ACOUSTIC MODEL ADAPTATION FOR RECOGNITION OF DYSARTHRIC SPEECH

BY

HARSH VARDHAN SHARMA

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Associate Professor Mark Hasegawa-Johnson, Chair
Professor Stephen Levinson
Professor Thomas Huang
Assistant Professor Torrey Loucks

Abstract

Speech production errors characteristic of dysarthria are chiefly responsible for the low accuracy of automatic speech recognition (ASR) when used by people diagnosed with the condition. The results of the small number of speech recognition studies, mostly conducted by assistive technology researchers, are a testimony to this statement. In the engineering community, substantial research has been conducted to find algorithms that adapt models of speech acoustics trained on one dataset for use with another. They are mostly mathematically motivated.

A person with dysarthria produces speech in a rather reduced acoustic working space, causing typical measures of speech acoustics to have values in ranges very different from those characterizing unimpaired speech. It is unlikely then that models trained on unimpaired speech will be able to adjust to this mismatch when acted on by one of the above-mentioned adaptation algorithms. The creation of acoustic models trained exclusively on pathological speech too is a task difficult to achieve: members of this population find it tiring to pursue physical activities for sustained periods of time, including speech production. While this makes speaker adaptation an approach worthy of pursuit, almost no research has been conducted so far on acoustic model adaptation methods for recognition of dysarthric speech.

This dissertation presents a study of acoustic model adaptation for recognition of dysarthric speech. First, it investigates the efficacy of a popular adaptation algorithm for dysarthric speech recognition. It then proposes an additional step in the adaptation process, to separately model 'normal' and pathology-induced variations in speech characteristics, and does so by trying to account for a recently proposed view of the acoustics of motor speech disorders in the clinical research community. Results show that explicitly addressing the population mismatch helps to increase the recognition accuracy.

To my mother Kumkum Sharma, and my father Krishna Kant Sharma.

Acknowledgments

I would like to first thank my adviser, Prof. Mark Hasegawa-Johnson; but for his generosity, patience, encouragement, invaluable advice and insightful suggestions, all that I have done during the course of my PhD program would not have been possible. My very first interaction with him in the summer of 2005, when I came to the University of Illinois as an undergraduate student from India, is what motivated me to return here for my graduate degree.

I am also very grateful to Prof. Stephen Levinson and Prof. Thomas Huang for agreeing to be on my doctoral committee. It is an honor to have had the privilege of presenting my work to you.

Thanks are also due to Prof. Torrey Loucks and Prof. Gary Weismer. As a student of engineering, jumping head-first into reviewing clinical research literature was scary. I cannot thank you enough for making the work accessible and helping me comprehend its implications for my research.

Thanks to all members of the Statistical Speech Technology group, the Beckman Institute, the Department of Electrical and Computer Engineering, and the University of Illinois, for providing me with the facilities and the environment to pursue my work.

I would like to thank Sharon Collins, secretary to Prof. Hasegawa-Johnson, for helping me out with various administrative tasks. I also thank Paritosh Garg, the network administrator for the IFP group at Beckman Institute. He was always available to handle my needs and requests for computing resources.

Thanks to all my friends whose company kept me cheerful during my pursuit of academics. It will be difficult to name all of you here, but surely you know who you are.

Finally, I would like to thank my parents. Their never-ending love and support is the very foundation of all my achievements. I cannot thank them enough. Ever.

Table of Contents

Chapter 1

INTRODUCTION

After more than two decades of research, automatic speech recognition (ASR) is a well-established and reliable human-computer interaction technology. The accuracy of the newest generation of large vocabulary speech recognizers, after adaptation to a user without speech pathology, is high enough to provide a useful human-computer interface especially for people who find it difficult to type with a keyboard.

For creating a speech recognizer for a particular speaker, there are two approaches: one is to create a speaker-dependent (SD) system by utilizing speech of that speaker alone to train the acoustic model; the other is to create a speaker-adapted (SA) system by first training the acoustic model in a speaker-independent (SI) fashion by using speech of several speakers, and then customizing the model to the characteristics of the particular speaker by using training examples of their speech to modify the model parameters. The parameter values do not get overwritten; they are adjusted using a regularized or constrained learning algorithm. Regularization or constraints allow the SA model to use far more trainable parameters per minute of training data without over-training the system.

Despite the advances in speech technology, their benefits have not been available to people with gross motor impairments mainly because these impairments include a component of *dysarthria*: a group of motor speech disorders resulting from disturbed muscular control of the speech mechanism, due to damage of the peripheral or central nervous system. Symptoms of dysarthria vary from talker to talker, but typical symptoms include strained phonation, imprecise placement of the articulators, incomplete consonant closure resulting in sonorant implementation of many stops and fricatives, and reduced voice onset time distinctions between voiced and unvoiced stops. Although the imprecise articulations of dysarthria are noticeable, and may even impair intelligibility, the articulation errors are usually neither random (unlike, for example, in the case of apraxia) nor unpredictable.

Dysarthria itself is often a symptom of a gross motor disorder, whose other symptoms often make it hard to use a keyboard and mouse. Published case studies have shown that some dysarthric users may find it easier to use an ASR system [1, 2, 3], instead of a keyboard.

One of the issues with developing ASR systems for dysarthric talkers is that speaking for long periods of time is very tiring. As a result it is difficult for a person with dysarthria to provide sufficient speech samples to train an SD ASR system. Speaker adaptation then seems a useful method to overcome this obstacle in developing dysarthric speech recognizers.

Although a substantially large amount of research has been conducted on methods for adaptation of ASR acoustic models, there has hardly been any study that evaluated their performance on recognition of dysarthric speech. However, even if one applied such adaptation methods, there exists a second obstacle: SI and SA systems of the kind used by talkers with no pathology are of less use to talkers with dysarthria, because the substitution errors characteristic of dysarthria dramatically increase word error rates. The goal of the study described in this dissertation is to test the hypothesis that *explicitly modeling the difference between unimpaired and dysarthric speech characteristics as a step in the adaptation technique should yield better recognition accuracy compared to using conventional adaptation methods as-they-are.*

This dissertation is organized as follows: in Chapter 2, a non-exhaustive overview of speech recognition technology is presented. It covers briefly the typical approach to modeling speech acoustics and language, and then reviews in somewhat greater detail acoustic modeling techniques, particularly with respect to model adaptation.

Chapter 3 provides an overview of the large amount of clinical research on motor speech disorders, especially acoustic analyses of dysarthric speech. For quite some time now, a debate has been going on in the clinical research community about an adequate theory of the acoustics of motor speech disorders. The long-held opinion of the majority of the community that “speech pathology reflects neuropathology” has been challenged by recent efforts of some researchers, particularly those of Gary Weismer’s group. The chapter ends by addressing this debate and attempting a justification of this study’s approach to investigating ASR system development for talkers with dysarthria.

Chapter 4 describes the speech corpus used for this study, and some preliminary experiments. These experiments were run to determine whether speaker adaptation should be pursued at all for recognition of dysarthric speech. These initial results indicate that SA systems do have the potential to achieve higher recognition accuracies than SD systems.

In Chapter 5, we present the main contribution of this study. We motivate the technique of background interpolation, based on the discussion in Chapter 3, and attempt a mathematical validation by exercising it on an artificial toy problem. The chapter ends with the mathematical derivation of equations for updating model parameters, when background interpolation is used in conjunction with MAP adaptation (a popular adaptation algorithm).

Chapter 6 presents the main experiments of this study. Results indicate that background interpolation achieves statistically significant improvements in recognition performance, across a range of intelligibility levels for speakers with dysarthria. Chapter 7 presents a more qualitative analysis of these results. It turns out that acoustic models adapted from an interpolated prior model do learn significantly different spectral representations of the modeled word-units.

Finally, Chapter 8 concludes with a review of the work's key findings and presents some possible directions for future work.

Chapter 2

BACKGROUND: AUTOMATIC SPEECH RECOGNITION

This chapter presents a concise review of research in the domain of automatic speech recognition. The emphasis is mostly on adaptation techniques for acoustic modeling.

2.1 Traditional ASR

Statistical ASR systems trace their beginnings to the work of Jelinek and others at IBM [4]. Such systems generally assume that the speech signal is a realization of some message encoded as a sequence of one or more symbols. To effect the reverse operation of recognizing the underlying symbol sequence given a spoken utterance, the continuous speech waveform is first converted to a sequence of equally spaced discrete parameter vectors which try to capture only the information relevant to speech and discard other acoustic information such as room and microphone effects and non-speech sounds. This sequence of parameter vectors is assumed to form an exact representation of the speech waveform on the basis that for the duration covered by a single vector (typically 10 ms or so), the speech waveform can be regarded as being spectrally stationary. Although this is not strictly true, it is a reasonable approximation. Much research has been conducted to determine the best (in terms of speech recognition accuracy) parameterization of speech. Typical parametric representations in common use are smoothed spectra or linear prediction coefficients plus various other representations derived from these [5, 6, 7]. Throughout this study, speech is parameterized in terms of Perceptual Linear Prediction (PLP) coefficients [7], a representation used by most ASR systems today.

The objective of a speech recognizer is to effect a mapping between sequences of speech vectors and the wanted underlying symbol sequences. In other words, the ASR system needs to come up with a sequence of N_W words $\mathcal{W} = \{w_k\}_{k=1}^{N_W}$ that is more likely (in a probabilistic sense) than any other to have generated the observed sequence of T PLP vectors $\mathcal{O} = \{\vec{o}_t\}_{t=1}^{T}$. This word sequence is therefore obtained as

$$\mathcal{W}^* = \arg\max_{N_W, \mathcal{W}} p\left(\mathcal{W}|\mathcal{O}\right)$$

which by application of Bayes' rule yields

$$\mathcal{W}^* = \arg\max_{N_W, \mathcal{W}} p\left(\mathcal{O}|\mathcal{W}\right) \cdot p\left(\mathcal{W}\right) \tag{2.1}$$

The $p\left(\mathcal{W}\right)$ component is called the *Language Model*, and the $p\left(\mathcal{O}|\mathcal{W}\right)$ component is called the *Acoustic Model*. The word sequence corresponding to the sequence of acoustic observations can be either modeled as-is at the word level itself or at a sub-word level as a sequence of sub-word units (e.g., phones, triphones, syllables, etc.); in the latter case, one makes use of a pronunciation dictionary to expand each word into its constituent sequence of sub-word units.

The ASR goal of determining $\mathcal{W}^*$ is made difficult by the fact that the boundaries between words cannot be identified explicitly from the speech waveform. This problem can be avoided by restricting the task to isolated word recognition: the acoustic observation sequence $\mathcal{O}$ corresponds to a single word from a fixed vocabulary. Despite the fact that this simpler problem is somewhat artificial, it nevertheless has a wide range of practical applications. Secondly, since dysarthric subjects find it physically exhausting to talk for long periods of time, the speech corpus that we make use of also consists of recordings of isolated words (described in more detail in Chapter 4). From here on, we thus focus on isolated word recognition: $N_W = 1$ and

$$\mathcal{W}^* = w^* = \arg\max_{w} p\left(\mathcal{O}|w\right) \cdot p\left(w\right) \tag{2.2}$$

The Language Model: Unless one has reason/evidence to believe that certain words are more likely than others to have been uttered by a talker, the prior probability $p\left(w\right)$ of a hypothesized word w is set to be the same for all the words in the task-vocabulary. Then

$$w^* = \arg\max_{w} p\left(\mathcal{O}|w\right) \tag{2.3}$$

The Acoustic Model: Given the dimensionality of $\mathcal{O}$, direct estimation of the joint conditional density $p\left(\mathcal{O} = \vec{o}_1, \vec{o}_2, \ldots | w\right)$ is not practicable. However, if a parametric model of word production such as a hidden Markov model (HMM) is assumed, then estimation from data is possible since the problem of estimating the class conditional observation densities $\{p\left(\mathcal{O}|w_k\right)\}_{k=1}^{N_V}$ (where N_V is the number of words in the task-vocabulary) is replaced by the much simpler problem of estimating the Markov model parameters. Most speech recognizers today are based on the HMM paradigm: each word w_k (or each sub-word unit, if one is modeling at the sub-word level) in the task-vocabulary is modeled by an HMM M_k — a finite state machine which changes state once every time unit — and each time t that a state j is entered, a speech vector $\vec{o}_t$ is generated from the probability density $b_j(\vec{o}_t)$, which is a mixture of multivariate Gaussians for most standard systems. The transition from state i to state j is also probabilistic and is governed by the discrete probability a_{ij}. Figure 2.1 shows an example of this process where the five state model moves through the state sequence $X = 1, 2, 2, 3, 3, 4, 4, 4, 5$ in order to generate the sequence $\vec{o}_1$ to $\vec{o}_7$. The entry and exit states $(1, 5)$ are non-emitting. This is to facilitate the construction of composite models: most systems use HMMs to perform modeling at the phone-level rather than word-level; as such, word-level models are constructed by stringing together phone-level HMMs for the constituent phones.

Let $\mathcal{S}_k = s_{k_1}, s_{k_2}, \ldots, s_{k_T}$ denote one of the many possible state sequences in M_k corresponding to the observation sequence $\mathcal{O}$. Then the probability of M_k generating $\mathcal{O}$ is obtained by summing up over all possible state sequences $\mathcal{S}_k$, the joint probability of $\mathcal{O}, \mathcal{S}_k$ conditioned on M_k:

$$p\left(\mathcal{O}|M_k\right) = \sum_{\mathcal{S}_k} \pi_{s_{k_1}} \prod_{t=1}^{T} b_{s_{k_t}}(\vec{o}_t) a_{s_{k_t} s_{k_{t+1}}} \tag{2.4}$$

where $\pi_{s_{k_1}}$ is the probability that the model starts off in state s_{k_1} before emitting the first observation vector. In practice, only $\mathcal{O}$ is known and the underlying state sequence $\mathcal{S}_k$ is hidden.

The recognition problem (Eq. 2.3) is then solved by considering the given set of models M_k corresponding to words w_k and setting

$$w^* \equiv \arg\max_k p\left(\mathcal{O}|M_k\right) \tag{2.5}$$

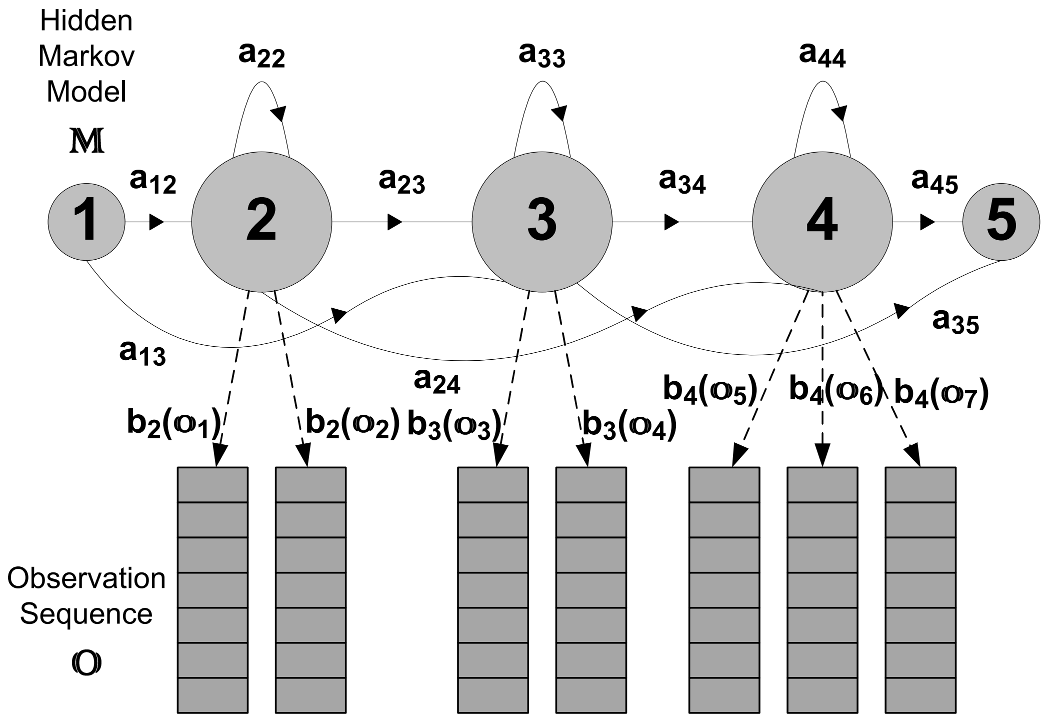

Figure 2.1: The Markov generation model

Making the acoustic model context dependent greatly improves the word error rate (WER; a measure of recognition accuracy) of an ASR system. We can make the acoustic model context dependent by including in the observation vector features that depend on neighboring frames: frequently the feature vector is augmented with the first and second order temporal differences of the features. Another way to make acoustic models context dependent is to use a triphone model. In this case, an HMM models a particular phone in context of the two phones immediately preceding and succeeding that particular phone. When training data sparsity becomes an issue (which is very likely to happen given that the possible number of triphones is large), the learned parameters of the triphone models can be shared across sets of triphones by tying them together in a data-driven way [8].

The Viterbi algorithm can be used to solve Eq. 2.5 and the Baum-Welch algorithm [9] can be used to find maximum likelihood (ML) estimates of parameters of the HMMs given some training data and an initial guess for the model parameters [10]. The Baum-Welch algorithm is a specific instance of the Expectation-Maximization (EM) framework [11]. These algorithms have been implemented in the HTK Toolkit [12] with speech recognition in mind.

2.2 Speaker Adaptation Methods for HMM-Based ASR

This section reviews some popular speaker adaptation schemes that can be applied to continuous density HMMs. These fall into three families based on Maximum A Posteriori (MAP) adaptation, linear transforms of model parameters such as maximum likelihood linear regression (MLLR), and speaker clustering/speaker space methods such as eigenvoices.

Speaker adaptation has been an area of speech recognition technology that has attracted much attention over the last decade. While SI speech recognition systems can show impressive performance, SD systems can provide an average WER a factor of two to three lower than an SI system if both systems use the same amount of training data. Hence the major rationale for investigating SA systems is that they promise to produce a final system that has desirable SD-like properties but requires only a small fraction of the speaker-specific training data needed to build a full SD system. Of course, as mentioned earlier, speaker adaptation seems useful for dysarthric speech recognition since dysarthric subjects will find it very exhausting to record large amounts of speech for training ASR systems.

SA systems operate in a number of *modes*. If the (word-level) transcription of the speaker-specific adaptation data is known then the adaptation is *supervised*, otherwise it is *unsupervised*: if the transcription is needed it must be estimated. While such an estimate may just be the errorful recognition output, some researchers have used confidence measures to ensure the adaptation process uses the most reliable material. Also adaptation modes are described as *static* (or block) in which all adaptation data is presented to the system before the final system is produced, or alternatively *dynamic* (or incremental) in which only part of the total adaptation data is available before use of the adapted system starts and the system continues to adapt over time. In this study, we concentrate on supervised static adaptation methods.

2.2.1 Maximum A Posteriori (MAP) adaptation

Let $\mathbf{\Lambda}$ denote the set of HMM parameters for an ASR system. If M is the number of d-dimensional multivariate Gaussians modeling the state-specific observation probability distribution, N the number of hidden states per HMM, and N_M the number of HMMs in the system, then

$$\begin{aligned} \mathbf{\Lambda} &= \left\{ \{\pi_i^n\}_i \ , \ \{a_{ij}^n\}_{i,j} \ , \ \{c_{il}^n, \vec{\mu}_{il}^n, \mathbf{\Sigma}_{il}^n\}_{i,l} \right\}_{n=1}^{N_M} \\ i, j &\in \{1, \ \dots \ , N\} \\ l &\in \{1, \ \dots \ , M\} \end{aligned} \tag{2.6}$$

where $\{c, \vec{\mu}, \mathbf{\Sigma}\}$ are respectively the mixture weight, mean vector and covariance matrix of the state-component-specific Gaussian density.

