(y - \hat{y}) | x] P(x) \\ &= E E[(y - \hat{y}) | x] \end{aligned} \tag{4.34}$$

conditional on the input, x. Minimizing this loss can be expressed in terms of minimizing $E[(y - \hat{y}) | x]$ which is the optimal hyperplane. Furthermore, the probabilistic support vector machine based ML estimator can be expressed in terms this expression. As the expected value of $logP(Y|X)$ is the conditional entropy, it implies that the large margin ML estimator satisfies Equation 4.1 and so is capable of modeling the observation distribution.

4.5 Findings and Summary

Problem :
To devise and develop effective stochastic models for modeling the speech process.
Dissertation Contributions :

- Development of objective function relating true observation distribution to an estimate in terms of the directly observable measurements and latent hidden variables.
- Formulation and development of an optimal stochastic model for the speech with noise problem. Proposal of a combination of recognizers that through a simple system fusion, combines multiple speech processes at the decision level. This is a novel stochastic method devised to combine a parameterized spectral missing data, MD, theory based and a cepstral based speech process using a coupled hidden variable topology. In using a fused coupled hidden Markov model, HMM, topology, an optimal stochastic model is proposed that is inherently more robust than single process models under noisy conditions.

- A novel analysis and comparison of the capabilities of coupled hidden variable topologies to model the speech process.
- Through the maximization of the devised objective function, Equation 2.16, it is shown that the resultant optimal combined acoustic space contains greater information content of the true observation distribution. Thus is capable of improved recognition accuracies.
- Segmentation of the speech acoustic space in a manner that can represent the speech process effectively and can be modeled with discriminative learning methods.
- Devising an optimal discriminant ML estimator to model the speech observation distribution.

In this chapter an optimal discriminant ML estimator is proposed that satisfies the objective function, Equation 2.16, devised to effectively model the speech process. This is formulated with the large margin discriminant function. Specifically, the methodology presented poses the speech modeling problem as that of,

- segmenting the acoustic space in a manner such that the stochastic model can encode the speech process without loss
- to effectively model the true observation distribution of the speech process

The first is formulated as a time series modeling problem. As is detailed in Section 4.1, the speech process can be modeled effectively through factorizing the acoustic space over multiple states. The hidden variable topology that represents this stochastic construct consists of a 1st order chain of latent ***rvs***, U, generating observations, O. The temporal aspect of the speech signal is captured within the hidden variables, U, (Section 2.1). A segmentation of the speech observation process, or acoustic space that is necessary to encode the process is expressed in terms of the number of hidden variables. This devised relation, Equation 4.5, relates the number of hidden states to the accuracy of the observation distribution modeled. This observation distribution is represented by the hidden variable generative process that is modeled as ML estimators.

The second involves modeling these ML estimators. Through using the segmentation of the acoustic space, as proposed in this chapter, the observation distribution can be modeled with ML estimators that are formulated as discriminatively trained estimators of the speech process. By accurately modeling the speech

observation distribution with the ML estimators, the speech process may be effectively encoded. The estimators are typically a problem of generative density estimation. Such a problem may be construed as ill posed. Discriminative techniques, Section 2.2, however, provide methods to model these estimators with powerful classifiers, or decision functions, that distinguish between separate clusters of data. Such techniques may form discriminant functions with effective optimization techniques. This chapter devises a large margin discriminant ML estimator. Such an estimator is not only formed with optimal decision functions but is also shown to effectively model the speech process. Unlike previous research efforts that have proposed discriminatively trained acoustic models, this work is presented as that of optimizing the objective function of Equation 2.16 and thus capable of effectively modeling the speech process without loss(Section 4.3).

Several discriminative methods are compared and contrasted in Section 4.2. From this comparison it is reasoned that the discriminant functions formed with support vector machines, SVMs, using the large margin training method are the most suitable for this proposed model. Not only are the discriminant functions optimal in the sense of minimizing the classification error, but also have strong generalization ability. Formed with the large training method, the resultant proposed ML estimators are shown to have increased information content of the true observation distribution. Moreover the formulation is shown to maximize the entropy of the observation distribution and thus satisfies the devised objective function of Chapter 2, Equation 2.16.

