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 process 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

Abstract . . . . **iii**

Dedication . . . . **v**

Table of Contents . . . . **vi**

List of Tables . . . . **viii**

List of Figures . . . . **ix**

1 Introduction **1**
1.1 Research focus 6
1.2 Research Contributions 9
1.3 Organization of thesis 10
1.4 Guide to the reader 11
Dissertation in five minutes 11
For the eager reader 11
1.5 Commonly used notation and symbols 13

2 Acoustic Modeling **16**
2.1 Hidden variable acoustic models 19
Assessing hidden variable model capabilities 22
Parameter training 24
2.2 Discriminative acoustic models 26
2.3 Findings and Summary 28

3 Speech with Noise: Combination of Recognizers **31**
I Combination of Recognizers 32
3.1 Speech with Noise 33
3.2 Missing Data Theory 37

3.3 Missing Data Theory: Pattern Recognition 40
3.4 Missing Data Theory: The Case for Cepstral Features 43
3.5 Modeling Interacting Processes 47
3.6 Coupled Stochastic Time Series 49
3.7 Combination of Recognizers 57
3.8 Findings and Summary 63
II Experiments 68
3.9 Combination of Recognizers: Experiments Setup 69
3.10 Combination of Recognizers: Stationary Noise Experiments . . . 72
3.11 COR: Non Stationary Noise Experiments 76

4 ML :A Large Margin Approach 84
III Discriminant ML Estimator 85
4.1 Acoustic Space 89
4.2 Discriminative Techniques 93
Logit Regression 96
Linear Discriminant Analysis 98
NN 99
Large Margin 100
4.3 Discriminative Speech Modeling: Past and Present 103
4.4 Large Margin Discriminant ML Estimator 105
4.5 Findings and Summary 109
IV Experiments 114
4.6 Large Margin Discriminant ML Estimator: Experimental Results . 115

5 Conclusions and Future Directions 119
5.1 Contributions 121

A Previously Published Work 123

B Abbreviations 124

Bibliography 126

List of Tables

3.1 COR theoretical recognition capacities 62
3.2 HMM States per Word in Experiments Recognizer Vocabulary . . 70
3.3 Experiments Corpus Training Sets 71
3.4 Experiments Corpus Test Sets 71
3.5 Stationary Noise: Baseline Recognizer Results, Clean Data 73
3.6 Recognizer Results With Test Corpus + Stationary Noise 74
3.7 Stationary Noise 6dB SNR Rankings 74
3.8 Stationary Noise 0dB SNR Rankings 75
3.9 Non Stationary Noise: BaseLine Recognizer Results, Clean Data . 77
3.10 Recognizer Results With Test Corpus + Destroyer Noise 77
3.11 Recognizer Results With Test Corpus + Factory Noise 78
3.12 Destroyer Noise 18dB SNR Rankings 79
3.13 Factory Noise 18dB SNR Rankings 80
3.14 Destroyer Noise 12dB SNR Rankings 81
3.15 Factory Noise 12dB SNR Rankings 81
3.16 Destroyer Noise 6dB SNR Rankings 82
3.17 Factory Noise 6dB SNR Rankings 83

4.1 Sample Corpus Vocabulary 115
4.2 Support Vectors per classifier 116
4.3 Test SVM Classification Rates 117

List of Figures

1.1 Trivial Speech Recognition Network 3
1.2 Parameterization of speech signal 5
1.3 Illustration of probabilistic space containing ***rvs*** *U* 15
1.4 Illustration of probabilistic space containing ***rvs*** *U* and *O* 15

2.1 1st order Markov chain stochastic graph 20
2.2 Hidden markov model 21

3.1 Speech with additive noise 33
3.2 CASA groupings of spectral utterance 38
3.3 Spectral utterance and binary MD mask 39
3.4 MD data imputation pattern recognition 42
3.5 MD marginalization pattern recognition 42
3.6 HMM topology 50
3.7 Multiple, g, stochastic time series 50
3.8 Coupled HMM 51
3.9 Mixed memory HMM 53
3.10 Fused HMM 56
3.11 Combination of recognizers pattern recognition 58
3.12 Combination of recognizers acoustic model 59
3.13 Stationary noise recognizer configuration rankings 75
3.14 Destroyer noise recognizer configuration rankings 82