In MAP parameter estimation, the parameters are set at the mode of the posterior distribution $p\left(\mathcal{O}|\mathbf{\Lambda}\right) \cdot p_0\left(\mathbf{\Lambda}\right)$ where $p_0\left(\mathbf{\Lambda}\right)$ is the prior distribution of the HMM parameter set $\mathbf{\Lambda}$. The use of the prior distribution in MAP estimation means that less data is needed to get robust parameter estimates and hence it is a useful and widely used technique in speaker adaptation.

It is convenient if the prior density is from the same family as the posterior distribution (the conjugate prior) if it exists. For mixture Gaussian HMMs such a conjugate prior of finite dimension does not exist and an alternative approach is usually used: the key idea here is to interpret a finite mixture density with M components as one associated with a statistical population that is a mixture of M component populations [13, 14]. Doing so permits us to use conjugate prior densities individually for HMM parameter subsets, as described below.

Each of $\{\pi_i\}_i$, $\{a_{ij}\}_{i,j}$, and $\{c_{il}\}_{i,l}$ can be modeled with a Dirichlet prior distribution, which is the conjugate prior for the multinomial distribution. For instance, the Dirichlet prior for the mixture weights of state i, $\vec{c_i} = \{c_{il}\}_l$ can be written as

$$p_c(\vec{c_i}) \propto \prod_{l=1}^{M} c_{il}^{\nu_{il}-1} \tag{2.7}$$

where $\{\nu_{il}\}_l$ are the prior's hyperparameters. If x_l is the number of observations generated by component l, then the posterior distribution of $\vec{c_i}$ can be written using the multinomial density and the above prior as

$$p\big(\vec{c}_i \big| \left\{x_l\right\}_l\big) \propto p\big(\left\{x_l\right\}_l \big| \vec{c}_i\big) \cdot p_c(\vec{c}_i) \propto \prod_{l=1}^{M} c_{il}^{x_l} \cdot \prod_{l=1}^{M} c_{il}^{\nu_{il}-1} \tag{2.8}$$

which is a Dirichlet distribution with hyperparameters $\left\{\hat{\nu}_{il} = \nu_{il} + x_l\right\}_l$.

Similarly, the parameters of each Gaussian component l of state i $\{\vec{\mu}_{il}, \mathbf{\Sigma}_{il}\}$ can be modeled with a suitable prior distribution: a normal-Wishart joint density (for the case of full covariance matrix), or a set of Gamma-normal joint densities (for the case of a diagonal covariance matrix, since now each dimension of the observation vector is being modeled independently). For instance, in the case of D-dimensional diagonal covariance Gaussian with $\mathbf{\Sigma}_{il} = \mathrm{diag}(r_{il_d}^{-1})$, the overall prior for the Gaussian's parameters (for state i, component l) is obtained as

$$p_{\mu,r}\big(\vec{\mu}_{il}, \mathbf{\Sigma}_{il}\big) \propto \prod_{d=1}^{D} r_{il_d}^{\alpha_{il}-\frac{1}{2}} \cdot \exp\big\{-\beta_{il_d} r_{il_d} - \frac{\tau_{il} r_{il_d}}{2}(\mu_{il_d} - \mu_{il_{0_d}})^2\big\} \tag{2.9}$$

where $\{\alpha_{il}, \tau_{il}, \{\beta_{il_d}\}_d, \vec{\mu}_{il_0}\}$ is the prior's hyperparameter set. The observations $\{\vec{y}_t\}_{t=1}^{T}$ are governed by the Gaussian distribution $\mathcal{N}(\vec{y}; \vec{\mu}_{il}, \mathbf{\Sigma}_{il})$. The posterior distribution can then be written as

$$p\big(\vec{\mu}_{il}, \mathbf{\Sigma}_{il} \big| \left\{\vec{y}_t\right\}_t\big) \propto p\big(\left\{\vec{y}_t\right\}_t \big| \vec{\mu}_{il}, \mathbf{\Sigma}_{il}\big) \cdot p_{\mu,r}\big(\vec{\mu}_{il}, \mathbf{\Sigma}_{il}\big) \tag{2.10}$$

It can be shown that this posterior distribution is also a product of D Gamma-normal joint densities (i.e., of the form in Equation 2.9) with a hyperparameter set $\left\{\hat{\alpha}_{il}, \hat{\tau}_{il}, \{\hat{\beta}_{il_d}\}_d, \hat{\vec{\mu}}_{il_0}\right\}$ where

$$\hat{\alpha}_{il} = \alpha_{il} + \frac{T}{2} \tag{2.11}$$

$$\hat{\tau}_{il} = \tau_{il} + T \tag{2.12}$$

$$\hat{\vec{\mu}}_{il_0} = \frac{\tau_{il}}{\tau_{il} + T} \cdot \vec{\mu}_{il_0} + \frac{\sum_{t=1}^{T} \vec{y}_t}{\tau_{il} + T} \tag{2.13}$$

$$\hat{\beta}_{il_d} = \beta_{il_d} + \frac{\tau_{il_d}}{2}(\hat{\mu}_{il_{0_d}} - \mu_{il_{0_d}})^2 + \frac{1}{2}\sum_{t=1}^{T}(y_{t_d} - \hat{\mu}_{il_{0_d}})^2 \tag{2.14}$$

Having obtained the posterior distributions as above, the final step in MAP estimation involves setting the parameters to the modes of these posteriors. It must be noted that since the generation of observations is a hidden process in HMMs, we really do not have hard 'labels' describing which observations were generated by a particular component of a particular hidden state. This is where the EM algorithm comes to our rescue. Gauvain and Lee [13, 14] have shown that the use of prior distributions above with the standard auxiliary function leads to a MAP version of it, and that this MAP version is identical in form to the original auxiliary function. To borrow the notion of conjugateness, conjugate priors thus lead to conjugate auxiliary functions.

For instance, the MAP estimate of the Gaussian mean vector for state i, component l with prior mean $\vec{\mu}_{il_0}$ is

$$\hat{\vec{\mu}}_{il} = \frac{\tau}{\tau + \sum_{t=1}^{T} P\left(i, l|\vec{o}_t\right)} \cdot \vec{\mu}_{il_0} \; + \; \frac{\sum_{t=1}^{T} P\left(i, l|\vec{o}_t\right) \cdot \vec{o}_t}{\tau + \sum_{t=1}^{T} P\left(i, l|\vec{o}_t\right)} \tag{2.15}$$

where τ is the regularization meta-parameter which governs the weightage given to the prior mean with respect to the ML estimate of the mean from the adaptation data; $\vec{o}_t$ is the adaptation data observation vector at time t; and $P\left(i, l|\vec{o}_t\right)$ is the probability that the observation $\vec{o}_t$ was generated by state i, component l. One can see that this estimate is as per the one in Equation 2.13 except that soft 'labels' have been applied to the observations: each observation was generated by each state and component with a particular probability. Similar formulae can be used to also update the transition probabilities, covariance matrices and mixture weights in the system [13, 14]. The hyperparameters (of the prior distributions) that are generally used are the SI model parameters (empirical Bayes approach). Typically, values of τ between two and twenty are used.

One key advantage of the MAP approach is that as the amount of training data increases towards infinity the MAP estimate converges to the ML estimate. Its main drawback is that it is a local approach to updating the parameters; i.e., only parameters that are observed in the adaptation data will be altered from the prior value. As a result, MAP adaptation can be slow. Shinoda and Lee [15] tackled the adaptation speed issue by organizing all the Gaussians in the system into a tree structure, and then recursively computing a mean offset and a diagonal variance scaling term for each layer of the tree, starting at the root node (which contained all the Gaussians) and then descending the tree. At each level of the tree, the distribution from the node above was used as the MAP prior. They showed that structural MAP increases the adaptation speed while converging to the MAP solution as the amount of adaptation data is increased.

2.2.2 Linear transformation based methods for adaptation

Another approach to the problem of speaker adaptation is to estimate a linear transformation of the model parameters (or sometimes, the observation feature vectors), to construct a more appropriate model. The advantage is that the same transformation can be used for a large number of (or even all) Gaussians in an HMM system and this sharing of transformation parameters provides a route towards rapid adaptation. This section reviews Maximum Likelihood Linear Regression (MLLR) and some of its variants.

Maximum Likelihood Linear Regression

In standard MLLR [16], the Gaussian mean vectors are updated as per

$$\hat{\vec{\mu}} = \mathbf{A}\vec{\mu} + \vec{b} \tag{2.16}$$

where $\mathbf{A}$ is a $d \times d$ matrix and $\vec{b}$ is a $d \times 1$ vector, d being the dimensionality of the observations. This equation is more often written as

$$\hat{\vec{\mu}} = \mathbf{W}\vec{\xi} \tag{2.17}$$

where $\mathbf{W}$ is a $d \times (d+1)$ matrix and $\vec{\xi} = [1\ \vec{\mu}^T]^T$ is the extended mean vector.

In MLLR, the transformation matrix $\mathbf{W}$ is estimated such that the likelihood of the adaptation data is maximized. It has been shown [16] that there is a closed form solution to the $\mathbf{W}$ matrix estimation problem using the EM algorithm. Furthermore, under certain circumstances (where the initial models can provide good Gaussian-frame alignments) only a single iteration of EM is required to estimate the matrix. Usually the transformation matrix is tied over a number of Gaussians. This transform sharing can allow all the Gaussians in a system to be updated with only a relatively small amount of adaptation data.

However, there is a tradeoff between robust adaptation via a global transform and using precise transforms that apply to a smaller number of (e.g., phone-specific) Gaussians. One solution that allows a good compromise to be drawn is to use a Regression Class Tree [17]: the Gaussians that are close in acoustic space are clustered together and undergo the same transformation (these groups are known as base classes). If the clustered components are then arranged into a tree structure (with all at the root node), then, depending on the amount of adaptation data available, the tree may be descended to an appropriate depth and a set of transformations generated where each transformation will be for a set of base classes.

While the most important speaker specific effect concerns the Gaussian means, the Gaussian variances can also be updated [18, 19]. The variance transforms $\mathbf{H}$ are estimated after the mean transforms have been estimated. Originally the form

$$\hat{\mathbf{\Sigma}} = \mathbf{L}\mathbf{H}\mathbf{L}^T \tag{2.18}$$

was used where $\mathbf{L}$ is the Choleski factor of the original covariance matrix $\mathbf{\Sigma}$. For the case of a diagonal variance transform (with a simple bias for the mean) this is the same as the variance transform suggested in [20].

A variance transform of the form

$$\hat{\mathbf{\Sigma}} = \mathbf{H}\mathbf{\Sigma}\mathbf{H}^T \tag{2.19}$$

is proposed in [19] which has the advantage that it can be applied efficiently by transforming the mean parameters and the observations even for full variance transformations. However the transformation elements need to be estimated using an iterative procedure given the sufficient statistics.

Typically means-only MLLR gives a 15% reduction in WER on large vocabulary clean speech tasks over the most accurate SI models available, using about a minute of adaptation data; and SD performance can often be achieved with perhaps thirty minutes of speech and many adaptation transforms [16].

Constrained MLLR

The MLLR formulation described above estimates independent transforms for the means and the variances. The constrained transform case (introduced in [21] for the diagonal transform case and extended in [19] to full transforms) is of the form

$$\hat{\vec{\mu}} = \mathbf{A_c}\vec{\mu} - \vec{b}_c \tag{2.20}$$

$$\hat{\mathbf{\Sigma}} = \mathbf{A_c}^T\mathbf{\Sigma}\mathbf{A_c} \tag{2.21}$$

This can be convenient since this is equivalent to transforming the observation vectors such that

$$\hat{\vec{o}_t} = \mathbf{A_c}^{-1}\vec{o}_t + \mathbf{A_c}^{-1}\vec{b}_c \tag{2.22}$$

noting that a factor of $|\mathbf{A_c}|$ is also needed when calculating the Gaussian likelihood. The maximum likelihood solution for this form requires iterative optimization given the sufficient statistics, but gives similar performance to using standard unconstrained MLLR with the same form of transformation matrix.

MLLR robustness

For adaptation methods relying on linear transformation(s), it is necessary to have sufficient data points to robustly estimate the transform(s). Performance even worse than that of SI systems can result if appropriate thresholds/forms of transforms are not used (due to over-training on the adaptation data).

Several solutions to this problem have been suggested and all increase the applicability of MLLR for rapid adaptation. Chesta, Siohan and Lee [22] proposed a MAP version of MLLR (MAPLR) where it was suggested that one make use of a prior from the family of elliptically symmetric distributions for the overall transformation matrix (which includes the bias vector in the matrix itself). In related work, they had made use of a special case of this family, the matrix variate normal density. The advantage of using this density is the existence of ML estimates of the prior density's hyperparameters; MAPLR was found to improve performance when very small amounts of data were available. As the number of adaptation utterances grew large, the MAPLR performance approached MLLR performance. Siohan, Myrvoll and Lee [23] proposed a MAPLR extension to the SMAP technique reviewed earlier: the idea was to hierarchically constrain the priors on the transformation matrices using a tree structure, in which the prior on a leaf node (which has a cluster of Gaussians associated with it, and to which the transformation is being applied) is obtained by propagating the root-nodes prior down the branch that ends in that leaf node. At the root node, the prior is a matrix variate normal distribution. However, by using Bayes rule at each node below the root node, one does not obtain a distribution from this family. Secondly, by propagating the obtained distribution down would keep increasing the number of terms in the sum-expression of the Bayes-posterior. So, this posterior at a particular node was approximated by a distribution from the matrix variate normal family that was "closest" to the Bayes-obtained density in the KL-divergence sense. Experiments with non-native speaker adaptation showed that SMAPLR outperformed the MLLR and MAPLR methods.

Both MAPLR and SMAPLR used a MAP-style estimation approach for MLLR parameters. Alternatively a variant of the EM algorithm that optimizes a discounted likelihood criterion and does not quickly overtrain was suggested by Gunawardana and Byrne [24]. DLLR was also found to improve robustness for small amounts of adaptation data when many transforms were to be trained.

2.2.3 Eigenvoice adaptation

MAP and MLLR families of adaptation methods do not explicitly use information about the characteristics of an HMM set for particular speakers. The simplest instance of such an approach is the use of gender dependent models which are widely used in SI systems. Traditional speaker clustering (e.g., [25]) goes a step further and estimates HMMs for a number of speaker groups. However the problem with this type of approach is that by taking hard decisions about speaker type, the training data is fragmented and it is possible to make a poor choice of speaker group when in use.

Recently there has been interest in the eigenvoice technique [26] (EV) which can be viewed as a generalization of the speaker clustering idea. The EV technique forms a weighted sum of "eigenvoice" HMMs, and uses this interpolated model to represent the current speaker. The parameters of the eigenvoice models that are estimated can be viewed as representing the axes of a "speaker space" and then the model for a particular speaker is found by estimating the appropriate point for the speaker in this speaker space. The canonical speakers (eigenvoices) are found using principal component analysis (PCA) of sets of "supervectors" constructed from all training set speakers' SD HMM systems. The eigenvoices with the largest eigenvalues are chosen as a basis set. Specifically, if $\mathbf{\Lambda}^k$ represents the SD HMM set for speaker k (see Eq. 2.6), then one first obtains D eigenvoice models $\{\mathbf{\Lambda}_{e_r}\}_{r=1}^{D}$ using PCA; next, the SA HMM set for a test speaker u is obtained as

$$\mathbf{\Lambda}^u = \sum_{r=1}^{D} \alpha_r \cdot \mathbf{\Lambda}_{e_r} \tag{2.23}$$

During adaptation the maximum likelihood eigen-decomposition algorithm as proposed in [26] is used to estimate the weights $\{\alpha_r\}_{r=1}^{D}$.

Kuhn et al. [26] evaluated the EV technique for a small vocabulary task using simple HMM models and produced impressive performance with small amounts of data. Unfortunately for large HMM systems (with several thousand Gaussians) the construction of separate HMM systems for all speakers and subsequent PCA analysis is particularly difficult. There are two main issues here: firstly, if mixture distributions are used then these must be "aligned" between the various sets of models; secondly, the large number of parameters to estimate in the full (context-dependent triphone based) SD models can result in both estimation issues and storage problems. The alignment issue can be solved by obtaining the SD supervectors from each speaker's SA model obtained by adapting an SI model trained on speech of all speakers in the training set. The estimation issue which makes the PCA difficult (eigen-decomposition of a huge supervector-covariance matrix is very computationally expensive) can be resolved by performing probabilistic PCA instead [27].

The EV approach can be effective for a small amount of adaptation data but the gains available with added data are more limited and in such cases a technique such as MAP may be preferable. Indeed, Sproat et al. [28] have demonstrated that MAP used together with another speaker adaptation algorithm can decrease the WER.

Chapter 3

BACKGROUND: MOTOR SPEECH DISORDERS

Several neurological diseases produce symptoms of disordered speech production. These symptoms indicate deficits in the "control" of any or some or all levels of the human speech production mechanism. These deficits are here referred to as "motor speech disorders". This chapter gives an overview of the study of dysarthria, reviews literature on the related acoustic analyses, and ends by attempting a justification of the approach this study took to investigating ASR system development for speakers with dysarthria.

3.1 The Mayo Clinic System of Classifying Dysarthrias

Darley, Aronson, and Brown of Mayo Clinic can be reasonably credited with the beginning of systematic investigations of motor speech disorders [29, 30]. Prior to the publication of their twin papers in 1969, there had been a few infrequent studies of various neuromotor speech disorders (Zentay [31, 32]; Canter [33, 34, 35]; Lehiste [36]); however, no rationale had been presented for the study of dysarthria (other than the clinical one). Darley et al.'s study provided this rationale in the form of a hypothesis regarding the localizing value of perceptual impressions of dysarthric speech. They used the term *dysarthria* as "a collective name for a group of speech disorders resulting from disturbances in muscular control over the speech mechanism due to damage of the central or peripheral nervous system. These disorders were characterized by problems in oral communication due to the resulting paralysis, weakness, or incoordination of the speech musculature." Darley and colleagues hypothesized the association of different types of neurological pathology with unique kinds of speech production phenomena which in turn would be revealed in the aforementioned perceptual impressions. Based on their extensive clinical experience with dysarthric patients, they chose 38 dimensions of speech production performance to quantify those impressions; with each dimension's prominence represented on a seven-point scale. These dimensions, listed in Table 3.1, were capable of providing a comprehensive profile of the speech production deficit in neurogenic speech disorders.