- **The proposed acoustic model is optimal**,
  The ML estimators are modeled with the large margin discriminative learning method. The decision boundaries form a unique hyperplane that separate data[68] with the largest margin between classes of data. This is formulated as the optimization problem of minimizing a quadratic functional subject to a linear constraint. Solved with Lagrangian multipliers, its dual form is used to determine the parameters for the resultant hyperplane. This dual form (Equation 4.26), W, is maximized:

$$W = \sum_i \alpha - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \left(x_i \cdot x_j \right), \ \alpha \geq 0, \sum_i y_i \alpha_i = 0$$
$$i, j \in \mathscr{I}, y \in \{-1, 1\}$$

  where, x_i, x_j are learning examples, y_i, y_j are the corresponding classification results(-1,1), and α_i, α_j are Lagrangian multipliers. It is a convex opti-

mization problem that is subject to the Kuhn Tucker conditions. It therefore has a global maximum. The resultant determined parameters for this hyperplane are a linear combination of support vectors and Lagrange multiplier weights, Section 4.2. The hyperplane through its construction is optimal both in the sense of minimizing the classification error and it having the largest margin, or distance, between classes of data. Therefore the proposed discriminant ML estimator is optimal in this sense.

The acoustic model proposed in this chapter is composed of ML estimators Section 4.1. The discriminant ML estimators are optimal. Therefore the acoustic model is optimal in that sense.

- **The proposed model is capable of effectively modeling the speech process**,
  The proposed model is an acoustic space modeled with optimal discriminatively trained ML estimators. Specifically discriminant ML estimators formed using the large margin learning method. The capability of the proposed methodology to model the speech process can be described in the following.

  The ML estimators model the acoustic space segmented by the number of hidden states within the hidden variable topology. If this model is capable of modeling the speech it should, as described in Chapter 2, have sufficient hidden states, U and accurate observation distributions that are modeled with ML estimators. The necessary number of these states is expressed in Section 4.1, in the final expression of Equation 4.5 that defines this number, m_U, in terms of the upper bound of the entropy of the observation distribution that is modeled, $H(\vartheta)$,

$$m_U \geq 2^{H(\vartheta)}$$

  This relation is a result of the mutual information, $I()$, or temporal, relation between successive observation measurements. Here expressed as ***rvs***, O and U that represent the observations and latent variables of the hidden variable topology,

$$I\left(O_n U^{(n)}\right) \geq I\left(O_n O^{(n-1)}\right)$$

  This expression indicates that the mutual information between an observation at time, or stage, n, O_n and the process information that is encoded in n

realizations of U, $U^{(n)}$, is stronger than that of O at n and the previous $n-1$ observations. The relation can be further related to the devised objective function, Equation 2.16, in terms of ML estimators. Thus, $m_U \geq 2^{H(\vartheta)}$, expresses the relationship between the number of hidden states and the quality of the model of the observation distribution.

As the effectiveness of encoding the observation process is dependent upon both the ML estimator and the number of hidden states that is also dependent upon the former. By optimizing the ML estimator and using the proposed segmentation of the acoustic space, the resultant model is capable of effectively modeling the speech process.

- **The proposed model increases the accuracy, or information content of the true speech process in the resultant models**,
  The proposed model maximizes the entropy of the observation distribution.

  The speech process, in this proposed acoustic model, is modeled with discriminant ML estimators. Each estimator is a decision boundary that is formed with the large margin discriminative method. This decision function may be expressed in terms of the loss function, $L()$, Equation 4.11,

$$L = (y - \hat{y}),$$

  where y is the true decision function and $\hat{y}$ its estimate. The optimal decision boundary, y^*, is the boundary the minimizes the expected value of the error between the true decision boundary and its estimate conditional on the learning examples, x,

$$E[(y - \hat{y}) \,|x]$$

  The expected value of the log value of the conditional is the conditional entropy. Maximizing this function maximizes the devised objective function of Equation 2.16. Since the proposed discriminant ML estimator maximizes this conditional, the resultant estimator satisfies the objective function, Equation 2.16. In doing so it maximizes the information content of the true speech process in its resultant models.