4.1 Linear combination decision boundary 94
4.2 Decision functions 95
4.3 Discriminant ML Estimator 106

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

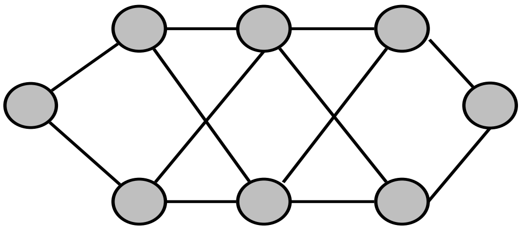

Figure 1.1: Trivial Speech Recognition Network

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 posteriors where,

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

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,

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

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

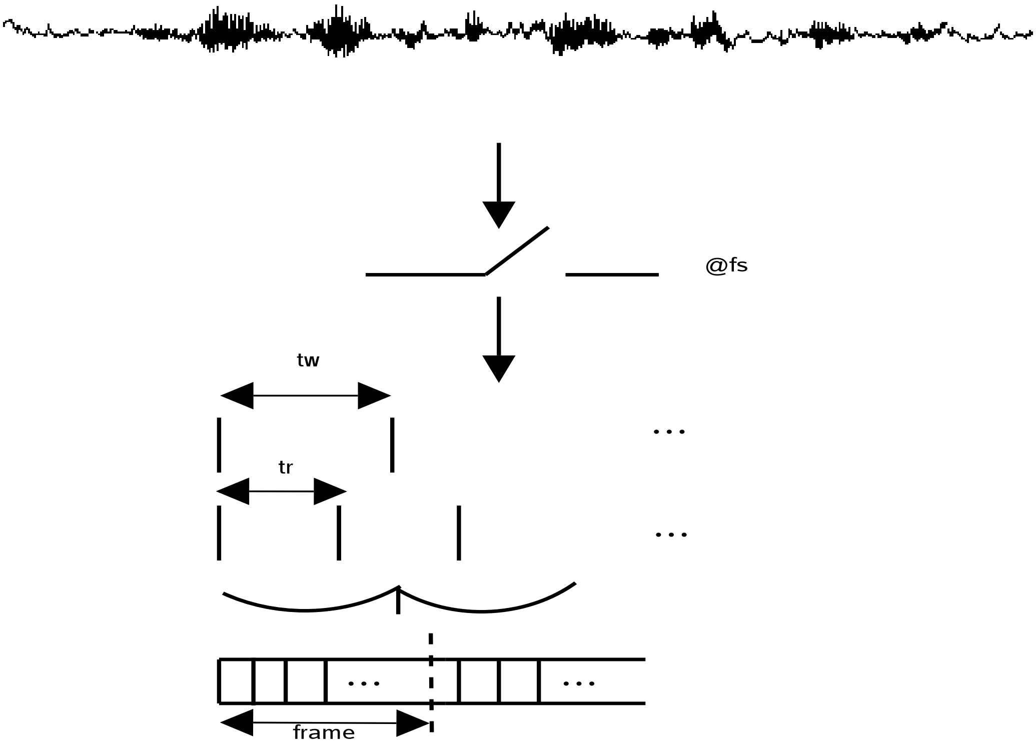

Figure 1.2: Parameterization of speech signal

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,

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

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
$\iff$	if and only if
$\implies$	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, $\sum_{l=1}^{k} \pi_{r\,l\,\theta_i} N(O_r \,|\, \mu_{r\,l\,\theta_i}, \Sigma_{r\,l\,\theta_i})$. Where $\pi_{rl\theta_i}$ is the l^{th} mixture weight, and $\mu_{r\,l\,\theta_i}$, $\Sigma_{r\,l\,\theta_i}$ 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