Table 3.1: Perceptual dimensions used in the Mayo Clinic studies of dysarthria by Darley et al. [29]

Articulation dimensions	**Respiration dimensions**
imprecise consonants	forced inspiration-expiration
irregular articulatory breakdown	audible inspiration
phonemes prolonged	grunt at end of expiration
phonemes repeated	
vowels distorted	
Prosodic dimensions	**Voice-Quality dimensions**
rate	strained-strangled voice
variable rate	harsh voice
increase of rate overall	hoarse voice (wet)
increase of rate in segments	breathy voice (continuous)
reduced stress	breathy voice (transient)
excess and equal stress	voice stoppages
phrases short	hypernasality
intervals prolonged	hyponasality
short rushes of speech	nasal emission
inappropriate silences	
Pitch dimensions	**Loudness dimensions**
pitch level	loudness level (overall)
pitch breaks	alternating loudness
monopitch	monoloudness
voice tremor	excess loudness variation
	loudness decay
"Overall" dimensions	
intelligibility	
bizarreness	

These perceptual dimensions were then combined in various ways to produce unique *clusters* and Darley et al.'s hypothesis was confirmed not on the basis of the individual dimensions, but on this unique clustering of the multiple dimensions. Table 3.2 lists these clusters along with their respective perceptual dysarthria types. They also studied several disorders in which the neurological deficit was more diffuse than in the five dysarthrias listed in

Table 3.2: Clusters of perceptual dimensions reported by Darley et al. [30] for the five dysarthria types

Flaccid:	phonatory incompetence; resonatory incompetence; phonatory-prosodic insufficiency
Spastic:	prosodic excess; prosodic insufficiency; articulatory-resonatory incompetence; phonatory stenosis
Ataxic:	articulatory inaccuracy; prosodic excess; phonatory-prosodic insufficiency
Hypokinetic:	prosodic insufficiency
Hyperkinetic:	phonatory stenosis; prosodic insufficiency; resonatory incompetence; articulatory-resonatory incompetence; prosodic excess; articulatory inaccuracy

the table – for example, amyotrophic lateral sclerosis (ALS) is a disease in which both lower and upper motoneuron lesions are common; in keeping with their system, Darley et al. identified the dysarthria in ALS as *spastic-flaccid*. Likewise, multiple sclerosis (MS) was said to cause a dysarthria labeled as *spastic-ataxic* because it frequently involved cerebellar and upper motoneuron lesions. These labeling decisions were consistent with their belief that "... speech pathology reflects neuropathology" ([37]; page 229).

The classification system developed by Darley et al. (hereafter referred to as the Mayo Clinic system) was appealing because it significantly reduced the dimensionality of perceptual analysis. Secondly, it provided a reasonable guide for the clinicians attempting to modify the speech production deficit in dysarthria: they could identify the most prominent cluster in the speech production deficit profile, and devote therapeutic efforts there to make maximal gains in correcting/reducing the deficit.

3.2 Dysarthria and Acoustic Analyses

Although there is much to be said about the value of perceptual analysis (ease of interpretation, for instance), the relation between perceptual impressions of impaired speech and the underlying speech pathology is quite complex and poorly understood. Moreover, one's perceptual abilities are influenced by linguistic exposure such that the acoustic distinctions that are not phonemic in the ambient language (e.g., voiced versus pre-voiced sounds in English) are difficult to perceive [38]. Acoustic analysis of speech is quite advantageous for the following reasons: firstly, the acoustic output of the vocal tract can be thought of as a bridge between speech production and perception, and therefore the acoustic speech signal can shed light on both the mechanism associated with disordered speech and the effect of those problems on speech intelligibility. Secondly, the acoustic output of the vocal tract contains the product of the entire speech system's effort, rather than an isolated component of the apparatus. To the extent that a speech disorder is defined by its anomalous communication product, acoustic analysis may then prove to be valuable. Thirdly, acoustic analysis is noninvasive. Finally, acoustic analysis has the potential to provide insights for the purpose of acoustic modeling in ASR.

The literature on acoustic analysis for motor speech disorders is vast, so a condensed and selective overview of the same is given below. Studies pertaining to acoustics of vowel and consonant articulation are reviewed. Those concerned with aspects which are not easily defined (e.g., prosody and coordination) are not covered here. Also, studies that deal with obtaining insights into motor speech disorders (and their acoustics) using other modalities (e.g., kinematics of articulators, imaging of brain regions, etc.) are not covered here: multi-modal ASR is not a practical option for day-to-day use. Similarly, studies that investigate acoustic performance under manipulations (e.g., those of speaking rate and/or loudness) are not emphasized here, as ASR users typically prefer to use the technology in *habitual* speaking conditions.

3.2.1 Vowel articulation

Although the common view of speech intelligibility is that consonants contain most of the information-bearing elements in speech, there is accumulating evidence that vowel characteristics contribute heavily to speech intelligibility deficits [39, 40, 41, 42, 43]. In the Mayo studies too, the perceptual dimension *distorted vowels* was a prominent component of the clusters associated with several different dysarthrias (e.g., spastic, ataxic, hyperkinetic, spastic-flaccid). A common finding across studies of acoustic characteristics is that speakers with motor speech disorders often produce individual movements or changes in overall vocal tract shape with reduced displacements and velocities (see Weismer [44]; Table 7.3). This often results in their having a compressed phonetic working space for speech production. Vowel formant frequencies and characteristics of formant transitions may therefore serve respectively as static and dynamic indices of this space.

Formant specification can be used to make inferences about the vocal tract configuration. There is a rich tradition of using vowel formant frequencies as an index of the vocal tract shape [45, 46]. They form a useful low-dimensional description of vowels, and their relationships to vowel articulation are fairly well understood: (1) advancement of the tongue from a posterior to anterior location within the vocal tract results in an increase of the second formant (F2) frequency and a decrease of the first formant (F1) frequency; (2) lowering of the tongue from high to low positions within the vocal tract increases F1; and (3) elongation of the vocal tract by lip protrusion and/or lowering of the larynx tends to result in a decrease of all formant frequencies. In other words, the F2−F1 difference can be interpreted as tongue advancement/retraction and the F1 value as a measure of tongue height.

The most frequently reported abnormalities of vowel production in speakers with dysarthria include:

- *large deviations in formant frequencies*: Watanabe et al. [47] measured F1 and F2 for the five Japanese vowels in 5 men with ALS and 5 normal subjects; they found F1 values for /i/, /u/ to be significantly higher than normal, and F2 values for /i/, /e/ to be significantly lower than normal.

- *centralization of formant frequencies*: Ziegler and von Cramon [48] measured F1 and F2 for three German vowels, produced by 8 male subjects with closed head trauma and found that the vowel articulation was characterized by a centralized formant pattern – convergence of frequencies to formant targets for /ə/. More recently, Sapir et al. [49] have proposed formant centralization ratio (FCR) as an alternative metric to vowel space area (VSA; see next item below) for reliably distinguishing between dysarthric and unimpaired speech ("probably because of reduced sensitivity to inter-speaker variability and enhanced sensitivity to vowel centralization").

- *change in vowel space area*: Turner et al. [41] measured formant frequencies for the four 'corner' vowels (/i/,/æ/,/ɑ/,/u/) of English in 9 subjects with ALS and 9 age- and gender-matched controls. The area of the vowel quadrilateral in the F1-F2 plane was calculated at three different speaking rates (habitual, fast, slow) and the dysarthric speakers were found to exhibit smaller corner vowel space areas (CVSA) and less systematic changes in CVSAs as a function of speaking rate. It was hypothesized that lax vowels may be relatively unaffected by dysarthria, owing to the reduced vocal tract shapes required for these phonetic events. Tjaden et al. [50] studied the lax vowel space areas (LVSA) – vowel space area for the lax vowels (/ɪ/,/ɛ/,/ʊ/) – at three different speaking rates (habitual, fast, slow) for speakers with ALS, speakers with Parkinson's disease (PD), and healthy controls. LVSAs for speakers with ALS but not speakers with PD differed from those for the appropriate control group. Thus, only the results for the PD group support the hypothesis that LVSAs for speakers with dysarthria should be similar to those for neurologically normal speakers. Compared with the habitual condition, rate reduction was associated with an expanded LVSA for all of the healthy speakers but for only about half of the speakers with dysarthria. Other studies [51, 52] indicate that for some speakers with dysarthria, a slower-than-normal rate and increased vocal loudness are associated with an expanded CVSA relative to habitual speech, with rate reduction more strongly affecting CVSA. The relationship between vowel space area and perceptual impressions of intelligibility for speakers with dysarthria has also been explored, but the strength of the relationship varies among studies [41, 52, 53, 54] and one study found no relationship between vowel space and perceptual impressions of intelligibility [55].

- *shallower formant slopes and greater inter-speaker formant transition variability*: In terms of dynamic articulatory behavior, formant transitions (as indexed by formant slopes, and particularly that of F2), have been the focus of most studies. Weismer et al. [56] showed that the F2 slope, calculated for the rapidly changing segment of F2, was quite uniform across neurologically normal speakers. Weismer et al. [57] found in a study of 25 men with ALS and 15 controls that the former (a) pro-

> duced formant transitions having shallower slopes than transitions of normal speakers, (b) tended to produce exaggerations of formant trajectories at the onset of vocalic nuclei, and (c) had greater inter-speaker variability of formant transition characteristics than normal speakers. Kent et al. [40] found a moderately high correlation (a Spearman rank-order correlation coefficient of 0.86) between speech intelligibility on a single-word identification test and the average second-formant (F2) slope of selected test words, for a group of 25 men and 10 women with ALS. Some other studies have also demonstrated a strong relationship between speech intelligibility and average F2 slope [58, 59, 60, 61]. More recently, Kim et al. [42] have found that distributional characteristics of acoustic variables, such as F2 slope, could be used to develop a quantitative metric of severity of speech motor control deficits in dysarthria, when the materials are appropriately selected.

The acoustic-articulatory formant relationships, although useful, come with an associated challenge: formant frequencies vary with the length of the vocal tract, and therefore with the speaker's age and gender. The formant frequency patterns for a particular vowel as produced by a man, woman, and child are not identical; and this dependence/variation hinders the comparison of formant data from speakers representing different age-gender combinations. Secondly, the interaction between speech acoustics and perception is complex. As there are multiple cues to vowel distinction (e.g., vowel duration, formant frequency variation, fundamental (F0) frequency differences), a particular acoustic dimension will usually not be able to account for all the variance in listeners' judgments.

3.2.2 Consonant articulation

The impression of *imprecise consonants* is common to nearly all perceptual types of dysarthria identified by Darley et al. and presumably is influenced by a range of anomalies of consonant production (e.g., omissions, substitutions, distortions). Consonants are quite diverse in their perceptual, acoustic, and physiologic properties. A useful distinction can be made between sonorant and non-sonorant consonants: the former (liquids, glides, nasals) can be described by patterns of formants and antiformants in either steady-state or transitional segments. Hence, the data are similar to those for vowels. The latter (stops, affricates, fricatives) involve some kind of frication event: a burst or transient noise (stops), a brief noise interval (affricates), or a longer noise interval (for fricatives). No single acoustic measure, or even a small set of measures, is adequate for the purpose of describing all consonants. However, there is a possibility that select attributes are correlated with neurologic impairment and severity of speech disorder. Most published information on consonant production in dysarthria has focused on non-sonorants, and the acoustic variables that have been usually studied are the Voice Onset Time (VOT) and spectral moments. These are discussed in more detail next.

Voice Onset Time

Voice Onset Time (VOT) has been the subject of numerous investigations of both normal and pathological speech, largely on the assumption that this acoustic interval between the burst and the onset of periodic energy corresponds to the physiological interval between the release of the consonantal constriction and the onset of vocal-fold vibration. Hence, VOT is a possible index of intersystem coordination or timing. It is typically measured from the burst to the first full glottal pulse of the following vowel. VOTs are usually in excess of 35 ms for voiceless stops, and less than 20 ms for voiced stops. Affricates have greater VOTs than those observed in stops of the same voicing status.

VOT abnormalities, especially those in a region in which the voicing characteristics of the sound are ambiguous (e.g., between 20 ms and 40 ms for stops), may be a component of imprecise consonants.

VOT has been studied both for identifying subtypes of dysarthria and for investigating the relationship between VOT values and speech intelligibility. Kent, Netsell, and Abs [62] reported that lengthening of segments is a fundamental property of ataxic dysarthria, and severe dysarthric speech was marked by increased durations of all segments, including VOT. Caruso and Burton [63] compared VOT values in patients diagnosed with ALS, with eight age-matched controls. They found no significant differences among the mean VOTs of the six stop consonants, but the variability of VOT was greater in the ALS group for all consonants except /p/. Morris [64] measured VOT for voiceless stops (/p/,/t/,/k/) in twenty speakers with dysarthria: five spastic, five flaccid, five ataxic, and five hypokinetic. He found that mean VOT values increased as the position of occlusion moves posteriorly, with much overlap of values among the three consonants. He also found that the VOTs produced by the spastic dysarthrics for /t/ were significantly shorter than those produced by the flaccid and ataxic dysarthrics. This result is similar to that of Hardcastle, Morgan Barry and Clark [65], who reported that spastic patients produce shorter VOTs than normally speaking control subjects. Weismer [66] and Kent and Rosenbek [67], on the other hand, observed unusually long VOT durations in spastic dysarthria. Morris additionally reported that patients with flaccid and ataxic dysarthria exhibited significantly greater VOT variability than those with spastic and hypokinetic dysarthria. In the case of the flaccid dysarthria group, the inter-speaker VOT variability was large whereas intra-speaker VOT variations were similar to those produced by spastic and hypokinetic dysarthrics. In contrast, the speakers with ataxic dysarthria exhibited not only inter-speaker but also intra-speaker VOT variability. Ackermann and Hertrich [68] have found similar results showing that there is great VOT variability among ataxic patients.

Spectral Moments

A natural step in the study of non-sonorant articulation is characterization of the noise. One would then expect the spectra of stop bursts and fricative noises to be a valuable source of information. However, there is little literature concerning spectral analysis of consonant production in motor speech disorders. The lack of such studies can be explained by considering the measurement issues in quantifying consonant noise spectra: these spectra are typically multi-peaked, with energy spread widely throughout the frequency range, and there is no generally accepted means of summarizing noise spectra as a small number of quantitative indices.

Some investigators have used a categorical system to measure spectral shape, wherein spectral templates related to place of articulation are used to categorize consonant spectra [69, 70]. Shinn and Blumstein [71] demonstrated the use of the Stevens and Blumstein [69] template system in understanding stop consonant production in aphasia, but little other work has been done in this area. The application of the template system to persons with motor speech disorders is very time-consuming, requiring a human observer to generate and classify the spectra on an individual basis.

Forrest et al. [72] developed a quantitative, observer-free approach to measurement of consonant noise spectra: the spectrum is treated as a statistical distribution that can be described in terms of its mean, variance, skewness, and kurtosis – the first four moments. The mean quantifies the central tendency of the energy in the spectrum and appears to be sensitive to certain kinds of fricative misarticulation. The most likely deviation in dysarthric samples is a reduction of the first moment [73]. McRae et al. [53] used the spectral mean to show a reduction in articulatory working space, in people with PD compared with age-matched control subjects. There also appears to be some relationship between "articulatory precision" (as obtained from perceptual ratings) and the spectral mean: Tjaden and Turner [74] compared fricative spectra in the speech of speakers with ALS and neurologically normal controls. They showed that a significant amount of variance in listeners' perceptual judgments was determined by the difference in spectral means of /s/ and /ʃ/ (however, McRae et al. [53] and Tjaden and Wilding [52] did not find this relationship to be strong). The second moment (variance) expresses the distribution of spectral energy around the mean, and has been found to be useful for differentiating place of articulation for fricatives produced by neurologically healthy adults [75] and children [76]. The third moment (skewness) is a measure of the degree to which the spectral energy is tilted towards low or high frequencies; and the fourth moment (kurtosis) expresses the degree to which the spectrum has sharp peaks or is relatively flat.

3.3 Drawbacks of the Mayo Clinic System and the Call for a New Approach

In spite of the obvious heuristic value of the Mayo studies, the framework of the system is also constraining. One limitation is the assumed independence of the 38 perceptual dimensions. It is now known that these dimensions are neither pyschophysically uniform nor independent [77]. Rather, they are different in kind, are often interdependent, and are sometimes hierarchical. Secondly, the notion that "speech pathology reflects neuropathology" has been responsible for the continued use of diagnostic tests in the speech clinic that emphasize classic signs of neuropathology in the orofacial system, rather than systematic evaluation of the *speech production* deficit. Contrary to the Mayo perspective, the classic symptoms of certain neurological diseases do not necessarily appear to account for deviant speech production characteristics: for instance, Neilson and O'Dwyer [78, 79] showed that electromyograms obtained from orofacial structures in adult speakers with athetoid cerebral palsy and dysarthria were as stable across repetitions as those obtained from neurologically normal speakers. Thirdly, the Mayo Clinic system has cultivated a scientific concern with oromotor, nonverbal performance of persons with motor speech disorders. However, there is little evidence that such a concern has produced insights to the *speech production* disorder in dysarthria of speech. Weismer [80] has reviewed the 40-year history of empirical work in this area, and concluded that there is no compelling case for conducting the non-speech evaluations, if one is interested in understanding the speech production deficit. Fourthly, one can question the effectiveness, reliability and validity of the Mayo Clinic system, especially when used by different groups of raters. Zyski and Weisinger [81] attempted to replicate the 'localizing' finding of Darley et al. by blinding graduate students and trained speech-language clinicians to patients' neurological diagnosis, and asking them to identify dysarthria type based on perceptual analysis of speech samples. They found the reliability and classification accuracy to be quite low for the system to be suitable for clinical purposes. Zeplin and Kent [82] found that reliability varied across speech tasks and perceptual features. Bunton et al. [83] found that when average parameter ratings were in the mid-range rather than the extremes, lower reliability was obtained. More recently, Fonville et al. [84] reported on classification accuracy for neurologists and neurology trainees, and Van der Graaf et al. [85] reported on classification accuracy for neurologists, residents in neurology, and speech therapists: both studies found the accuracy to be quite low for classification by perceptual judgment alone. Finally, it is very likely that the typically large inter-speaker variability observed in almost any study of a particular perceptual type of dysarthria is due, at least partly, to variations in severity of speech disorder. In the original Mayo Clinic studies, there was considerable overlap in perceptual characteristics, across the types of dysarthria. The Mayo Clinic system does not address the possibility that variation in speech severity within a particular dysarthria type could explain as much variability in physiological, perceptual, and/or acoustic data as variation across dysarthria types. In fact, Kim et al. [86] found that classification accuracy (using spectral and temporal acoustic measures) by dysarthria type was typically worse than by disease type or severity level, and concluded that when severity is indexed by speech intelligibility scores, the measure is equally or more explanatory of variation in acoustic measures of speech as is the perceptual dysarthria type.

Given the above, and their observation that the existing literature on speech production characteristics in speakers with dysarthria suggests notable similarities across speakers with different disease and dysarthria types (e.g., slower-than-normal speaking rates, compressed vowel space, reduced formant transitions and articulatory velocities, reduced phonetic contrasts, etc.), Weismer and Kim [87] have argued for abandoning the Mayo Clinic's classification-based approach in favor of a taxonomy-based approach to the study of motor speech disorders. They propose that it is reasonable to expect the various neurological diseases to produce a core of *similar* speech symptoms, along with some set of symptomatic differences that distinguish the various types of motor speech disorders: "To use a statistical metaphor, these differences may be considered as the residuals of the model fit to the core symptoms."

3.4 Dysarthria and Automatic Speech Recognition

In this section, some previous studies on the performance of ASR systems for speakers with dysarthria are reviewed; and acoustic variability is discussed from the perspective of ASR system design: previous studies, and the need for addressing variability in acoustic modeling.