Book IV

Experiments

Table 4.1: Sample Corpus Vocabulary

Grammar Type	Grid Dictionary Word
verb	bin lay place set
color	blue green red white
prep	at by in with
letter	a-v, y-z
number	zero one two three four five six seven eight nine
adv	again now please soon

4.6 Large Margin Discriminant ML Estimator: Experimental Results

Experiments have been conducted to ascertain the validity of the methodology presented in the previous section. This was achieved though analyzing the task of recognizing words from sentences derived from the Grid Corpus [18]. The corpus used consisted of 33000 unique sentences generated from 33 different speakers. Each sentence or utterance is derived from the following grammar,

$verb sp $color sp $prep sp $letter sp $number sp $adv,
where sp is a silence marker. The vocabulary is illustrated in Table 4.1.

HMM word models were constructed and trained using half of the corpus composed of utterances from 17 different speakers. Each HMM model consists of 32 Gaussian Mixtures, diagonal covariance matrices, with the number of states corresponding to its word phoneme representation as per the CMU dictionary[5]. Each utterance of the training set are transformed to a cepstral, MFCC, format consisting of 39 dimensions with energy, delta and acceleration coefficients. A test set was formed using the unused half of the corpus composed of utterances from 16 speakers differing from the training set. The HTK toolkit [77] was used for creating the HMM models, cepstral features and for determining the time alignment of

[5]www. speech. cs. cmu. edu/cgi-bin/cmudict

Table 4.2: Average Number of Training Examples for SVMs and corresponding number of SVMs per Learning Machine Partitioned by Number of States per Word

No. States	Training Examples	SVMs
2	33197.4	3754
4	33856.6	3102
6	38239.4	3047
8	39642.3	2870
10	33991.3	2848

the features. The baseline recognizer produced recognition accuracies of 95.15% over a subset of the test set.

SVM classifiers were constructed using the HMM word models to provide the initial segmentation and alignment of the utterances. Each state of the HMM models were mapped to form a separate SVM classifier creating 256 distinct learning machines. The HMM models provided the time alignment of each utterance of the training set to a corresponding state. This time alignment was within a sampling of 10ms corresponding to a frame. Each frame of the training set served as the generator set with the target labeling derived from the corresponding state label. The training strategy adopted was *one versus all* so as the supervisor formed target responses to each frame of the training set corresponding to Equation 4.30. Each learning machine was constructed with Gaussian, or Radial Basis Function, RBF, kernels. The parameters of the model and the parameters of the sigmoid function of Equation 4.33 were determined using a cross validation procedure, *N-fold*, composed of three sets. In this parameter determination scheme, each target generator training set was split into three parts. Permutations of two out of three served as training examples with the third part used for testing the performance of the resultant SVM. The training examples for each SVM classifier were formed from the training set of the HMM models. Each frame was inserted into a database and labeled with its corresponding word model state. From this database, each SVM training set was formed using all of the features for the target class together with a fixed number from each other class including examples from every speaker within the training development set. All SVM classifiers were constructed using the Torch3 machine learning development suite [17]. The average number

Table 4.3: SVM Classifier Error Rates over Training Database and Test Database Test Sets, Partitioned by Number of States per Word

No. States	Train Db Test Set (55512 examples), Error Rate (x100) %	Test Db Test Set (46260 examples), Error Rate (x100) %
2	0.0135	0.0292
4	0.0104	0.0240
6	0.0101	0.0258
8	0.0090	0.0228
10	0.0099	0.0230

of training examples and the resultant number of SVMs per classifier partitioned by the number of states per word are illustrated in Table 4.2.

A test database was formed with MFCC features that were time aligned to word state labels using the HMM acoustic models by the discriminative space mapping technique described in the methodology. This database consisted of frames from 16 different speakers taken from the test set formed from the corpus.