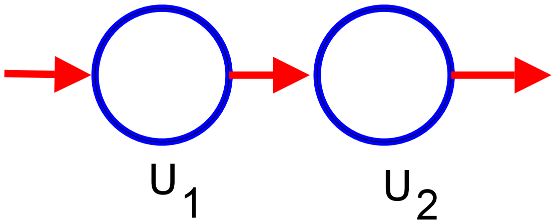

Figure 1.3: Illustration of probabilistic space containing ***rvs*** U

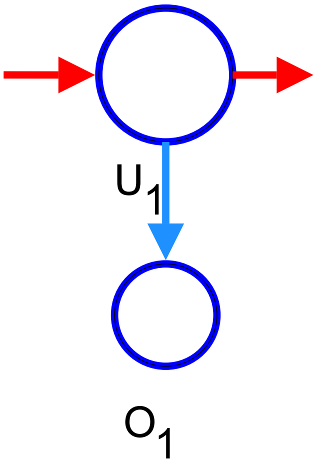

Figure 1.4: Illustration of probabilistic space containing ***rvs*** U and O

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],

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

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

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

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,

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

[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,

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

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

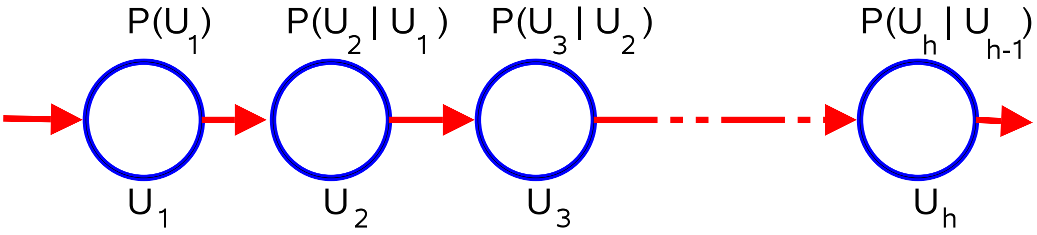

Figure 2.1: 1st order Markov chain stochastic graph

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.

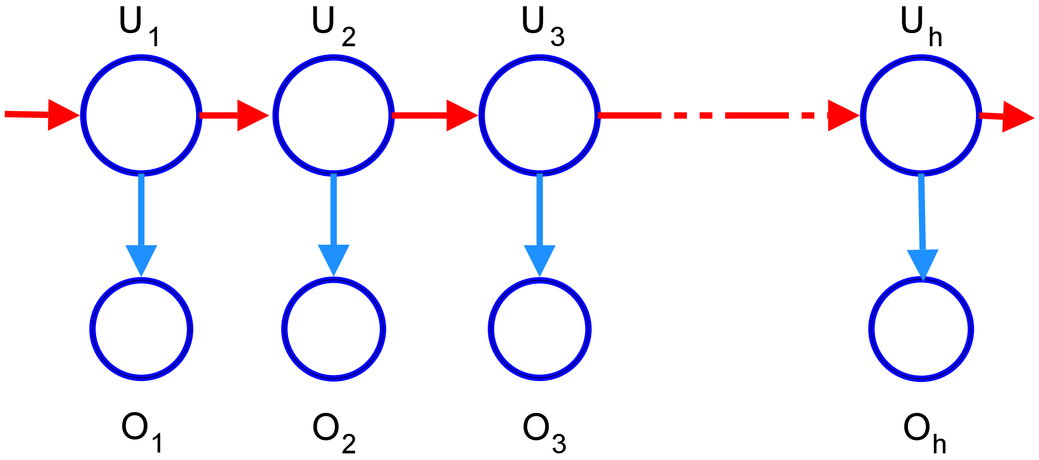

Figure 2.2: Hidden Markov model with hidden variable U and observations O

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,

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

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 $\boldsymbol{\pi_0}$ represents the steady-state initial state probabilities and the final product factor, is indicative of the *emission probabilities*, or rather the probability 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, $U_i \longrightarrow U_{i+1} \longrightarrow U_{i+2}$,

$$\begin{aligned} & I(U_i U_{i+1} U_{i+2}) \\ & = I(U_i U_{i+2}) + I(U_i U_{i+1} | U_{i+2}) \\ & = I(U_i U_{i+1}) + I(U_i U_{i+2} | U_{i+1}) \\ & \\ & \iff U_{i+2} \perp\!\!\!\perp U_i \,|\, U_{i+1} \\ & \implies I(U_i U_{i+1}) \geq I(U_i U_{i+2}) \end{aligned} \tag{2.11}$$

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,

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

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,

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

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,

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

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, $KL\big(P(O\,|\,\theta_i) \parallel \sum_j P(O\,|\,\theta_j)P(\theta_j)\big)$.

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,

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

The solution to this form of *quadratic programming* problem is a linear solution of the form $\mathbf{A}\,\mathbf{x} + \mathbf{b}$, 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, $\mathbf{W} : \mathbf{w} \in \mathscr{R}^m$,

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

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. Effective 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

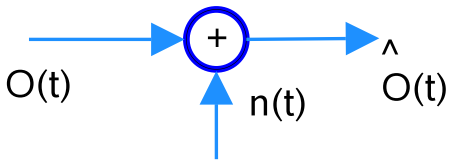

Figure 3.1: Speech with additive noise

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