3.4.1 Previous studies

Only a small number of studies so far have investigated a variety of acoustic modeling techniques in terms of their usability for recognition of dysarthric speech. Ferrier et al. [88] examined the relationship between speech intelligibility level and ASR accuracy (for the DragonDictate system) by analyzing repeated recordings of the Pledge of Allegiance from 10 speakers with spastic dysarthria due to cerebral palsy. Results indicated that speech intelligibility ratings were generally positively correlated with ASR accuracy for the same reading passage over multiple trials. Perceptual review of the recordings indicated that speakers with more consistent articulatory productions had higher ASR accuracies. Chen et al. [89] studied the speech of a subject with intelligibility (as rated by human listeners) of only 15%, and found that after ten iterations of each word in a ten-word vocabulary, automatic word recognition accuracy was raised to 90%. Deller, Hsu and Ferrier tested dynamic time warping [90] and HMMs [91]. Polur and Miller studied the development of HMM-based small vocabulary (eight repetitions each of ten digits and fifteen 'command' words in English) SD systems for three male subjects subjectively classified by a trained clinician as moderately dysarthric [92, 93]. They found that an ergodic HMM with a slight left-to-right character (called a *transition-interpolated* HMM from hereon) provides lower WER than a standard left-to-right HMM, apparently because the transition-interpolated HMM is able to capture outlier events (e.g., syllable repetitions, phone insertions, pause insertions) as a backward or nonlinear progress through the intended word. The benefit of using ergodic modeling over left-to-right modeling in distorted speech applications with disruption events, pause events, and limited training data has also been noted earlier by Deller, Hsu and Ferrier [91]. Jayaraman and Abdelhamad tested an automatic neural network (ANN) [94], while Hasegawa-Johnson et al. tested support vector machines [95]. Polur and Miller demonstrated improved performance using a hybrid ANN/HMM [96].

Fewer studies exist on model adaptation for dysarthric speech: Raghavendra et al. [97] compared recognition accuracy of an SA system and an SD system. They found that the SA system adapted well to the speech of speakers with mild or moderate dysarthria, but the recognition scores were lower than those for an unimpaired speaker. The subject with severe dysarthria was able to achieve better performance with the SD system than with the SA system. These findings were also supported by Rudzicz [98] who compared the performance of SD and "SA" systems on the Nemours database [99] by varying independently the amount of data for training and the number of Gaussian components used for modeling the output probability distributions. The "SA" technique implemented is not speaker adaptation in the conventional sense: it uses the parameter values for the SI system as the starting point to train HMMs for a particular dysarthric speaker. In a training algorithm without regularization or constraint terms, it is possible for a system of this type to over-train, resulting in loss of accuracy on test data from the same speaker, and Rudzicz's results suggest that such over-training may have occurred in some cases. He further concluded that there was not enough data in the database to represent intra-speaker variation.

While these studies had reasonably good ASR accuracies, it should be noted that they all utilized corpora with very small vocabularies, often to control assistive technology such as environmental control systems. Recently, Sharma and Hasegawa-Johnson [100] investigated the development of medium vocabulary HMM recognizers for dysarthric speech of various degrees of severity with the following aims: (1) to test the performance of MAP-adapted systems relative to SD systems, for various degrees of dysarthria severity, (2) to test the performance of an SD system employing transition-interpolated HMMs relative to an SD system using strictly left-to-right HMMs, (3) to test the performance of a MAP-adapted system with transition-interpolated HMMs relative to an SD system having strictly left-to-right HMMs and, (4) to see if the results in the above three cases are essentially a function of the speaker's dysarthria severity. They found that performing transition-interpolation generally worsens recognition performance when compared to left-to-right HMMs. Performing both MAP adaptation and transition-interpolation results in higher recognition accuracy compared to the SD system with left-to-right HMMs, but adaptation-only systems have still better performance. This implies that state-transitions not accounted for in left-to-right HMMs do not capture (or capture rather poorly) the outlier events that differentiate dysarthric speech from unimpaired speech at the sub-phone level. The most interesting outcome of their study was that for subjects with very severe dysarthria, MAP adaptation was able to achieve substantial improvement in recognition accuracy, compared to the SD systems. This finding is significant in that it is contrary to the conclusions of previously published studies. These results therefore suggest that the severity of dysarthria as quantified by the subject's intelligibility rating is not a sufficient indicator of the relative performance of SD and SA systems.

3.4.2 Acoustic variability due to dysarthria characteristics: Impact on ASR

Generally speaking, acoustic variability has a substantial impact on ASR accuracy. Parker et al. [101] found the consistency of phonetic representation over time to be crucial for accurate recognition. Excessive instance-to-instance intra-speaker variability hinders the stabilizing of parameter values for the acoustic model, in the ASR training stage. The ASR accuracy goes down when the utterance being decoded is quite ‘far’ from the representation learned by the ASR system (as reflected by the learned model-parameters’ values) for it.

One of the first studies on acoustic variability was performed by Kent et al. [62]. They examined the acoustic characteristics of speakers with ataxic dysarthria. Speakers with dysarthria were found to exhibit increased intra-speaker variability (though not consistently) in lax vowel segment durations, compared to controls – especially the ones who had more severe dysarthria. Kent et al. [102] found in their study of 14 speakers with ataxic dysarthria, significant intra-speaker variability for the measures of F0, energy maxima and minima across repeated syllables, and for syllable duration. Ziegler et al. [103] studied syllable timing in speech of speakers with dysarthria due to traumatic brain injury (TBI) or cerebro-vascular accidents (CVA). They found that more severely impaired individuals demonstrated greater intra-speaker variability on syllable duration, than controls and speakers with mild dysarthria.

Blaney and Wilson [104] tried to explain the source of intra-speaker variability for dysarthric (3 individuals with ataxic dysarthria) and normal (6 age- and gender-matched controls) speakers, and to identify its relationship with ASR accuracy (for the DragonDictate, version 3.00 ASR system). Variability of acoustic measures (VOT, vowel duration, fricative duration, word stem duration, vowel formant frequencies) was determined using the mean, standard deviation and coefficient of variability (CoV = standard deviation/mean) values. Significant intra-speaker variability was noted for speakers with dysarthria, with regards to VOT for voiced plosives, vowel duration, and fricative duration. Speech from speakers with moderate dysarthria exhibited greater variability across all acoustic measures, compared to the speaker with mild dysarthria and the controls. In addition, minimal-pair categories were not preserved (merging of acoustic space, and violation of minimal-pair contrasts) and timing discrepancies were observed for word stem durations. Finally, some correlations were found between ASR accuracy and variability in VOT, vowel duration and fricative duration; but the authors noted that the small number of tokens may have contributed to the limited number of correlations.

More recently, Fager [105] investigated the durations of single words and sound types with acoustic analysis as well as the variability of word durations of ten participants with dysarthria due to TBI and ten control participants. The study also examined the relationships between word intelligibility and word duration, and between word intelligibility and variability for the participants with TBI. Results showed statistically significant differences on word and sound type durations between the dysarthric and control participants. Specifically, a pattern of doubling or near-doubling of durations was consistent across word and sound type durations. Extraneous features were identified from the acoustic signals of the dysarthric speakers and included pre-vocalizations, insertions, omissions, substitutions, and voicing of voiceless consonants; however, there was a lack of consistency in the occurrence of these features. When controlling for these features, the word and sound type durations remained significantly greater for the dysarthric compared to the control participants. Differences existed in standard deviations but not the CoV between dysarthric and control speakers. There was no correlation between word intelligibility and word duration or word intelligibility and variability. Fager concluded that "investigations with larger number of individuals with a wide range of dysarthria severity levels is warranted before a clear need to attempt to account for variability in (A)SR algorithms is identified."

3.5 Conclusion

From the literature reviewed so far, at least one thing is clear: speech production in speakers with dysarthria exhibits certain phenomena that do not fit current theories of unimpaired speech production. Large-scale studies of acoustic characteristics of motor speech disorders, and especially of how they are *similar* or *different* (see Section 3.3), within and across dysarthria or disease types, have not been undertaken; and a comprehensive theory of acoustics of motor speech disorders (or perhaps, one that explains characteristics of both unimpaired and pathological speech) has not been developed yet. In a perfect world, such a theory would provide crucial knowledge for ASR system design, for speakers with dysarthria. Regardless of the availability of such a theory, and given the findings of the studies discussed above, there are aspects of ASR system design that beg discussion. Some of these are briefly discussed below:

Firstly, the ASR system for speakers with dysarthria needs to handle the reduction in acoustic working space. We have seen that this reduction amounts to overlapping of classes in the frequency domain (as indicated by the reduction in vowel space area). So, the feature-space of the acoustic representation would probably need to be a transformation of the frequency domain where this overlap is absent and separability (at least from the perspective of the acoustic model's classification ability) of observations/features is restored. We have also seen that duration measures are markedly different in speech of speakers with dysarthria as compared to that of controls. If the feature-extraction module in the ASR system involved compacting a sequence of successive intermediate observation 'frames' into a final feature-space, then one other question worth investigating would be whether such a compaction scheme needs to, or should, account for these durational deviations. Yet another project to undertake can be the search for a feature extraction or feature transformation scheme, that absorbs the ill effects of excessive intra-speaker variability on stabilization of model-parameters' values that some of the above mentioned studies have discussed. Of course, reduction in the acoustic space and the resultant class-overlapping in feature-space is (most likely) not entirely attributable to disordered speech production: there have been studies that support and oppose the potential role speech perception plays here. For example, Tjaden and Sussman [106] studied anticipatory coarticulation and its inter-speaker variation in CVC words produced by controls and speakers with dysarthria, and found that "listeners appear to be tuned to similar types of information in the acoustic speech stream irrespective of the source or speaker, and any perceptual effects of inter-speaker variation in coarticulation are subtle." Weismer and Martin [107] on the other hand argue that "there is much in the speech perception literature to suggest that the listener has more trouble with a disordered speech signal than would be predicted solely from the mismatch between ‘normal’ and degraded acoustic-phonetic events.” Work by Liss and her colleagues [108, 109, 110, 111] also considers speech perception issues to be a critical component of an adequate theory of motor speech disorders.

Secondly, similar research questions can be asked for the acoustic modeling framework itself: issues of consistency of phonetic representation, class separability, and durational discrepancies may need to be addressed by the mathematical model that one uses (if not the HMM formalism), if the acoustic feature extraction does not address the same.

Thirdly, the ASR system designer would also need to address the possible presence of Fager’s extraneous features discussed above. One possible approach to account for such insertions, omissions, substitutions, and dysarthria-related “dysfluencies” (e.g., intra-word pauses) could be to model them in the ASR lexicon or in some sort of pathological “language” model.

Each of these questions is quite hard to answer, but is also a potentially promising avenue of ASR research. Although, there is no *a priori* reason to abandon/explore one or some of these questions, attempting to answer them in a single effort would be undoubtedly very challenging. The study in this dissertation has chosen to investigate ASR system development research by addressing intra-speaker variability in the acoustic model, on account of the following:

- The deviant phenomena that plague speech acoustics in speakers with dysarthria are mutually exclusive in subsets of this population. As such, it would be difficult to develop an ASR system development algorithm/recipe for dysarthria in general. A more speaker-specific approach is required, at least until the clinical research community obtains (and achieves consensus on) an adequate theory of acoustics of motor speech disorders. Addressing intra-speaker acoustic variability in the meantime, then, is worth attempting.

- Weismer and Kim [87] have recently proposed a starting point for the development of that adequate theory in the form of the hypothesis that “some *normal* bounds of variability can be determined for selected movement and/or acoustic measures from word and/or sentence productions, and that when the measure is made for a speaker with dysarthria, its *distance* from this normal range of variability will have some meaning. That distance is hypothesized to index something about the speaker's speech motor control capabilities." It is assumed that this hypothesis is worthy of pursuit by the clinical research community.

- Since we have already chosen to focus on developing SA systems for this population (because of the general inability of the subject to provide sufficient data to train acoustic models from scratch, as described in Chapter 1), there is the constraint of using the feature representation from the initial SI system. Hence, between feature-representation design and acoustic modeling, we choose the latter.

Chapter 5 presents the proposed approach for modeling Weismer and Kim's "distance from normal range of variability."

Chapter 4

THE UA-SPEECH CORPUS AND PRELIMINARY EXPERIMENTS

This chapter describes the speech corpus used for the experiments performed in this study, and presents results from some preliminary experiments. The latter were conducted to determine whether speaker adaptation is a direction worth pursuing for people diagnosed with dysarthria.

4.1 The UA-Speech Corpus of Dysarthric Speech

The studies reported in Chapters 2 and 3 either used (a) commercial off-the-shelf ASR software (which have models trained from several hours of unimpaired speech); (b) speech from four or fewer speakers; or (c) either the Whitaker database [112] or the Nemours database [99]. The Whitaker database contains 30 repetitions of 46 isolated words (10 digits, 26 alphabet letters, and 10 'control' words) and 35 words from the Grandfather passage produced by each of six individuals with cerebral palsy. The Nemours database contains 10 sentences read by each of 20 different speakers, representing a wide range of speech pathology diagnoses; only a fraction of the speakers show symptoms of spastic dysarthria.

To our knowledge, the first publicly available database suitable for training medium-vocabulary automatic dysarthric speech recognition for speakers with high, moderate, low, or very low intelligibility is the UA-Speech database, reported by Kim et al. in [113]. The UA-Speech corpus contains recordings of 16 subjects informally diagnosed with dysarthria. Each speaker recorded three blocks of words: each block contained the same 155 core words, plus 100 "uncommon words" that differed across blocks. The core words included the 10 digits ("zero" through "nine") – `D`, the 26 letters of the international radio alphabet ("alpha, bravo, charlie,...") – `L`, 19 computer commands ("command, enter, paragraph,...") – `C`, and the 100 most common words in the Brown corpus of written English ("is, it,...") [114] – `CW`. The uncommon words – `UW` – were selected from children's novels digitized by Project Gutenberg (e.g., *Wizard of Oz*, *Peter Pan*) to maximize phoneme-sequence diversity. Digits and common words were primarily composed of monosyllables, computer commands and radio alphabet letters of bisyllables, and uncommon words of polysyllabic words (more than half of the uncommon words were trisyllabic or longer). Each subject recorded a total of 765 words, including 455 distinct words. Most of the subjects in the corpus have spastic dysarthria (12 of the 16). This was done to generate consistency in specific dysarthric characteristics. Intelligibility assessment is described in [113].

Table 4.1 lists the speakers in the corpus along with their respective human listener intelligibility ratings (in percent) and intelligibility categories. All experiments in this study have made use of the UA-Speech corpus.

Table 4.1: Intelligibility characteristics of the speakers in the UA-Speech corpus

Speaker	**Average Intelligibility (%)**	**Intelligibility Category**
M01	17	Very Low
M04	2	Very Low
M05	58	Mid
M06	39	Low
M07	28	Low
M08	95	High
M09	86	High
M10	93	High
M11	62	Mid
M12	7	Very Low
M14	90	High
M16	43	Low
F02	29	Low
F03	6	Very Low
F04	62	Mid
F05	95	High

More recently, Rudzicz et al. [115] have acquired the TORGO database of dysarthric speech which includes aligned acoustic and articulatory data from seven individuals diagnosed with cerebral palsy or ALS, as well as from age-

and gender-matched control subjects. At the time of writing, this database is available by contacting the authors.

The following section describes some preliminary experiments in acoustic model adaptation for speech of speakers with dysarthria.

4.2 State-Transition Interpolation and MAP Adaptation

This study is reported in [100]. We investigated the development of medium vocabulary HMM recognizers for dysarthric speech of various degrees of severity with the following aims: (1) to test the performance of MAP-adapted systems relative to SD systems, for various degrees of dysarthria severity, (2) to test the performance of an SD system employing transition-interpolated HMMs relative to an SD system using strictly left-to-right HMMs, (3) to test the performance of a MAP-adapted system with transition-interpolated HMMs relative to an SD system having strictly left-to-right HMMs and, (4) to see if the results in the above three cases are essentially a function of the speaker's dysarthria severity.

Transition Interpolation: Figure 4.1 illustrates the topologies of strictly left-to-right (LR) and transition-interpolated (TI) HMMs with 3 emitting states. If $\mathbf{A} = \{a_{ij}\}$ is the N × N transition probability matrix for an N-state HMM, then we have for an LR HMM: for each state i, $0 < a_{ii}\ ,\ a_{i,i+1} < 1$; $a_{ii} + a_{i,i+1} = 1$ and $a_{ij} = 0$ for $j \neq i, i+1$. In other words, each emitting state has only two possible state-transitions: given the current state, the HMM either remains in the same state or moves into the succeeding state; it will not jump over states or go to a preceding state.

The TI model is an LR model which has non-zero transition probabilities for jumps and transitions to preceding states from a particular state (for emitting states). These probabilities are, however, small compared to self-transition and next-state-transition probabilities. A TI HMM is initialized as follows: for each emitting state i, $a_{ij} = \epsilon$ for $j \neq i, i+1$ where $0 < \epsilon << 1$; $a_{ii}\ ,\ a_{i,i+1} >> \epsilon$ and $\sum_{j=1}^{\mathrm{N}} a_{ij} = 1$. After this initialization, the transition probability matrix is re-estimated for SD systems using the standard Baum-Welch algorithm, and for MAP-adapted systems using the MAP variant of the Baum-Welch algorithm.

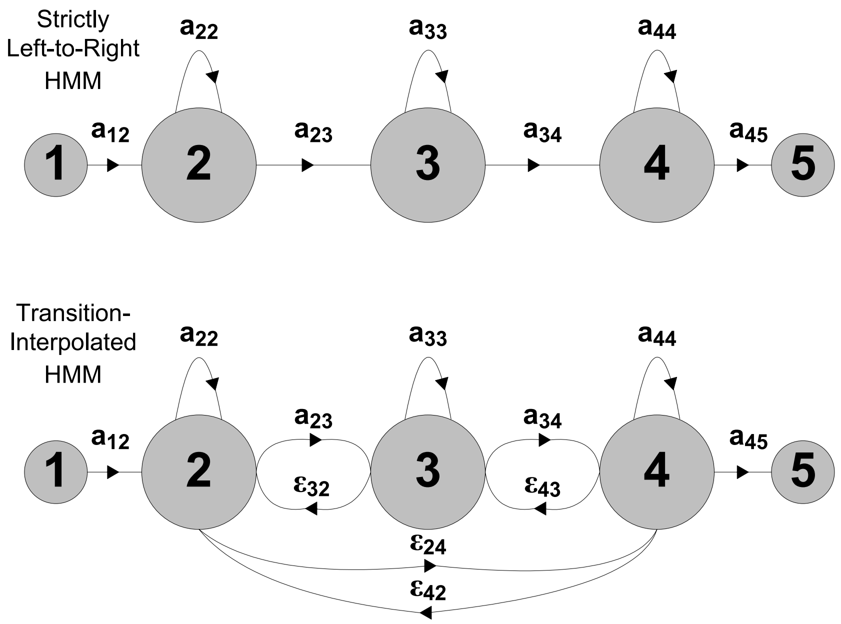

Figure 4.1: Difference between strictly left-to-right and transition- interpolated HMM topologies

These experiments utilized speech of 7 subjects from the UA-Speech database: M09, M05, M06, F02, M07, F03, and M04. For building the MAP prior SI system, the unadapted HMMs were trained on speech from the TIMIT corpus.

Table 4.2 lists the characteristics of the various system configurations that were studied: SD stands for speaker-dependent, MAP for MAP-adapted; LR implies use of strictly left-to-right HMMs, TI for transition-interpolated HMMs; 'm','v','w','t' respectively denote means, variances, mixture-component weights and transition probabilities.

Table 4.2: Summary of ASR system configurations

System (Type)	**HMM**	**Parameters adapted**
C00 (SD)	LR	—
C01 (SD)	TI	—
C11 (MAP)	LR	m
C12 (MAP)	LR	m,v
C13 (MAP)	LR	m,v,w
C14 (MAP)	LR	m,v,w,t
C15 (MAP)	TI	m,v,w,t

These systems were developed for each of the seven speakers and employed word-internal, context-dependent triphone HMMs, with three hidden states and observations modeled as mixture-of-Gaussians. Configuration C00 is the 'standard' SD system using LR HMMs, and is the baseline configuration for this study. For configurations C11 through C15, the SI systems trained on TIMIT employed left-to-right HMMs. For systems C15, the transition interpolation was performed after obtaining the SI TIMIT-trained left-to-right HMMs and before adaptation to the UA-Speech speaker's data: the original non-zero entries in the transition probability matrices were scaled down so that the sum of each row was unity after changing the zero-entries to ϵ. For each speaker, all of blocks 1 and 3 were used as training data (systems C00, C01) or adaptation data (systems C11-C15) and all of block 2 was used for testing. The SI system was trained on all of TIMIT's training data and was tested on speech of 32 randomly chosen speakers from its test data. The features extracted from the speech waveform were comprised of 12 PLP coefficients [7] for 25 ms Hamming-windowed segments obtained every 10 ms, plus the energy of the windowed segment. 'Velocity' and 'Acceleration' components were also calculated for this 13-dimensional feature, which finally results in a 39-dimensional acoustic feature vector.