Two sets of test examples were constructed to ascertain the performance of the speech classifiers. The first, a set composed of features or frames from the training database. This set of examples consisted of a small sample of features representative of each class and each speaker within the training database. The distribution of frames taken from each class and speaker to form the set of examples was uniformly distributed. Similarly, a test set of examples was constructed from the database of test features indicative of unknown speaker examples to the learning machines. Table 4.3 reflects the average error rates of each classifier with outputs subjected to Equation 4.33 grouped by the number of states within the respective word models. A positive classification is associated with $p > 0.5$. The error rates are with respect to the classifier performance over the sampling of the features from speakers and features the classifiers were presented and over completely unseen examples. A lower rate corresponds to fewer misclassifications. As such, a 0% classification error rate would correspond to a perfect classifier.

The constructed classifiers possess the ability to differentiate between the speech patterns presented. This is evident from the results of the classification

exercises conducted over the entire set of SVMs. The strength and flexibility of the classifiers are reflected in the classification error rates determined in Table 4.3. The number of SVMs per classifier depicted in Table 4.2 demonstrate the compactness of the learning machines. Only a small fraction of the examples presented are required to form the decision functions reflecting the distribution of $f(x, y)$. In the case of the methodology and corresponding experiment setup, the SVMs are formulated with approximately *10%* of the examples presented.

The results of Table 4.3 demonstrate the strong generalization and differentiating capabilities of the constructed learning machines. The classification error rate results of the SVMs subjected to the test set derived from the training database provide evidence that the models do not suffer from *overfitting*. Furthermore, they are able to generalize well over unseen examples generated by known speakers. With respect to the test set formed from the training database, a majority of the examples chosen were not used to train the classifiers but do originate from speakers of which examples were presented. The classification error rates from the experiments conducted with the test database set provides empirical evidence that the constructed learning machines possess powerful distinguishing characteristics. This can be stated as all of the examples presented are generated by speakers other than were used to train the classifiers. Each separate speaker and corresponding utterance provides a significant variation to the speech pattern presented to the classifier. Relative to the baseline error rate established by the training test set, the classification of these unseen, significantly variant, examples experience only a small increase in classification error rate. Inferred in the results of the experiments conducted is the ability to overcome the problems which hinder other established discriminative techniques. Those problems being, controlling the complexity of the learning machine while maintaining a low error rate, as in Equation 4.14, over known and unknown data points.

As specified in the methodology, the outputs of each classifier are subjected to post-processing with a sigmoid function to obtain a probabilistic representation. It had been observed that with the use of this technique, classification error rates were lower in comparison to unaltered classifier error rates. This is an additional benefit to the original intention of creating probabilistic measures to aid in analyzing the performance of the classifiers. The probabilistic measures derived from the SVM outputs served as a measure of confidence in the classification ability of a given SVM classifier for a speech pattern. It had been observed that the probabilistic measures were calibrated across all classifiers. This quality makes them appropriate for discerning sequences of speech patterns for speech recognition.

Chapter 5

Conclusions and Future Directions

Engineering a solution to the speech recognition problem has inspired many research efforts. It was anticipated that the problem would be solved in its entirety shortly after the first speech recognizer was devised. The complexity of the problem, though, has prevented that realization. Over past few decades a rich and vibrant research community has been formulating innovative solutions to solve various facets of the problem. Many of these techniques have been subsequently applied in other scientific fields. Arguably, the speech research community's most significant contributions to the scientific process has been its work on stochastic modeling. Stochastic modeling of the speech process is fundamental to the speech recognition problem. This thesis investigates this problem. It formulates and presents methodologies to effectively model the observation process with optimal stochastic acoustic models. Two such models are proposed. The first, uses a coupled time series topology that is capable of effectively encoding the speech process for robust speech recognition. The second models the observation process with discriminatively trained ML estimators.

The speech process time series can be modeled with a hidden variable stochastic model. This model is capable of effectively representing the speech process. In its manner of encoding the mutual information, or the correlation between successive samples, with a first order Markov chain, the hidden variables capture the temporal aspect of the signal. It can be shown through information theoretic concepts that the observation distribution of the input parameterized samples can be represented within its ML estimators, or, emission densities without loss of information.

This thesis investigated optimal stochastic models for the speech recognition problem. Using the time series analysis techniques shown in this thesis, a cou-

pled time series topology was proposed to increase the information content of the acoustic space encoded within it for robust speech recognition. It was shown that the resultant model optimized a derived relation[1] that described the statistical process. In doing so, it increased the true speech acoustic content of the model. Similarly, a model based on this relation was formed in terms of discriminately trained ML estimators. Specifically ML estimators that are devised in terms of discriminant decision boundaries formed with the large margin discriminative method. It was also shown that in formulating the problem in this manner, the resultant models are capable of modeling the observation process without loss.

The combination of recognizers was shown to be an effective approach to the speech with noise problem. This method used an optimal coupled stochastic model to represent the joint space of two representations, or two parameterized streams of the speech signal. This combined classifier technique utilized both missing data, MD, techniques for noise robust recognition and combined it with cepstral based techniques at the decision level. Through an optimal stochastic coupled time series acoustic model, the joint space of both processes was captured. Representing the combined classifiers in this manner allowed the statistical dependencies between both processes to be inferred from the coupled topology. Moreover, the combined space increased the models' capacity to capture and represent the true speech process. This became evident during a series of speech recognition experiments. Under all tested conditions, recognition results from the combined joint space fused model outperformed those from all other single process acoustic model recognizers tested.

The joint space itself is currently implemented with Gaussian densities that represent the observation space. The effectiveness of this fused model can be related to how well it captures the information content of the true speech process. This can be shown in relation to the objective function of Equation 2.16. Since this objective function is not dependent upon modeling with Gaussian densities (as is demonstrated with the discriminant ML estimators), the effectiveness of the fused model in not dependent on the use of Gaussian densities.

This methodology may be further enhanced by investigating improvements to the MD mixture model. Further refinements to the MD acoustic models can only improve the combination of recognizers' capability to represent the speech process. Currently the MD mixture model has some modeling limitations due to some necessary assumptions. These conditions can be potentially overcome through the use of variational methods.

[1] The relation is the objection function $\approx H(O|U)$, Equation 2.16

Discriminative models of ML estimators within a hidden variable topology have the potential to effectively capture the speech process. Discriminatively trained ML estimators can be devised from a segmentation of the acoustic space that permits the speech process to be captured and segmented into unique segments suitable to form discriminant decision boundaries. An ML estimator formed with a large margin decision boundary was shown to be both the optimal ML estimator and one that maximizes the entropy of the speech acoustic model. Through non linear decision functions formed from parameterized speech frames, these classifiers were shown theoretically and empirically to be capable of representing the observation distribution of the speech process without loss. Empirically this was evident with the strong classification performance (2-3% error rate) of the classifiers over a test set of cepstral based speech frames. This set of research may be further enhanced though introducing either regressive large margin discriminant methods or relevance vector machines into the methodology. These discriminative techniques may offer further refinements to modeling the ML estimators.

5.1 Contributions

The presented thesis contributes to stochastic modeling for speech recognition research in the following manner. Contributions:

- Development of objective function relating true observation distribution to an estimate in terms of the directly observable measurements and latent hidden variables.
- Formulation and development of an optimal stochastic model for the speech with noise problem. Proposal of a combination of recognizers that through a simple system fusion, combines multiple speech processes at the decision level. This is a novel stochastic method devised to combine a parameterized spectral missing data, MD, theory based and a cepstral based speech process using a coupled hidden variable topology. In using a fused coupled hidden Markov model, HMM, topology, an optimal stochastic model is proposed that is inherently more robust than single process models under noisy conditions.
- A novel analysis and comparison of the capabilities of coupled hidden variable topologies to model the speech process.

- Through the maximization of the devised objective function, Equation 2.16, it is shown that the resultant optimal combined acoustic space contains greater information content of the true observation distribution. Thus is capable of improved recognition accuracies.

- Segmentation of the speech acoustic space in a manner that can represent the speech process effectively and can be modeled

- Devising an optimal discriminant ML estimator to model the speech observation distribution.