The measure used for assessing the performance of the developed recognizers is the fraction of task-vocabulary words correctly recognized (in percent), defined in Equation 4.1. For isolated word recognition, it is also the word recognition accuracy (WRA).

$$WRA = \frac{\#\ words\ correctly\ recognized}{vocabulary\ size(\#\ words)} \times 100 \tag{4.1}$$

For each configuration, the number of Gaussian components in the state-specific observation probability densities was increased (in an iterative manner) in powers of 2, from 1 to 32 components (for C00 and C01) or 64 components (for C11-C15). In order to avoid over-tuning, the number of Gaussian components was constrained to be the same across all speakers. For the SD systems (C00 and C01), results are for HMMs with 2 Gaussian components per probability density. For the MAP-adapted systems (C11-C15), results are for HMMs with 32 Gaussian components per probability density: while training the SI TIMIT system, it was found that the phone recognition accuracy increased monotonically when going from 1 to 32 Gaussian components but decreased when going from 32 to 64 components.

Tables 4.3 and 4.4 list the WRA scores for the various system configurations developed. The speakers are listed in decreasing order of intelligibility rating.

Table 4.3: WRA scores for each speaker's configurations C00-C12

	System Configuration			
Speaker	**C00**	**C01**	**C11**	**C12**
M09	52.04	47.3	57.1	62.1
M05	35.52	33.7	31	39.4
M06	34.01	36.1	38.6	38.5
F02	35.06	32.8	20.8	26.9
M07	43.87	40.7	32	35.9
F03	12.61	11.3	17.4	22.2
M04	2.82	1.7	3.7	4.2

Table 4.4: WRA scores for each speaker's configurations C00, C13-C15

	System Configuration			
Speaker	**C00**	**C13**	**C14**	**C15**
M09	52.04	66.4	65.8	64.2
M05	35.52	45.2	44	38.1
M06	34.01	40.7	40.1	39.2
F02	35.06	30.4	29.7	26.6
M07	43.87	43	41.8	35.9
F03	12.61	27.7	26.2	25.7
M04	2.82	4.2	3.8	3.1

We see that SD systems with left-to-right HMMs (C00) have higher recognition accuracy than the SD systems with transition-interpolated HMMs (C01), for all speakers except M06. System C11 for a particular speaker, with adaptation of Gaussian means alone, performs either better or worse than both systems C00 and C01 for that speaker. System C12 with adaptation of Gaussian means and variances, has better recognition accuracy than both SD systems, for all speakers except F02 and M07 (worse than both SD systems). System C13 with adaptation of all parameters except transition-probabilities has the highest recognition accuracy for all subjects except F02 and M07 (highest among MAP-adapted systems only). System C14, which adapts all parameters including transition probabilities, always performs worse than the corresponding system C13, for all speakers. However, like system C13, it has better recognition accuracy than both SD systems for all speakers except F02 and M07. Finally, performing transition-interpolation and adaptation of all parameters (system C15) worsens the performance to below that of the corresponding system C14; additionally, C15 has better recognition accuracy than both SD systems whenever the corresponding C13 (and C14) system also performs better than them.

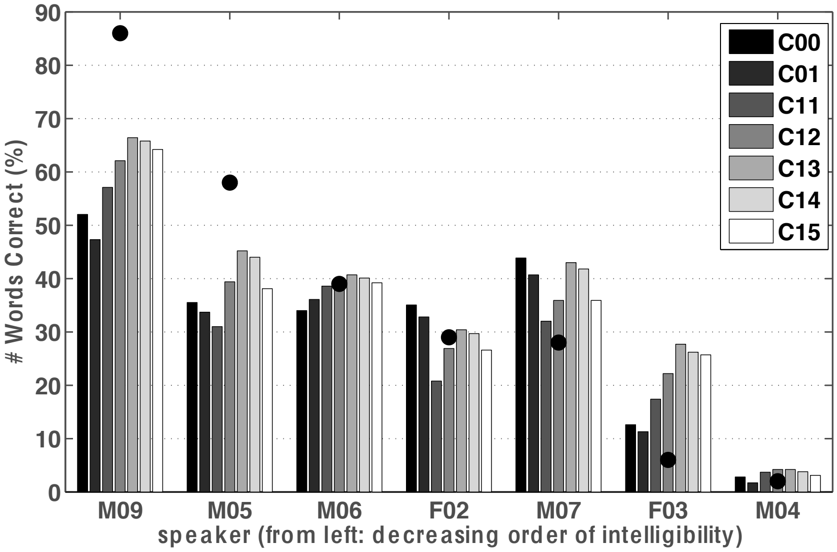

Figure 4.2: WRA scores for various system configurations (the black circles indicate speakers' human listener intelligibility ratings)

These results are plotted in Figure 4.2 along with the human listeners' intelligibility ratings of these speakers (the black circles). For speakers M09 and M05, system C13 with the best overall WRA score is still far from doing as well as human listeners. For the remaining subjects, it has however been able to do as well as or better than human listeners even when it performed worse than the corresponding SD systems (C00,C01): in fact, for speaker M06, it does better than human listeners when the SD systems do not.

Figure 4.3 plots, for all speakers, the WRA of system x ($x \in \{C01 - C15\}$), expressed relative to the WRA of system C00.

For speakers who have an intelligibility rating above 35% or below 25%, the MAP-adapted systems generally do better than their SD counterparts. System C01, with transition interpolation, performs worse than system C00 for all speakers except M06. The surprising result though is that for speakers

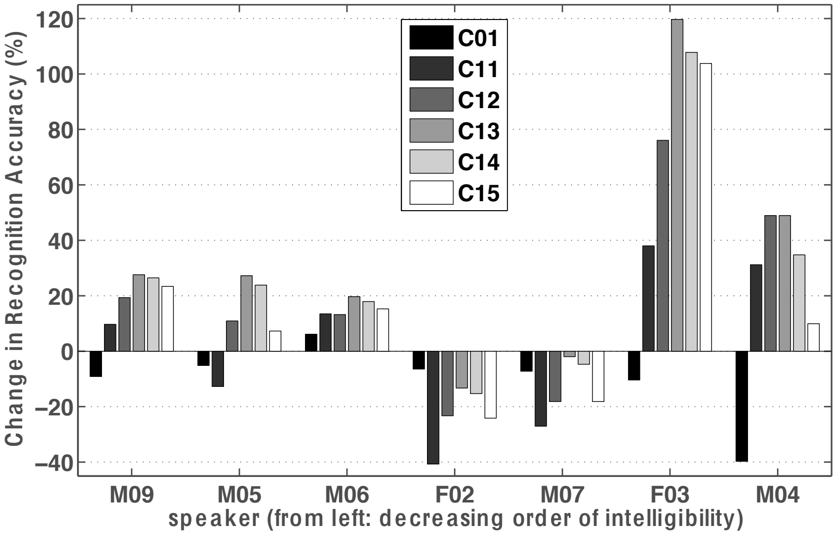

Figure 4.3: Percentage change in WRA scores for various system configurations relative to configuration C00's WRA score

with highly severe dysarthria (F03 and M04), MAP-adapted systems have substantially better recognition accuracies than their SD counterparts, when previous studies have indicated that for such subjects, SD systems perform better than speaker-adapted systems.

Conclusions: It was found that performing transition-interpolation generally worsens recognition performance when compared to left-to-right HMMs. Performing both MAP adaptation and transition-interpolation results in higher recognition accuracy compared to the SD system with left-to-right HMMs but adaptation-only systems have still better performance. This implies that state-transitions not accounted for in left-to-right HMMs do not capture (or capture rather poorly) the outlier events that differentiate dysarthric speech from unimpaired speech at the sub-phone level.

The most interesting outcome of our experiments is that for subjects that have very severe dysarthria, MAP adaptation was able to achieve substantial improvement in recognition accuracy, compared to the SD systems. This finding is significant in that it is contrary to the conclusions of previously published studies. These results therefore suggest that the severity of dysarthria as quantified by the subject's intelligibility rating is not a sufficient indicator of the relative performance of speaker-dependent and speaker-adapted systems.

Chapter 5

MODELING MISMATCH WITH BACKGROUND INTERPOLATION

In Chapter 2, some well-established techniques for adapting acoustic models were reviewed. They have been shown to perform well in data-rich situations but where the target populations were not drastically mismatched with that of the training data. This should not come as a surprise, because these techniques do not explicitly model the mismatch that exists between the speech characteristics of the target speaker population and those of the population used to train the to-be-adapted acoustic model. Of course, a model adaptation technique also requires substantial speech data to have good WER performance.

There is a limit to the amount of speech data that can be acquired from a speaker with dysarthria in order to train an ASR system for them – speaking for long periods of time is very tiring for members of this population. Also, in Chapter 3, it was seen that the acoustic characteristics are indeed very different for unimpaired speech and speech of speakers with dysarthria. The chapter ended with the suggestion that from the perspective of acoustic variability, one needs to explore a speaker-specific approach, for modeling the *distance* from the range of variability observed in unimpaired speech.

Considering the parameters of an ASR acoustic model (AM), every AM is a point in the space of AM parameters. Hence, if one wanted to obtain an AM at some 'distance' from another AM, one would also need to account for a 'direction' in which to go searching for that new AM. This study proposes to obtain a speaker-specific 'background' model and use that to determine the search direction in the AM parameter space. After reaching a suitable/desired point (AM) in this direction, adaptation is performed. In this sense, the task of modeling population mismatch can be viewed as one of designing a suitable prior AM for adaptation.

First, we define some notation that will be used in the remainder of this chapter. Let $\mathbf{\Lambda}$ denote the AM parameter set for the system in vector notation (same as $\mathbf{\Lambda}$ in Equation 2.6). In this study, the AM is a set of HMMs whose observation distributions are mixtures of multivariate Gaussian densities with diagonal covariance matrices. So, for a system with N_M N-state HMMs with M Gaussians per state, $\mathbf{\Lambda}$ has as its dimensions the initial state occupancy probabilities $\{\pi_i^n\}_i$, the transition probabilities $\left\{a_{ij}^n\right\}_{i,j}$, mixture weights, mean vector components, and variance vector components ($\{c_{il}^n, \vec{\mu}_{il}^n, \mathbf{\Sigma}_{il}^n\}_{i,l}$ respectively) — $i, j \in \{1, \ldots, N\}$; $l \in \{1, \ldots, M\}$; $n \in \{1, \ldots, N_M\}$. Speaker adaptation is performed in two stages: first, a model that accounts for the mismatch between speaker populations (see Sections 5.1.1 and 5.1.2) is obtained; in the second stage, this model is used as the prior or initial model (see Section 5.1.3) for the actual adaptation.

5.1 Background Interpolation: Formulation

5.1.1 Speaker background models

The Universal Background Model (UBM) is an effective and widely used framework [116] in the field of speaker verification when speaker-models are to be trained using limited per-speaker data. The UBM approach calls for pooling together data for all speakers to train a background model as a first step. This model is then adapted in a second stage, to each speaker using that speaker's data alone. Since speech from a large number of speakers is used to train the UBM, it can be a model with a high parameter count and still be mostly free from the risk of overfitting.

To develop a speaker-specific model of population mismatch, a speaker-dependent background (SDB) model, $\mathbf{\Lambda}_{SDB}$, is first created by training an HMM system using all speech of a particular speaker from the target population. This system is trained regardless of the actual words spoken in each utterance. The SDB does not learn any patterns that can discriminate between phones/words. It is a model of the general characteristics of the speaker from the target population. The intention behind using such a model is to capture aspects of time-frequency variation that depend on the speaker (rather than on what was spoken by him/her). As with the UBM, the SDB can have a high parameter count since all speech from the speaker is being used.

5.1.2 Combining speaker-independent and speaker-background models

Let $\mathbf{\Lambda}_{SI}$ denote the SI system trained on speech from a population that is very different from the target population in terms of speech characteristics. In our case, this would be the population of unimpaired speakers.

The explicit modeling of mismatch in AMs is now motivated. Figure 5.1 plots a fictitious posterior probability of the model parameters given the observations. Most algorithms for acoustic model adaptation start out with $\mathbf{\Lambda}_{SI}$ as the 'prior' for the parameter-set and try to reach a local maximum of the posterior probability, obtained at $\mathbf{\Lambda}^*_{SI}$.

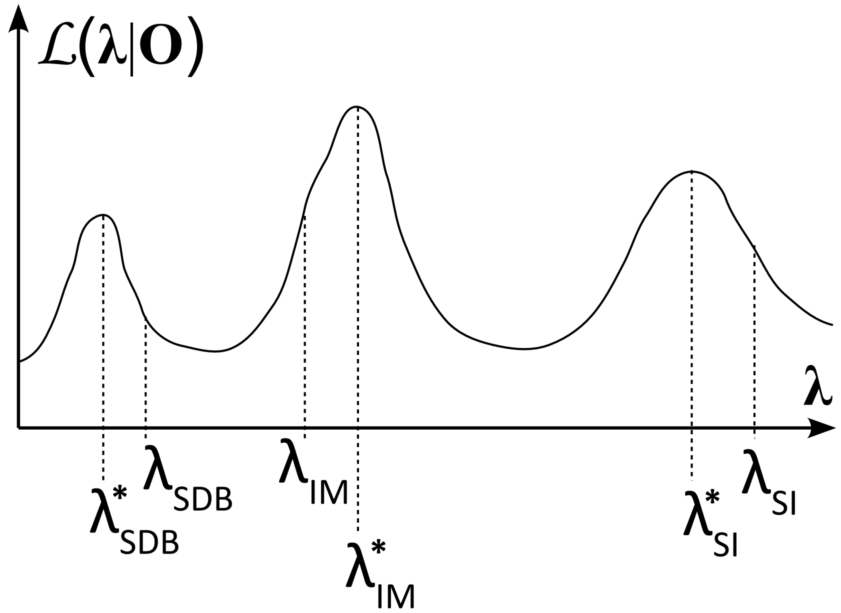

Figure 5.1: Searching for a better local maximum of the posterior probability of model

One can do something similar for the SDB from the target-population speaker, i.e. the speaker with dysarthria ($\mathbf{\Lambda}_{SDB}$), and reach a local maximum at $\mathbf{\Lambda}^*_{SDB}$. However, since the SDB does not learn any phone-discriminating patterns, the posterior at $\mathbf{\Lambda}^*_{SDB}$ is very likely to be much lower than that at $\mathbf{\Lambda}^*_{SI}$.

In general, because of the population mismatch, $\mathbf{\Lambda}_{SI}$ and $\mathbf{\Lambda}_{SDB}$ will be quite far away from each other in the AM parameter-space. However, *this large separation does not preclude the existence of an intermediate model* $\mathbf{\Lambda}_{IM}$ *which can reach a local maximum* $\mathbf{\Lambda}^*_{IM}$ *such that the posterior at* $\mathbf{\Lambda}^*_{IM}$

is higher than that at $\mathbf{\Lambda}^*_{SDB}$ *as well as* $\mathbf{\Lambda}^*_{SI}$. The 'in-between' model is formulated as a linear interpolation between $\mathbf{\Lambda}_{SDB}$ and $\mathbf{\Lambda}_{SI}$:

$$\mathbf{\Lambda}_{IM} = \mathbf{\Delta} \cdot \mathbf{\Lambda}_{SI} + (\mathbf{I} - \mathbf{\Delta}) \cdot \mathbf{\Lambda}_{SDB} \tag{5.1}$$

where $\mathbf{\Delta} = \operatorname{diag}(\delta_i)_i$ is a $P \times P$ diagonal matrix such that $0 \leq \delta_i \leq 1 \; \forall \; i$ (P being the dimensionality of the AM parameter-space); and $\mathbf{I}$ is the P-dimensional identity matrix. The locus of $\mathbf{\Lambda}_{IM}$ is the P-dimensional hypercube, two of whose vertices are $\mathbf{\Lambda}_{SI}$ and $\mathbf{\Lambda}_{SDB}$.

5.1.3 Intermediate model as prior for adaptation

In the second stage, adaptation is performed with $\mathbf{\Lambda}_{IM}$ as the prior or to-be-adapted model. One benefit of this two-stage approach is that once the mismatch has been accounted for, one should be able to employ any particular (classical) adaptation technique, be it MAP or MLLR or SMAP or EV.

5.2 Background Interpolation: An *A Priori* Empirical Study

To check the validity of the interpolation approach described above, a simulation was conducted for estimating the parameters of a mixture-of-Gaussians distribution [117]. Figure 5.2 displays the true distribution, two 2-component distributions that are obtained as local optima on the likelihood surface in the parameter space, and the globally optimal 2-component distribution. All learned distributions were estimated using the EM algorithm.

Figure 5.3 shows the likelihood contour plot for the same simulation, as a function of the means of the estimated mixture distribution. Points L_1^0 and L_2^0 are the initial mean values for EM estimation, and converge to local optima L_1^* and L_2^* respectively. The global optimum is indicated by G^*.

It can be seen that all points on the line-segment connecting L_1^0 and L_2^0 are interpolations between L_1^0 and L_2^0. If the interpolated point is very close to L_1^0, EM will converge to L_1^*; if it is very close to L_2^0, EM converges to L_2^*. However, if a suitable interpolation factor is chosen (an example of which is shown in the figure), EM can reach a better optimum (which in this case

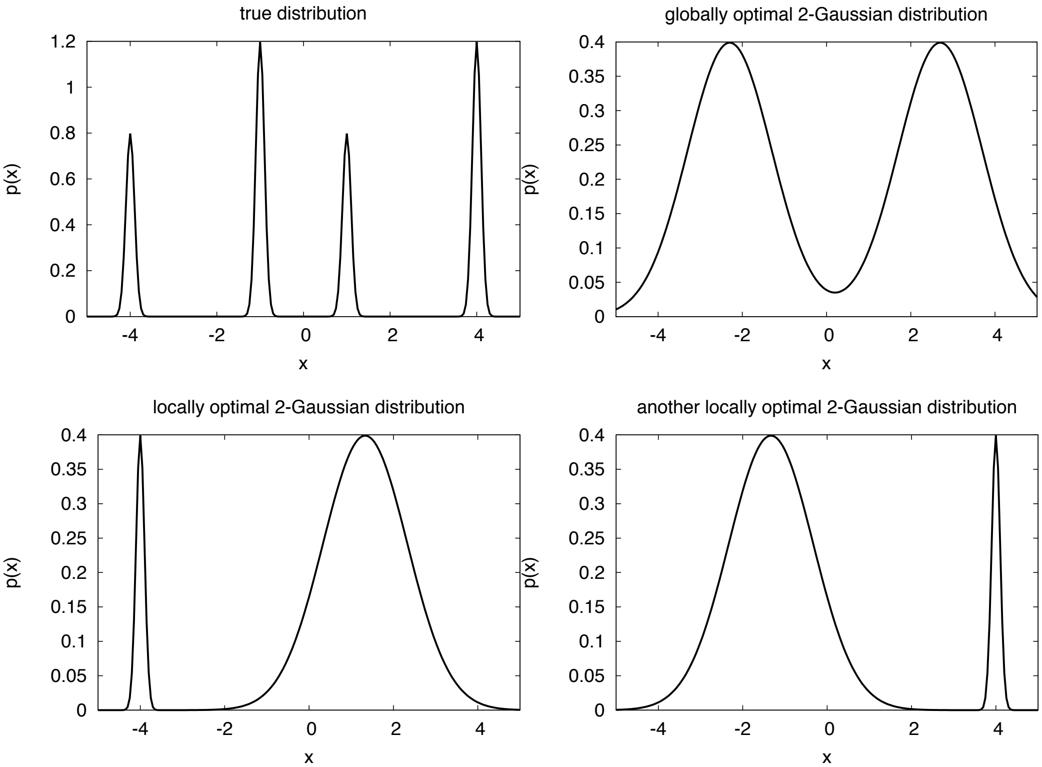

Figure 5.2: Gaussian mixture distributions from the simulation study

happened to be the global optimum) than both L_1^* and L_2^*.

5.3 Background Interpolated MAP Adaptation (BI-MAP)

This section presents a MAP adaptation scheme utilizing the interpolated speaker background model.

As mentioned earlier, the AM comprises HMMs whose observation distributions are mixtures of multivariate Gaussian densities with diagonal covariance matrices. So, for an N-state HMM with M p-dimensional Gaussians per state, the parameters are $\{\pi_i\}_i$, $\{a_{ij}\}_{i,j}$, $\left\{c_{il}, \vec{\mu}_{il}, \mathbf{\Sigma}_{il} = \mathrm{diag}\left(\sigma^2_{il_d}\right)\right\}_{i,l}$ — $i, j \in \{1, \ldots, N\}$; $l \in \{1, \ldots, M\}$; $d \in \{1, \ldots, p\}$.

Conventional/classical MAP adaptation utilizes Dirichlet distribution priors for $\{\pi_i\}_i$, $\{a_{ij}\}_{i,j}$, $\{c_{il}\}_{i,l}$ and a Gamma-Normal distribution prior for each $\left\{\mu_{il_d}, r_{il_d} = \sigma^{-2}_{il_d}\right\}$ pair. Ignoring constant terms, the overall prior for an HMM is (λ denoting the parameter set for a single HMM):

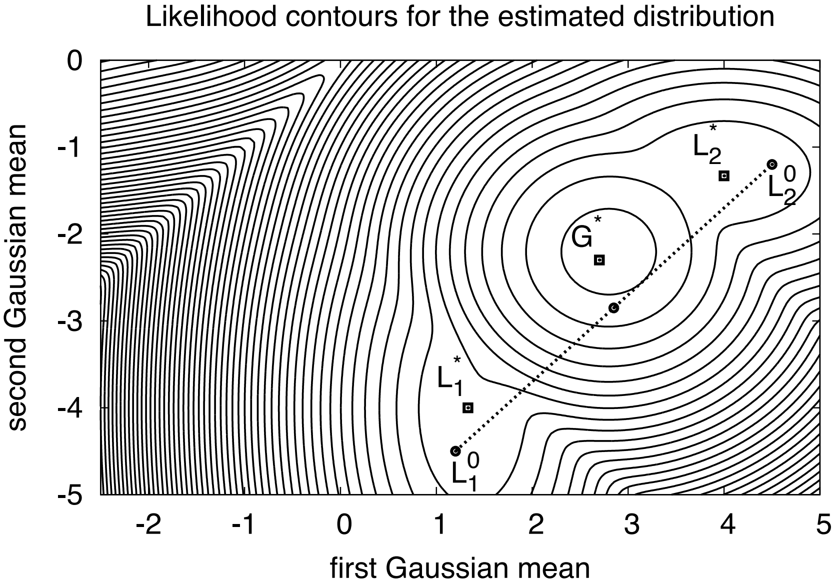

Figure 5.3: Likelihood contours, shown here as a function of the estimated Gaussians' means

$$\begin{aligned}\log G(\lambda) = &\sum_{i=1}^{N}(\eta_i - 1)\cdot\log(\pi_i) + \sum_{i=1}^{N}\sum_{j=1}^{N}(\eta_{ij}-1)\cdot\log(a_{ij}) \\ &+ \sum_{i=1}^{N}\sum_{l=1}^{M}(\nu_{il}-1)\cdot\log(c_{il}) \\ &+ \sum_{i=1}^{N}\sum_{l=1}^{M}\sum_{d=1}^{p}\Big[(\alpha_{il}-\frac{1}{2})\cdot\log(r_{il_d}) - \beta_{il_d}r_{il_d}\Big] \\ &- \sum_{i=1}^{N}\sum_{l=1}^{M}\sum_{d=1}^{p}\frac{\tau_{il}r_{il_d}}{2}(\mu_{il_d}-\rho_{il_{d_0}})^2\end{aligned} \tag{5.2}$$

where η_i, η_{ij}, ν_{il}, α_{il}, τ_{il}, β_{il_d} and $\rho_{il_{d_0}}$ are the prior's hyperparameters.

Combining $G(\lambda)$ with the maximum-likelihood (ML) auxiliary function [10], the MAP auxiliary function for iteration u+1 of EM is obtained (as a function of $\lambda^{(u)}$):

$$
\begin{aligned}
Q_{\mathrm{MAP}}^{(u+1)} &= \sum_{i=1}^{N} \Big(\eta_i - 1 + \sum_{k=1}^{K} \gamma_i^{uk}(1)\Big) \cdot \log(\pi_i) \\
&+ \sum_{i=1}^{N} \sum_{j=1}^{N} \Big(\eta_{ij} - 1 + \sum_{k=1}^{K} \sum_{t=1}^{T_k - 1} \xi_{ij}^{uk}(t)\Big) \cdot \log(a_{ij}) \\
&+ \sum_{i=1}^{N} \sum_{l=1}^{M} \Big(\nu_{il} - 1 + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)\Big) \cdot \log(c_{il}) \\
&+ \sum_{i=1}^{N} \sum_{l=1}^{M} \sum_{d=1}^{p} \Big[\Big(\alpha_{il} - \frac{1}{2} + \frac{1}{2} \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)\Big) \cdot \log(r_{il_d}) - \beta_{il_d} r_{il_d}\Big] \\
&- \sum_{i=1}^{N} \sum_{l=1}^{M} \sum_{d=1}^{p} \Big[\frac{\tau_{il} r_{il_d}}{2} (\mu_{il_d} - \rho_{il_{d_0}})^2 + \frac{r_{il_d}}{2} \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t) \cdot (o_{t_d}^{k} - \mu_{il_d})^2\Big]
\end{aligned}
\tag{5.3}
$$

where $\vec{o_t}^k$ is the observation vector at time t from the k^{th} observation sequence $\mathcal{O}_k = \{\vec{o_1}^k \ldots \vec{o_t}^k \ldots \vec{o_{T_k}}^k\}$; and $\gamma_i^{uk}(t)$, $\xi_{ij}^{uk}(t)$, $\gamma_{il}^{uk}(t)$ are respectively the posterior state occupancy, state transition, and mixture-component occupancy probabilities determined from iteration u (i.e., using $\lambda^{(u)}$). Maximizing $Q_{\mathrm{MAP}}^{(u+1)}$ gives us

$$
\begin{aligned}
\pi_i^{(u+1)} &= \frac{(\eta_i - 1) + \sum_{k=1}^{K} \gamma_i^{uk}(1)}{\sum_{j=1}^{N} (\eta_i - 1) + \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_j^{uk}(1)} \\
a_{ij}^{(u+1)} &= \frac{(\eta_{ij} - 1) + \sum_{k=1}^{K} \sum_{t=1}^{T_k - 1} \xi_{ij}^{uk}(t)}{\sum_{j=1}^{N} (\eta_{ij} - 1) + \sum_{k=1}^{K} \sum_{t=1}^{T_k - 1} \gamma_i^{uk}(t)} \\
c_{il}^{(u+1)} &= \frac{(\nu_{il} - 1) + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)}{\sum_{l=1}^{M} (\nu_{il} - 1) + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_i^{uk}(t)} \\
\mu_{il_d}^{(u+1)} &= \frac{\tau_{il} \rho_{il_{d_0}} + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t) o_{t_d}^{k}}{\tau_{il} + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)} \\
\sigma_{il_d}^{2\ (u+1)} &= \frac{2\beta_{il_d} + \tau_{il} (\mu_{il_d}^{(u+1)} - \rho_{il_{d_0}})^2}{(2\alpha_{il} - 1) + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)} \\
&+ \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t) \cdot (o_{t_d}^{k} - \mu_{il_d}^{(u+1)})^2}{(2\alpha_{il} - 1) + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)}
\end{aligned}
\tag{5.4}
$$

The prior's hyperparameters are chosen such that (1) if there is no adaptation data, the HMM parameter set should correspond to the initial parameter set; and (2) if there is an 'infinite' amount of adaptation data, the HMM parameter set should correspond/converge to its ML estimate [10]. Letting $\lambda_0 \equiv \left\{ \{\pi_{i0}\}_i \,,\ \{a_{ij0}\}_{i,j} \,,\ \left\{ c_{il_0}, \vec{\mu}_{il_0}, \mathbf{\Sigma}_{il_0} = \mathrm{diag}\left(\sigma^2_{il_{d_0}}\right) \right\}_{i,l} \right\}$:

$$\begin{gathered}
(\eta_i - 1) = \pi_{i_0} \left(\sum_{l=1}^{M} \tau_{il} \right) \quad ; \quad (\eta_{ij} - 1) = a_{ij_0} \left(\sum_{l=1}^{M} \tau_{il} \right) \\
(\nu_{il} - 1) = c_{il_0} \left(\sum_{l=1}^{M} \tau_{il} \right) \\
\beta_{il_d} = \frac{\tau_{il}\sigma^2_{il_{d_0}}}{2},\ \vec{\rho}_{il_0} = \vec{\mu}_{il_0},\ \alpha_{il} = \frac{\tau_{il}+1}{2}
\end{gathered} \tag{5.5}$$

Further, during implementation, τ_{il} are chosen to be a pre-specified constant τ. The MAP update expressions then become:

$$\begin{aligned}
\pi_i^{(u+1)} &= \frac{\pi_{i_0} M\tau + \sum_{k=1}^{K} \gamma_i^{uk}(1)}{M\tau + \sum_{j=1}^{N}\sum_{k=1}^{K} \gamma_j^{uk}(1)} \\
a_{ij}^{(u+1)} &= \frac{a_{ij_0} M\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k-1} \xi_{ij}^{uk}(t)}{M\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k-1} \gamma_i^{uk}(t)} \\
c_{il}^{(u+1)} &= \frac{c_{il_0} M\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)}{M\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_i^{uk}(t)} \\
\mu_{il_d}^{(u+1)} &= \frac{\tau\mu_{il_{d_0}} + \sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_{il}^{uk}(t) o_{t_d}^{k}}{\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)} \\
\sigma^{2\ (u+1)}_{il_d} &= \frac{\tau\sigma^2_{il_{d_0}} + \tau(\mu_{il_d}^{(u+1)} - \mu_{il_{d_0}})^2}{(2\alpha_{il} - 1) + \sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)} \\
&\quad + \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_{il}^{uk}(t) \cdot (o_{t_d}^{k} - \mu_{il_d}^{(u+1)})^2}{\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)}
\end{aligned} \tag{5.6}$$

The τ hyperparameter is a regularizer: it specifies the weight of the prior information relative to that of 'evidence' (the observations).

BI-MAP utilizes the same prior as conventional MAP, i.e. Equation 5.2 is still valid. The difference is that the starting/initial parameter set λ_0 is now an interpolation between the SI parameter set λ_0^{I}, and the SDB parameter set λ_0^{D}:

$$\begin{aligned}
\pi_{i_0} &= \delta_i \cdot \pi_{i_0}^{\mathrm{I}} + (1-\delta_i) \cdot \pi_{i_0}^{\mathrm{D}} \\
a_{ij_0} &= \delta_{ij} \cdot a_{ij_0}^{\mathrm{I}} + (1-\delta_{ij}) \cdot a_{ij_0}^{\mathrm{D}} \\
c_{il_0} &= \delta_{il}^{w} \cdot c_{il_0}^{\mathrm{I}} + (1-\delta_{il}^{w}) \cdot c_{il_0}^{\mathrm{D}} \\
\vec{\mu}_{il_0} &= \delta_{il}^{m} \cdot \vec{\mu}_{il_0}^{\mathrm{I}} + (1-\delta_{il}^{m}) \cdot \vec{\mu}_{il_0}^{\mathrm{D}} \\
\sigma_{il_{d_0}}^{2} &= \delta_{il}^{s} \cdot \sigma_{il_{d_0}}^{2^{\mathrm{I}}} + (1-\delta_{il}^{s}) \cdot \sigma_{il_{d_0}}^{2^{\mathrm{D}}}
\end{aligned} \tag{5.7}$$

where the δs are the interpolation factors (for the respective HMM parameter) in the range $[0,1]$. To give the same weight to the SDB prior relative to the SI prior for all parameters for all HMMs in the model, one can fix all the δs to be the same δ. This is the same as setting $\delta_i = \delta \; \forall \; i$ in Equation 5.1.

BI-MAP therefore has two types of regularizers: τ which determines the prior vs. evidence weighting; and δ which determines the SI vs. SDB weighting. The BI-MAP mean update, for instance, is:

$$\begin{aligned}
\vec{\mu}_{il}^{(u+1)} &= \frac{\tau_{\mathrm{I}}}{\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_{il}^{uk}(t)} \cdot \vec{\mu}_{il_0}^{\mathrm{I}} \\
&+ \frac{\tau_{\mathrm{D}}}{\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_{il}^{uk}(t)} \cdot \vec{\mu}_{il_0}^{\mathrm{D}} \\
&+ \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_{il}^{uk}(t)\vec{o}_t^{\,k}}{\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k}\gamma_{il}^{uk}(t)}
\end{aligned} \tag{5.8}$$

where $\tau_{\mathrm{I}} = \delta\tau$ and $\tau_{\mathrm{D}} = \tau - \tau_{\mathrm{I}} = (1-\delta)\tau$. BI-MAP updates for other parameters can be similarly obtained by using Equation 5.7 in Equation 5.6.

Chapter 6

EXPERIMENTS

In this chapter, we report on adaptation experiments in which the *prior* acoustic model is obtained as an interpolation of two models, to improve speech recognition for speakers with dysarthria. Given the way background interpolation was motivated, these are the first adaptation experiments of this kind.

The main idea of these experiments is to obtain an acoustic model that better captures the acoustic patterns in dysarthric speech. A search for such more optimal acoustic models is performed starting at a point in the vicinity of the usual SI model, in the parameter space. The starting point for this search is chosen while accounting for speaker-specific speech characteristics, and should in principle allow us to obtain a better local maximum of the likelihood function.

Section 6.1 discusses some issues related to the implementation of BI-MAP adaptation. Section 6.2 describes the experimental setup and the strategy for evaluating recognition performance. Sections 6.3 and 6.4 present the adaptation experiments performed and the results obtained.

6.1 Implementing BI-MAP Adaptation

All experiments in this study built acoustic models utilizing a 3-state HMM for each context-dependent triphone. Data-driven state tying (using decision trees) was performed to accommodate data sparsity, as discussed earlier in Chapter 2. All steps in the experiments (with the exception of model interpolation) were performed using the HTK toolkit [12].

6.1.1 AM-space dimensionality and Gaussian alignment

When re-estimating HMM parameters, HTK usually applies pre-specified floors to variance and mixture weight parameters of Gaussian distributions: variance estimates below a certain minimum variance are clamped to that floor, and Gaussian components whose mixture weight falls below a certain minimum weight are discarded from the corresponding mixture. Therefore, the resulting acoustic model may not have the same number of unique Gaussians as the one it was re-estimated from. This is a cause of concern if the AMs to be interpolated, $\mathbf{\Lambda}_{SI}$ and $\mathbf{\Lambda}_{SDB}$, do not have the same dimensionality in the AM parameter space, after they have undergone their respective re-estimation stages.

Secondly, even if one is fortunate enough to have the two AMs be of the same dimensionality, there exists the issue of Gaussian alignment: before the two AMs can be interpolated, it is required to know which Gaussian component in one of the unique HMM states of $\mathbf{\Lambda}_{SI}$ corresponds to which Gaussian component in the corresponding HMM state of $\mathbf{\Lambda}_{SDB}$. The number of all possible alignments is $(M!)^{N_s}$, where M is the number of Gaussian components per mixture (i.e., per HMM state) and N_s is the number of unique HMM states in the AM. Considering all these alignments is not practical. Moreover, doing so would require a metric to compare the 'fitness' of the candidate alignments: for mixture of Gaussians, the Kullback-Leibler divergence is not analytically tractable. However, one can use the K-L metric for aligning the Gaussian components within each pair of corresponding mixtures (HMM states): doing so reduces the complexity of determining the alignment but only to $O(M^2 \cdot N_s)$.

An approximate solution to the two issues discussed above was used: the two AMs were generated from the same prototype HMM. When learning an AM, the state-specific distributions are usually up-mixed (to more Gaussian components) *after* performing the data-driven state tying. Here, a single HMM with the final number of Gaussians was first obtained (using the mismatched population's training speech). By setting the mixture weight floor to machine-epsilon, it was ensured that no Gaussian components would be discarded in the successive re-estimation stages. To obtain $\mathbf{\Lambda}_{SI}$, it was then cloned to each unique HMM state (the unique HMM states were determined using a state-tying tree learned in a separate estimation of mismatched AM, performed in the usual manner) and this cloned AM was re-estimated completely. To obtain $\mathbf{\Lambda}_{SDB}$, this seed/prototype HMM was re-estimated completely using all speech from the training set of the speaker with dysarthria. This re-estimated HMM definition was cloned to each unique HMM state (also specified by the tree mentioned above).

6.1.2 Interpolation parameters – independent or dependent?

In principle, one can have as many interpolation factors (the δs) as there are AM parameters. Doing so would permit investigation of all possible $\mathbf{\Lambda}_{IM}$s. However, controlling/specifying such a large number of interpolation factors individually is not practical. The first BI-MAP adaptation experiments investigated the "same δ for all AM parameters" scenario. Fixing δ to be the same for all parameters solves the issue outlined above, but only explores a limited portion of the AM-parameter space (the line-segment joining $\mathbf{\Lambda}_{SI}$ and $\mathbf{\Lambda}_{SDB}$).

In order to investigate more $\mathbf{\Lambda}_{IM}$s, an in-between approach was taken for the second set of BI-MAP experiments: Gaussian means and variances were always interpolated; the mixture weights were either interpolated or came from $\mathbf{\Lambda}_{SI}$ ($\delta = 1$); and the transition probabilities were either interpolated, or came from $\mathbf{\Lambda}_{SI}$ ($\delta = 1$) or $\mathbf{\Lambda}_{SDB}$ ($\delta = 0$). The interpolation δ was fixed to be the same for the parameters that were interpolated. Although this parameter-type dependent interpolation does not cover all possible $\mathbf{\Lambda}_{IM}$s, BI-MAP was still able to outperform the conventional MAP technique for the UA-Speech corpus.

6.2 Evaluation

Architecture, speech features, and corpus for mismatched-population: These were identical to those for the preliminary experiments described in Chapter 4.
Baseline: The performance of BI-MAP adaptation was compared to that of the standard MAP adaptation (referred to as SI-MAP from hereon). SI-MAP adaptation was performed again for two reasons: (1) the preliminary experiments described in Chapter 4 were performed only for a subset of

the UA-Speech speakers; and (2) while implementing BI-MAP, a bug was discovered in HTK's code for MAP re-estimation of Gaussian variances.