The contributions are directly related to effectively modeling the speech process with stochastic acoustic models. These contributions form stochastic models for the speech recognition problem. An objective function is devised, that relates the observation distribution to its estimate. This objective function, formulated in this manner, represents both the transient properties of the signal and its information content. In maximizing the estimate term of this expression, the resultant model is capable of representing the observation process without loss of information.

A novel acoustic model is devised to model the speech process for the speech with noise problem. Here, an effective robust model is formed through a simple system fusion of multiple speech processes at the decision level. Its contribution to speech recognition stochastic modeling is in the approach that is taken for robust speech recognition. This approach combines two speech processes at the decision level. By inferring the statistical dependencies between a missing data, MD, ASR process, and a conventional cepstal process, the resultant acoustic model is robust under both stationary and non stationary noise conditions. The structure of the resultant model is formed such that it satisfies the devised objective function. Thus it effectively models the speech process.

An optimal acoustic model is devised that is composed of optimal discriminant ML estimators. Its contribution to the stochastic modeling of speech is in that the resultant model is optimized to effectively encode the speech process without loss. Such models not only benefits from having discriminant decision boundaries, but also by directly optimizing the devised objective function.

Appendix A

Previously Published Work

N. Joshi and L. Guan, "Missing data ASR with fusion of features and combination of recognizers," *IEEE Spoken Language Technology Workshop*, pp. 114–117, 2006.
N. Joshi and L. Guan, "Combination of Recognizers and Fusion of Features Approach to Missing Data ASR Under Non-Stationary Noise Conditions," *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007*, pp. 1041–1044, 2007.
N. Joshi and L. Guan, "Feature Fusion Applied to Missing Data ASR with the Combination of Recognizers," *Journal of Signal Processing Systems*, pp. 1–12, 2009.

Appendix B

Abbreviations

ASR Automatic speech recognition.
CMN Cepstral mean normalization.
COR Combination of recognizers.
HMM Hidden markov model.
KL Kullback-Leibler.
MD Missing data.
MF Mel-frequency.
MFCC Mel-frequency cepstral coefficients.
ML Maximum likelihood.
NN Neural network.
SVM Support vector machine.

Acronyms in order of appearance:
IBM International Business Machines
HMM Hidden Markov Model
NN Neural Network
DARPA Defense Advanced Research Projects Agency
CMU Carnegie Mellon University
BBN Bolt, Beranek and Newman
SRI Stanford Research Institute
MIT Massachusetts Institute of Technology
RASTA Relative Spectra
MD Missing Data
WWW World Wide Web

HLT Human Language Technology
NP Noun Phrase
ML Maximum Likelihood
MF Mel Frequency
MFCC Mel Frequency Cepstral Coefficient
CMN Cepstral Mean Normalization
PMC Parallel Model Combination
KL Kullback Leibler
SVM Support Vector Machine
EM Expectation Maximization
ASR Automatic Speech Recognition
CASA Computational Auditory Scene Analysis
CDHMM Continuous Density Hidden Markov Model
MAP Maximum Aposterior Probability
COR Combination of Recognizers
HTK Hidden Markov Model Toolkit
CTK Computational Auditory Scene Analysis Toolkit
SNR Signal to Noise Ratio
SFD Segment Fragment Decoder
MLP Multilayer Perceptron
RBF Radial Basis Function

Bibliography

[1] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, 1974.

[2] L. Bahl, P. Brown, P. D. Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," vol. 11, 1986.

[3] J. Barker, *RESPITE CASE Toolkit CTK v1.1.1 User's Guide*. University of Sheffield, 2001.

[4] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," *Sixth International Conference on Spoken Language Processing*, pp. 373–376, 2000.

[5] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, no. 1, pp. 5–25, 2005.

[6] Y. Bengio and P. Frasconi, "Input-output HMMs for sequence processing," *IEEE Transactions on Neural Networks*, vol. 7, no. 5, pp. 1231–1249, 1996.

[7] A. Betkowska, K. Shinoda, and S. Furui, "Speech recognition using FHMMs robust against nonstationary noise," vol. 4, 2007.

[8] J. Bilmes, "What HMMs can do," *IEICE Transactions on Information and Systems*, vol. 89, no. 3, pp. 869–891, 2006.

[9] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[10] H. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Kluwer Academic Pub, 1994.

[11] M. Brand, "Coupled hidden markov models for modeling interacting processes," *Tech. Rep. 405 – MIT Media Lab*, 1997.

[12] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

[13] C. Cerisara, "Towards missing data recognition with cepstral features," *Eighth European Conference on Speech Communication and Technology*, pp. 3057–3060, 2003.

[14] C. Cerisara, S. Demange, and J. P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Computer Speech & Language*, vol. 21, no. 3, pp. 443–457, 2007.

[15] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 23–37, 2002.

[16] P. Clarkson and P. Moreno, "On the use of support vector machines for phonetic classification," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 585–588, 1999.

[17] R. Collobert, S. Bengio, and J. Mariethoz, "Torch: a modular machine learning software library," *IDIAP Research Report*, vol. 2, 2002.

[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, pp. 2421–2424, 2006.

[19] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1997.

[20] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, Jun. 2001.

[21] M. Cooke and T. Lee, "The 2006 speech separation challenge," *Computer Speech & Language*, 2008.

[22] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley-Interscience, 2006.

[23] R. Cox, "Probability, frequency and reasonable expectation," *American Journal of Physics*, vol. 14, no. 1, pp. 1–13, 1946.

[24] N. Cristianini and J. Shawe-Taylor, *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, 2000.

[25] J. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *The Annals of Mathematical Statistics*, pp. 1470–1480, 1972.

[26] K. H. Davis, R. Buddulph, and S. Balashek, "Automatic recognition of spoken digits," *J. Acoustic Soc. Am.*, vol. 6, pp. 637–642, 1952.

[27] S. Demange, C. Cerisara, and J. P. Haton, "Missing data mask estimation with frequency and temporal dependencies," *Computer Speech & Language*, vol. 23, no. 1, pp. 25–41, 2009.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[29] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley-Interscience, 2001.

[30] W. Feller, *An Introduction to Probability Theory and Its Applications. Vol. 1*. Wiley, 1968.

[31] J. Ferguson, "Hidden markov models for speech," *IDA – Princeton, NJ*, 1980.

[32] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, Oct. 1996.

[33] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, Sep. 1996.

[34] A. Ganapathiraju, J. E. Hamaker, J. Picone, and R. Conversay, "Applications of support vector machines to speech recognition," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2348–2355, 2004.

[35] A. Gelman, J. Carlin, and H. Stern, *Bayesian data analysis*. CRC press, 2004.

[36] H. V. Hamme, "Robust speech recognition using missing feature theory in the cepstral or LDA domain," *Eighth European Conference on Speech Communication and Technology*, pp. 1973–1976, 2003.

[37] ——, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004.

[38] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 58, no. 1, pp. 155–176, 1996.

[39] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[40] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[41] M. M. Hochberg, G. D. Cook, S. J. Renals, A. J. Robinson, and R. S. Schechtman, "The 1994 ABBOT hybrid connectionist hmm large-vocabulary recognition system," *Spoken Language Systems Technology Workshop*, 1995.

[42] J. Häkkinen and H. Haverinen, "On the use of missing feature theory with cepstral features," *Proc.of Eurospeech, The Workshop on Consistent and Reliable Acoustic Cues, Aalborg, Denmark, September*, 2001.

[43] D. Jurafsky and J. H. Martin, *Speech and language processing*. Prentice Hall, 2008.

[44] S. Kapadia, "Discriminative Training of Hidden Markov Models [Ph. D. dissertation]," *Cambridge University Engineering Department, UK*, 1998.

[45] N. Karmarkar, "A new polynomial-time algorithm for linear programming," *Combinatorica*, vol. 4, no. 4, pp. 373–395, 1984.

[46] D. Klatt, "A digital filter bank for spectral matching," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 573–576, 1976.