Recognition performance evaluation: Word recognition accuracy, as defined in Equation 4.1, was used for these experiments as well.

Significance Testing: Statistical significance of the difference in ASR recognition accuracies between two ASR systems was compared at two levels. The Gillick-Cox matched-pairs test [118] was first used for each speaker with dysarthria to determine if the difference in the recognition accuracies using BI-MAP versus SI-MAP was statistically significant. Then, the Wilcoxon signed-rank test [119] is performed for the pairs of WRAs (for the speakers for which the Gillick-Cox test had rejected the null hypothesis) to determine if BI-MAP is overall significantly different from SI-MAP with respect to WRA. This second statistical test is needed in the event that SI-MAP has higher WRA than BI-MAP for one or more speakers, such that the difference in WRAs is significant (from the Gillick-Cox test).

6.3 Parameter-Type Independent Background Interpolation

This section describes the first set of BI-MAP experiments, where the interpolation factor δ was set to be the same for all parameters. Three configurations were studied: conventional MAP with TIMIT-trained $\mathbf{\Lambda}_{SI}$ as the prior model (i.e., SI-MAP); $\mathbf{\Lambda}_{SDB}$ of a particular UA-Speech speaker as the prior model for MAP adaptation (called SDB-MAP from hereon); and a linear interpolation $\mathbf{\Lambda}_{IM}$ of these two prior models (as per Equation 5.1) as the overall prior model for MAP (i.e., BI-MAP). The value of δ for BI-MAP was varied from 0 to 1 in steps of size 0.1; $\delta = 0$ corresponds to SDB-MAP and $\delta = 1$ corresponds to SI-MAP. All parameters were adapted in the second stage with the MAP hyperparameter τ set to 5.0 for all configurations. These systems were developed for each of the sixteen UA-Speech speakers. Since $\mathbf{\Lambda}_{SDB}$ does not learn any phone-discriminating information from the training data, one would expect it to have the poorest performance among the three configurations.

Table 6.1: Speaker intelligibility and WRAs for parameter-type independent BI-MAP experiments

Speaker	**Average Intell. (%)**	**SDB-MAP WRA (%)**	**SI-MAP WRA (%)**	**BI-MAP WRA (%)**
M04	2	0.6	2.98	**3.2**
F03	6	7.1	**21.4**	19.8
M12	7	4.6	14.77	**16.4**
M01	17	4.4	12.65	**14.1**
M07	28	17.7	38.99	**42.5**
F02	29	17.5	29.02	**31.1**
M06	39	13.5	36.75	**39.3**
M16	43	5.2	26.47	**32.1**
M05	58	15.4	**38.09**	36.8
M11	62	10.2	**29.8**	28.9
F04	62	10.6	32.88	**34.8**
M09	86	25.5	63.92	**70**
M14	90	29.1	60.73	**64.1**
M10	93	52.3	73.11	**74.2**
M08	95	21.2	**69.58**	66.9
F05	95	57.9	78.71	**80.7**

6.3.1 Results

Columns 3, 4 and 5 in Table 6.1 list the WRAs for each UA-Speech speaker, for the three system configurations, in increasing order of the speakers' average intelligibility. For BI-MAP, the score is listed for δ that gave the best WRA. Speakers with higher BI-MAP WRA than SI-MAP WRA have their BI-MAP WRAs listed in boldface. Speakers for whom the Gillick-Cox test rejected the null hypothesis have their BI-MAP WRAs highlighted in green color. We chose $\alpha = 0.1$ rather than $\alpha = 0.05$ because of the small amount of test set data.

As expected, SDB-MAP is drastically outperformed by both the baseline (SI-MAP) as well as BI-MAP. Further, for most of the speakers, BI-MAP was able to find a δ that gave a higher WRA than the corresponding SI-MAP system (all except F03, M05, M11, M08). It is possible that because we only searched at discrete values of δ, there may be some values of it in the unexplored intervals where one might get BI-MAP to have a better score than SI-MAP.

Of the 12 speakers for which BI-MAP had a higher WRA than SI-MAP, there are 10 speakers for which the difference in WRAs is significant (all except M04 and M01). For the remaining speakers, there was no significant difference. At $\alpha = 0.05$, 7 of these 10 speakers still had a significantly higher BI-MAP WRA than SI-MAP WRA.

Considering the speakers for which the difference in these WRAs was significant, the Wilcoxon signed-rank test rejected the null hypothesis "SI-MAP and BI-MAP are different only by chance" with 95% confidence.

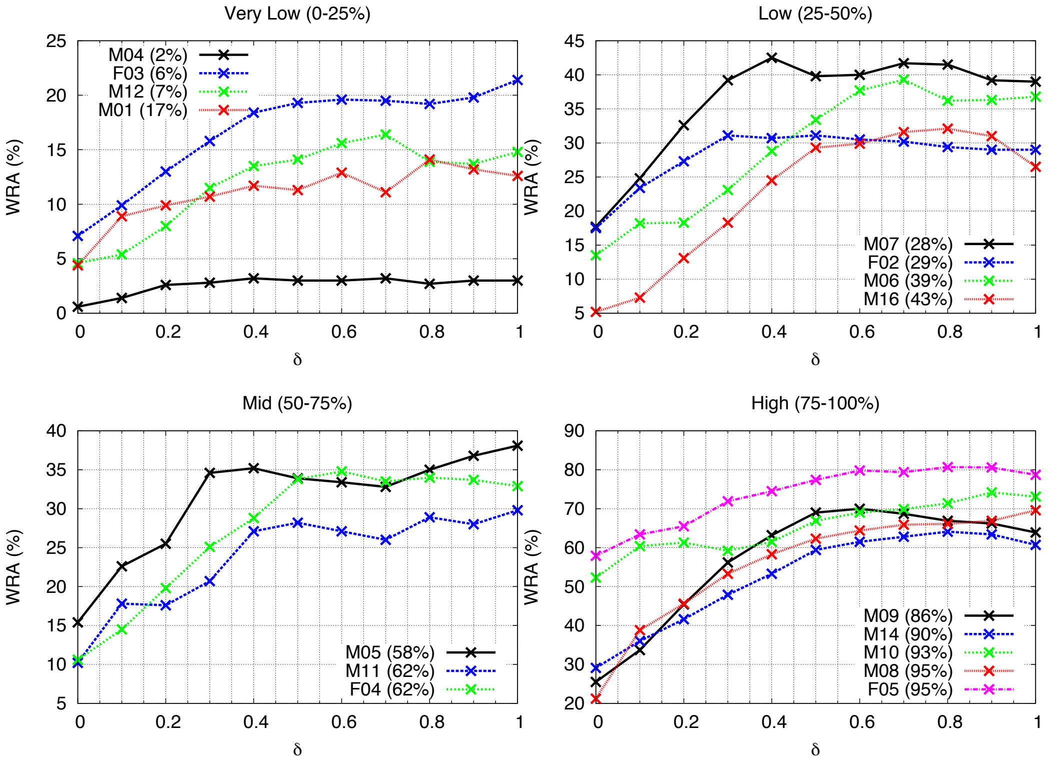

Figure 6.1: Validation experiments' recognition accuracies for various intelligibility categories

Figure 6.1 plots the WRA as a function of δ for the four intelligibility categories, obtained at steps of size 0.1 between 0 and 1. For all speakers except M04, we see gradual improvement as one moves away from $\mathbf{\Lambda}_{SDB}$ ($\delta = 0$) and towards $\mathbf{\Lambda}_{SI}$ ($\delta = 1$). For most of these speakers (twelve of sixteen), we see $WRA(\delta)$ peaking at an intermediate value of δ. Further, small values of δ generally perform more poorly than SI-MAP in terms of WRA. For speakers with higher BI-MAP WRA (compared to SI-MAP), the optimal δ occurs between 0.5 and 1. It can be seen that even though we are searching at discrete points (and that too only along the line-segment whose end-points are the SI and SDB prior AMs), the hypothesis that there can exist better local maxima for intermediate prior models stands validated.

6.4 Parameter-Type Dependent Background Interpolation

This section describes the second set of BI-MAP experiments, where the interpolation δs were set as described in Section 6.1.2. For these experiments, six BI-MAP configurations were studied. The system naming convention involves two digits preceded by the letter 'C'. The first digit indicates the source of prior mixture weights and the one following it indicates the source of prior transition probabilities (prior Gaussian means and variances were interpolated for all six systems). A '0' indicates that the associated parameter was interpolated, i.e. it came from $\mathbf{\Lambda}_{IM}$; a '1' indicates that it came from $\mathbf{\Lambda}_{SI}$; and a '2' indicates that it came from $\mathbf{\Lambda}_{SDB}$. These are listed in Table 6.2. Systems C00, C01 and C02 will be collectively referred to as the C0 subgroup (and systems C10, C11 and C12 as the C1 subgroup), when necessary. The value of δ for BI-MAP was varied from 0 to 1 in steps of size 0.05; all parameters were adapted in the second stage with the MAP hyperparameter τ set to 5.0 for all configurations.

Table 6.2: BI-MAP system configurations studied. Gaussian means and variances were always interpolated.

BI-MAP	Prior	Prior
configuration	Mixt. Weights	Trans. Probs.
C00	SI+SDB	SI+SDB
C01		SI
C02		SDB
C10	SI	SI+SDB
C11		SI
C12		SDB

6.4.1 Results

Columns 3 and 4 in Table 6.3 list the WRAs for each UA-Speech speaker, for SI-MAP and BI-MAP adaptation, in increasing order of the speakers' average intelligibility. For BI-MAP, the score is listed for δ and the system configuration that gave the best WRA. Speakers with higher BI-MAP WRA than SI-MAP WRA have their BI-MAP WRAs listed in boldface. Columns 5 and 6 indicate whether the Gillick-Cox test rejected the null hypothesis (at 95% and 90% confidence levels respectively): colored cells indicate that the difference in WRAs was significant, and white cells indicate otherwise.

Table 6.3: Speaker intelligibility and WRAs for parameter-type dependent BI-MAP experiments

Speaker	**Average Intell. (%)**	**SI-MAP WRA (%)**	**BI-MAP WRA (%)**	$\alpha = 0.05$	$\alpha = 0.10$
M04	2	2.98	**4.16**		🟧
F03	6	21.4	**22.35**		
M12	7	14.77	**16.41**		🟧
M01	17	12.65	**15.39**	🟩	🟧
M07	28	38.99	**42.8**	🟩	🟧
F02	29	29.02	**33.39**	🟩	🟧
M06	39	36.75	**40.67**	🟩	🟧
M16	43	26.47	**32.88**	🟩	🟧
M05	58	38.09	**38.88**		
M11	62	29.8	**30.91**		
F04	62	32.88	**35.74**	🟩	🟧
M09	86	63.92	**71.65**	🟩	🟧
M14	90	60.73	**64.2**	🟩	🟧
M10	93	73.11	**75.01**	🟩	🟧
M08	95	**69.58**	67.79	🟩	🟧
F05	95	78.71	**82.07**	🟩	🟧

For all speakers except M08, BI-MAP had a higher WRA than the corresponding SI-MAP system.

Of the 15 speakers for which BI-MAP had a higher WRA than SI-MAP, there are 12 speakers for which the difference in WRAs is significant at $\alpha = 0.10$. For the remaining speakers, there was no significant difference. At $\alpha = 0.05$, 10 of these 12 speakers still had a significantly higher BI-MAP WRA than SI-MAP WRA. Secondly, by exploring more starting points (than just the ones lying on the line-segment connecting SI and SDB prior AMs), BI-MAP has obtained higher WRA scores for more speakers, compared to the parameter-type independent scenario. Again, although not all possible values of δ have been tested yet, these numbers show that background-interpolated prior models can help to improve recognition accuracy.

Considering the speakers for which the difference in these WRAs was significant, the Wilcoxon signed-rank test rejected the null hypothesis "SI-MAP and BI-MAP are different only by chance" with 99% confidence, for both α levels of the Gillick-Cox test.

Figure 6.2 plots the WRA of the six BI-MAP configurations, as a function of δ for speakers in the *Very Low* and *Low* intelligibility categories (and Figure 6.3 for those in the *Mid* and *High* categories), obtained at steps of size 0.05 between 0 and 1. For all speakers except M04, we see gradual improvement as one moves away from $\mathbf{\Lambda}_{SDB}$ ($\delta = 0$) and towards $\mathbf{\Lambda}_{SI}$ ($\delta = 1$). For M04, the WRA curves fluctuate more and the improvement with increase in δ is less pronounced.

Secondly, for speakers with higher BI-MAP WRA than SI-MAP (all except M08), $WRA\,(\delta)$ peaks at an intermediate value of δ, with the optimal δ occurring between 0.4 and 1. Smaller values of δ generally perform more poorly than SI-MAP in terms of WRA. This is not counter-intuitive or unexpected: for small values of δ, $\mathbf{\Lambda}_{IM}$ is not very different from $\mathbf{\Lambda}_{SDB}$, and we have seen earlier that due to lack of phone-discriminating information SDB-MAP had rather poor recognition accuracies.

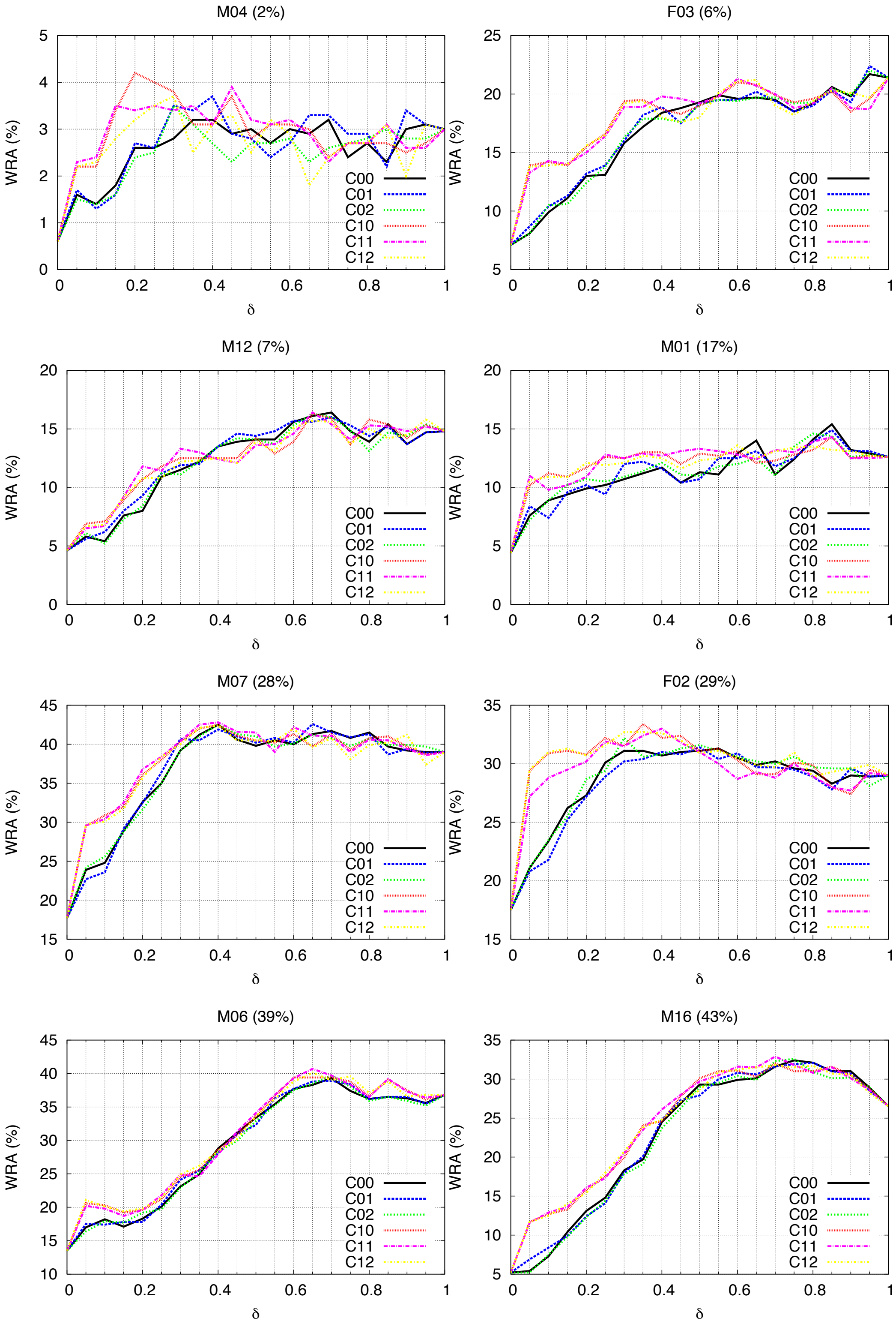

Figure 6.2: Recognition accuracies for Very Low (0–25%) and Low (25–50%) categories

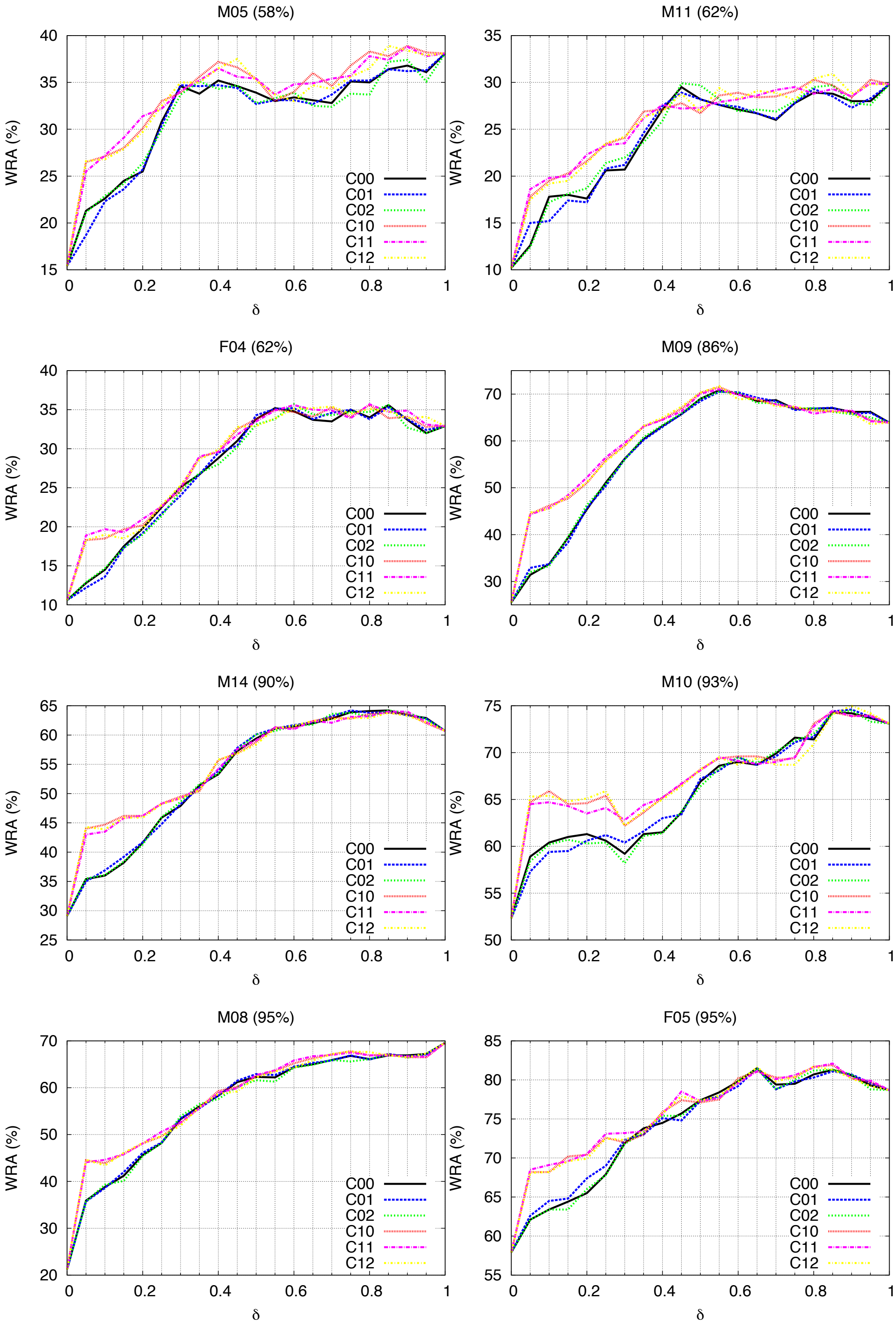

Figure 6.3: Recognition accuracies for Mid (50–75%) and High (75–100%) categories

Chapter 7

DISCUSSION

This chapter discusses the results of the study's experiments in more detail.