[47] T. T. Kristjansson, B. J. Frey, and T. S. Huang, "Event-coupled hidden markov models," *IEEE ICME Multimedia and Expo*, vol. 1, pp. 385–388, 2000.

[48] S. Kullback, "The kullback-leibler distance," *The American Statistician*, vol. 41, pp. 340–341, 1987.

[49] R. P. Lippmann, "An introduction to computing with neural networks," *IEEE ASSP Mag*, vol. 2, pp. 4–22, 1987.

[50] T. Marin, A. L. Nelson, and H. Zadell, "Speech recognition by feature abstraction techniques," *Tech. Report AL-TDR-64-176, Air Force Avionics Lab*, 1964.

[51] ——, "Speech discrimination by dynamic programming," *Kibernetika*, vol. 2, pp. 81–88, 1968.

[52] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., 1982.

[53] B. C. J. Moore, *An introduction to the psychology of hearing*. Academic press, 2003.

[54] R. Nopsuwanchai, "Discriminative training methods and their applications to handwriting recognition [Ph. D. dissertation]," *Cambridge University Engineering Department, UK*, 2005.

[55] H. Pan, S. E. Levinson, T. S. Huang, and Z. P. Liang, "A fused hidden markov model with application to bimodal speech processing," *IEEE Tranactions on Acoustics, Speech, and Signal Processing*, vol. 52, no. 3, pp. 573–581, 2004.

[56] J. Platt, "Probabilistic outputs for svms and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, pp. 61–74, 1999.

[57] W. Press, *Numerical recipes: the art of scientific computing*. Cambridge university press, 2007.

[58] L. R. Rabiner, "A tutorial on hidden markov models and selected application in speech recognition," *Proc. IEEE*, pp. 257–286, 1989.

[59] B. Raj, M. L. Seltzer, and R. M. Stern, "Robust speech recognition: the case for restoring missing features," *Proc.of Eurospeech, The Workshop on Consistent and Reliable Acoustic Cues*, 2001.

[60] B. Raj and R. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[61] D. R. Reddy, "An approach to computer speech reconition by direct analysis of the speech wave," *Tech. Report No. C549, Computer Science Dept. Stanford Univeristy*, 1966.

[62] D. F. Rosenthal and H. G. Okuno, *Computational auditory scene analysis*. Lawrence Erlbaum Associates, 1998.

[63] L. K. Saul and M. I. Jordan, "Mixed memory markov models: Decomposing complex stochastic processes as mixtures of simpler ones," *Machine Learning*, vol. 37, no. 1, pp. 75–87, 1999.

[64] E. Singer and R. P. Lippman, "A speech recognizer using radial basis function neural networks in an hmm framework," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 629–632, 1992.

[65] J. Stadermann and G. Rigoll, "A hybrid SVM/HMM acoustic modeling approach to automatic speech recognition," *Eighth International Conference on Spoken Language Processing*, pp. 661–664, 2004.

[66] T. G. Stockham, T. M. Cannon, and R. B. Ingebretsen, "Blind deconvolution through digital signal processing," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 678–692, 1975.

[67] V. Tyagi, I. McCowan, H. Bourlard, and H. Misra, "Mel-Cepstrum modulation spectrum (MCMS) features for robust ASR," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 399–404, 2003.

[68] V. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[69] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 2000.

[70] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," *DRA Speech Research Unit, Malvern, England, Tech. Rep*, 1992.

[71] A. P. Varga and R. K. Moore, "Hidden markov model decomposition of speech and noise," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 845–848, 1990.

[72] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[73] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[74] A. Weibel, T. Hanazawa, G. Hinton, K. Shirano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 393–404, 1989.

[75] P. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[76] P. Wolfe, "The simplex method for quadratic programming," *Econometrica: Journal of the Econometric Society*, pp. 382–398, 1959.

[77] S. J. Young, P. C. Woodland, and W. J. Byrne, *HTK User, Reference and Programming Manual*. Cambridge University, Engineering Department & Entropic Research Laboratory Inc, 1993.

December 18, 2009