7.1 Evaluating What the Acoustic Model Learned

From the previous chapter, it is clear that BI-MAP was able to obtain higher recognition accuracy compared to the conventional SI-MAP technique. An important question that automatically arises is why this happened, especially from the point-of-view of acoustic modeling. It would therefore be interesting to see if the acoustic models (HMMs for the sub-word units) learn significantly different spectral representations from different to-be-adapted prior models.

After tying and data-driven clustering of HMM states has been performed, the overall acoustic model will typically contain a smaller number of unique states or *senones*, than before (in addition to the transition probability matrices). The senone definitions describe the state-specific probability distributions that govern the observations' generation. The acoustic models generated in this study's experiments ended up having roughly 3000 senones (3041, to be precise). In HTK, these senone definitions are identified by a "∼`s`" symbol and a unique name, which will be referred to as the senone's label from hereon. Then, to answer the above mentioned question, we want to efficiently select those senones whose spectra are significantly different in SI-MAP and BI-MAP adapted acoustic models, for the case where BI-MAP had better WRA.

The following procedure was used for selecting *significantly different* HMM-state spectra:

1. In the reference transcription and the two hypothesis transcriptions (one each for SI-MAP and BI-MAP adaptation), the sequence-of-triphone

labels was mapped to sequence-of-senones labels for each test-set utterance. For BI-MAP the hypothesis transcription came from the configuration (C00, etc.) with the best WRA.

2. String alignment was performed for senone-label sequences, for (a) reference and SI-MAP hypothesis; and (b) reference and BI-MAP hypothesis. This step was performed using the `sclite` toolkit [120].

3. Gillick-Cox matched pairs test was performed for each senone label, to determine if it had been identified in significantly different locations in the test-set transcriptions (when comparing reference transcription to the SI-MAP and BI-MAP hypothesis transcriptions). This was done for senone labels that showed up at least 20 times in the transcriptions and for a 95% confidence interval. At this stage, 228 senones were found to be present in significantly different positions in the transcriptions.

4. For each of these 228 senones, the 32 Gaussian mean vectors were weighted with their respective mixture weights and added to obtain a weighted PLP mean vector for that senone. This PLP mean vector was inverted to obtain the spectral representation for that senone. This was done for 4 versions of the senone – one each from $\mathbf{\Lambda}_{SI}$, $\mathbf{\Lambda}_{IM}$ and the final adapted versions of these two (i.e., the final SI-MAP and BI-MAP acoustic models).

5. From these 228 spectral representations, the ones for which SI-MAP and BI-MAP spectra were visually more-or-less same were discarded. This resulted in a final set of 129 senones.

Enumerating the differences for all 129 senone spectra here would be difficult, so this discussion is restricted to a few interesting ones. First, some notation: in the figures that will be discussed shortly, each plot will compare spectra for a particular senone, for a particular speaker. This information is indicated in the plot's title as follows: for the twentieth version of the senone representing the middle emitting state of the phoneme '`t`' for speaker F02, the plot's title would be "`F02 :: t_s320`". The first digit after 's' indicates the HMM-state (in HTK, dummy 'entry' and 'exit' states are attached to each HMM; hence the middle three states are the actual emitting states),

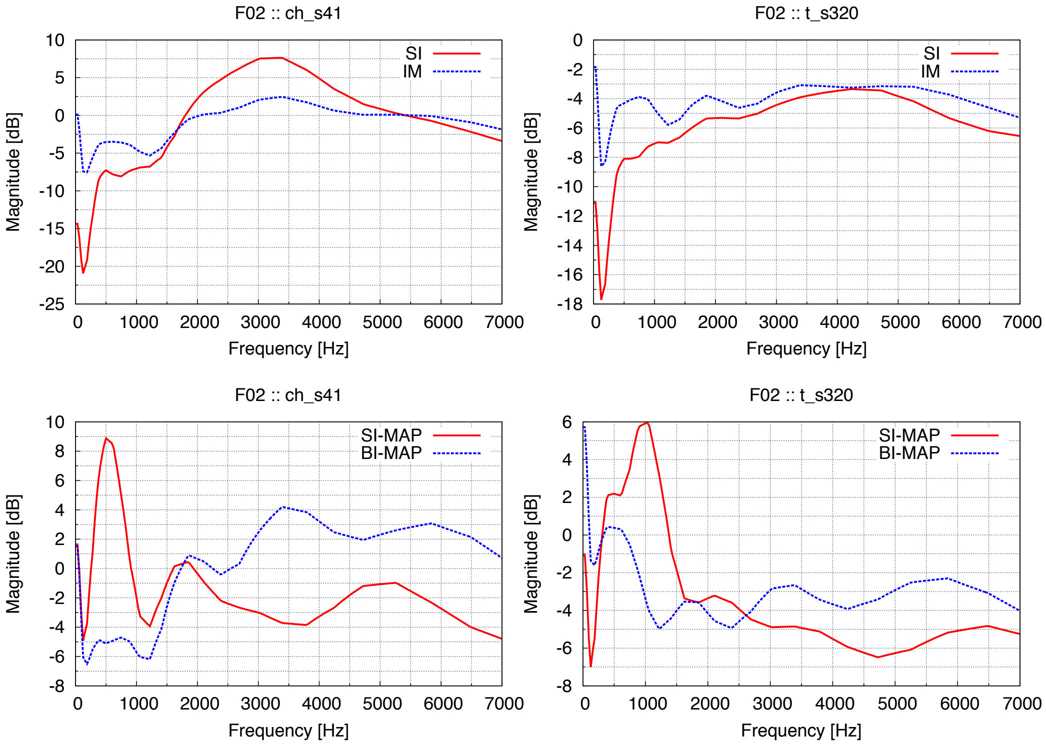

Figure 7.1: Senone spectra before and after adaptation, for standard MAP and BI-MAP

and the remaining digits indicate the particular version (of that state) that resulted from data-driven clustering.

Figure 7.1 shows the spectra for two senones of speaker F02: the first version of the last emitting state for the affricate ‘ch’, and the twentieth version of the middle emitting state for the unvoiced stop ‘t’. The plots in the figure’s top half compare the spectra from $\mathbf{\Lambda}_{SI}$ and $\mathbf{\Lambda}_{IM}$, the to-be-adapted prior acoustic models; and the bottom half’s plots do so for the final adapted acoustic models. The BI-MAP configuration was C10 with $\delta = 0.35$. For ‘ch_s41’, we see that the SI-MAP spectrum is indicative of the speaker already moving into the first emitting state of the vowel following this affricate. The BI-MAP spectrum indicates more of a pause between the ‘ch’ release and the onset of voicing. The de-emphasis of the spectral peaks in the 0-1500 Hz range is accounted for by the low energy of the turbulence source at low frequencies. For ‘t_s320’, the SI-MAP spectrum picks up the incomplete closure; the BI-MAP spectrum exhibits spectral peak emphasis (0-100 Hz) and de-emphasis (in the neighborhood of the first formant). The peak emphasis could be attributed to incomplete or no closure of vocal folds (for peak near the first harmonic), or to aperiodicity in voicing (for peak below the first harmonic; there could be bi-periodicity with a larger second period, causing an energy peak at about half the pitch frequency). Peak de-emphasis about the first formant is likely to have been caused by increased variability, or by increased damping (which in turn, is possible because of increased nasality or breathiness). Background interpolation with F02's SDB model has helped here because the SI-MAP prior model was trained using speech from TIMIT corpus, which is not only population-mismatched with UA-Speech but also style mismatched: TIMIT's speech is read sentences, and UA-Speech consists of isolated word utterances.

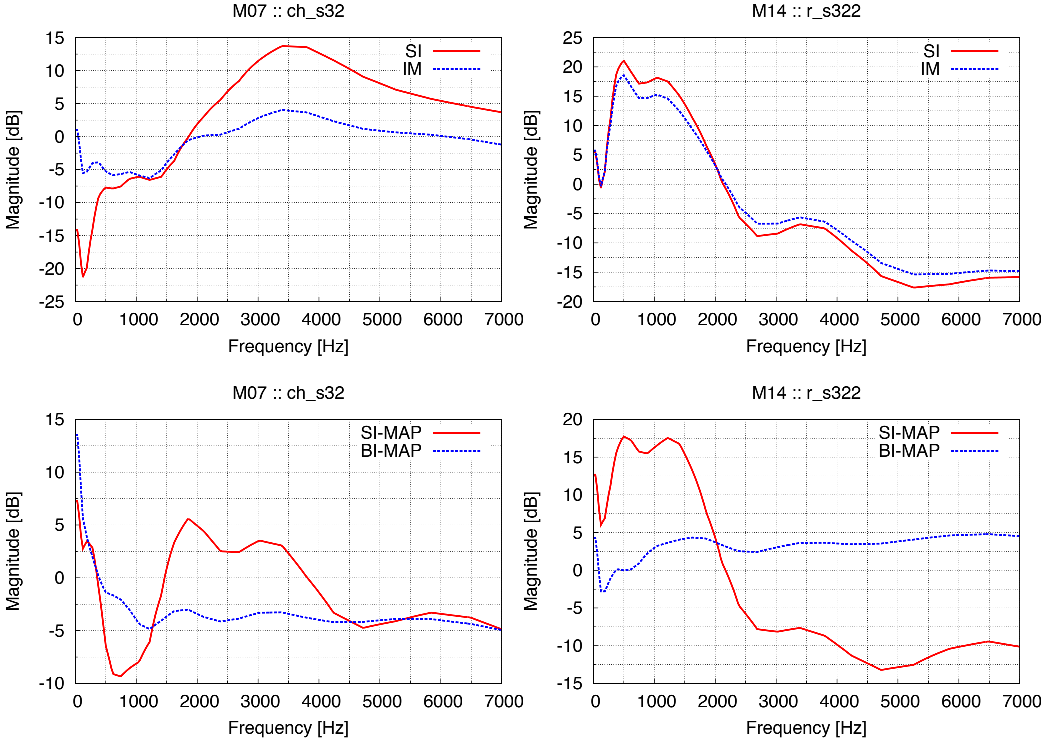

Figure 7.2: Senone spectra before and after adaptation, for standard MAP and BI-MAP

Figure 7.2 shows the spectra of a senone each from speakers M07 and M14: second version of the middle emitting state for the affricate '`ch`' for M07, and the twenty-second version of the middle emitting state for the liquid '`r`' for M14. The BI-MAP configuration was C11 with $\delta = 0.4$ for M07 and C02 with $\delta = 0.85$ for M14. In M07's case, the BI-MAP spectrum looks much more like that of silence and the SI-MAP version is more indicative of frication. Similarly for M14, there seem to have been enough observation-tokens that the BI-MAP model is representing as having come from the middle emitting state of the liquid, but which are more silence-like. On the other hand (depending on the triphone context), it could be the case that the tongue-tip is not going where it should, and not going reliably where it goes. What is interesting here is that the spectra prior to adaptation are almost identical (which is expected for the high value of δ), but they learned very different frequency-energy distributions during adaptation. This can be definitely attributed to the prior transition probabilities coming from M14's SDB model (configuration C02). So, background interpolation has helped again to better counter the mismatch.

7.2 Structurality of Model Parameters

The experiments described in the previous chapters also speak to the structurality of HMM parameters, particularly the mixture weights and transition probabilities. More specifically, the results indicate that the mixture weights are more structural parameters than transition probabilities, and that this is the case across all intelligibility categories. This is illustrated by the following observations regarding recognition accuracy:

1. From the results of the preliminary experiments (described in Chapter 4), it is clear that modifying and/or adapting the transition probabilities lowered the recognition accuracy, compared to the configuration where they were not changed. This happened with both speaker-dependent (C00 vs. C01) as well as speaker-adapted (C13 vs. C14 vs C15) systems. Among the speaker-adapted systems, the best accuracies were obtained by the configuration in which the mixture weights were adapted, but not the transition probabilities (apart from adapting the means and variances).

2. Looking at the WRA curves for BI-MAP adapted systems again (Figures 6.2 and 6.3), we see that the C1 subgroup of configurations had higher recognition accuracies compared to the C0 subgroup of configurations, for lower values of δ. In that range, the interpolated prior

model is not too far from the SDB prior model in the AM-parameter space and having the mixture weights come from the TIMIT prior model helps because they must incorporate some amount of phone-discriminating information.

3. The WRA curves for the various BI-MAP systems also exhibit tight coupling for most of the speakers: the curves for C0 subgroup are tightly coupled, and so are the ones within the C1 subgroup. Recognition accuracies within either subgroup do not appear to be impacted much by the source of prior transition probabilities (C_0 vs. C_1 vs. C_2). This can be expected for higher values of δ (but only for C_0 vs. C_1, because in these configurations the prior transition probabilities are not very different from each other when δ is high; coupling of C_2 WRA curve is still unexplained). However, this effect is observed for lower δ values as well.

Chapter 8

CONCLUSION

This chapter reviews the key findings and summarizes the experiments performed.

This study explored population-mismatch modeling for adaptation of acoustic models, particularly for recognition of dysarthric speech. From a peripheral view, population mismatch is an important problem because it is one of the major causes of poor ASR performance. Therefore, having an acoustic modeling technique that accounts for such mismatch is an important goal. This goal becomes even more important when speaker-dependent systems are hard to obtain due to scarcity of speech resources.

The experiments underlying this study investigated population mismatch modeling in a particular context – with a particular adaptation algorithm (MAP adaptation), and on the task of isolated word recognition. Recognition of dysarthric speech is a difficult task. It is made more difficult by the lack of sufficient speech data to model at a fine level of granularity the inconsistencies in acoustic features of this population. We have also seen in Chapter 3 that there exists much debate in the clinical research community on an adequate theory of motor speech disorders. Therefore, the technique developed in this study to obtain the prior acoustic model is designed with only a subtle connection to the hypothesis of Weismer et al. [87] about such a theory; yet, at this subtle level of connection, it stands in support of their hypothesis.

The central objective of this research was to find a procedure for obtaining a better starting point (i.e., a better prior acoustic model) for the adaptation algorithm than is used conventionally, with respect to recognition accuracy. For a parameter count almost identical to the baseline approach (BI-MAP has used only one additional hyperparameter: the interpolation factor δ), the experiments show that background interpolation has been able to achieve either at-par or better recognition performance (in terms of statistical significance), at least for the speech corpus used.

The fact that searching for alternative starting points (for optimization algorithms working on objective functions punctuated with local optima) can lead to a better local optimum is not new. Undoubtedly, the procedure presented in this work for doing so is not the be-all and end-all of research in adaptation for dysarthric speech. However, it has been hitherto unexplored and does appear promising: it provides a principled way of searching for prior acoustic models that account for population-mismatch. The positive results of parameter-type dependent BI-MAP adaptation suggest that making the interpolation factors specific to model parameters is helpful. Finding principled ways of doing so is an obvious direction for future work.

8.1 Directions for Future Work

The experiments described in this dissertation point towards WRA improvements through an interpolation-based technique for obtaining the to-be-adapted acoustic model. This section covers briefly some of the possible extensions for the proposed approach.

One of the possible directions to explore is that of *localized* SDB models. In this work, a single SDB model was obtained for a particular speaker. The SDB did not learn any patterns that can discriminate between phones/words. It was a model of the general characteristics of the speaker from the target population. The intention behind using such a model is to capture aspects of time-frequency variation that depend on the speaker (rather than on what was spoken by him/her). Given sufficient training tokens for each sub-word unit (for a particular speaker), one can obtain an SDB for each broad class of sub-word units. For example, if the sub-word units are phones, then one could in principle obtain a vowel SDB, a fricative SDB, a plosive SDB, etc.; doing so would be useful if the speaker's production of a particular broad class of sub-word units is characterized by some speech production deficit.

Testing background interpolation with other standard adaptation techniques (such as the MLLR algorithm) is an obvious next step. However, something more interesting would be to find a principled way of setting the interpolation factors (the δs) for each parameter of the acoustic model. This can be achieved for instance by jointly optimizing the δs and the HMM parameters, with respect to the adaptation data. For example, we have seen in Section 2.2.1 that MAP adaptation entails setting the HMM parameter set $\mathbf{\Lambda}$ to the mode of the posterior distribution $p\left(\mathcal{O}|\mathbf{\Lambda}\right)\cdot p_0\left(\mathbf{\Lambda}\right)$ where $p_0\left(\mathbf{\Lambda}\right)$ is the prior distribution of $\mathbf{\Lambda}$. If we denote the set of all interpolation factors by $\{\delta\}$, then one can set $\mathbf{\Lambda}$ to the mode of the posterior distribution $p\left(\mathcal{O}|\mathbf{\Lambda},\{\delta\}\right)\cdot p_0\left(\mathbf{\Lambda},\{\delta\}\right)$ where $p_0\left(\mathbf{\Lambda},\{\delta\}\right)$ is the prior distribution of $\mathbf{\Lambda}$. In other words, the interpolation factors are also chosen treating the adaptation data as 'evidence'.

The setup described in the last paragraph can be generalized and stated more formally for data-scarce learning as follows. Our objective is to leverage models trained on data from mismatched domain(s) to obtain a model for data from the target domain. We attempt to do so by starting with a prior estimate of the model's parameter set $\mathbf{\Lambda}_0$ and use the small amount of training/adaptation data from the task-at-hand $\mathcal{O}$, to obtain an updated parameter-set $\mathbf{\Lambda}^*$. The prior estimate $\mathbf{\Lambda}_0$ itself is obtained as some function of models $\mathbf{\Lambda}_{0_i}$ learned from data $\mathcal{O}_i$ in domains $\mathcal{D}_i$ (along with target-domain data $\mathcal{O}$, if necessary):

$$\mathbf{\Lambda}_0 = f\left(\left\{\mathbf{\Lambda}_{0_i}\right\}_i, \mathcal{O}\right)$$

The leveraging function f is chosen from a function class $\mathbb{F}$ using model selection criterion $\mathcal{C}$:

$$\hat{f} = \arg\max_{f\in\mathbb{F}} \mathcal{C}\left(\mathcal{O}, \left\{\vec{\lambda}_{0_i}\right\}_i, f\right)$$

The updated parameter-set $\mathbf{\Lambda}^*$ is then given by:

$$\mathbf{\Lambda}^* = \hat{f}\left(\left\{\mathbf{\Lambda}_{0_i}\right\}_i, \mathcal{O}\right)$$

The choice of criterion $\mathcal{C}$ can be very important, from the perspective of performance. We have seen in the previous chapter that adapting transition probabilities did not cause any significant change to recognition accuracy, regardless of their source (SI vs. SDB vs. IM). It can be explained by the fact that the traditional auxiliary likelihood function employed in HMM re-estimation is "mostly influenced by the emission distributions and almost not at all by the transition probabilities ... hence temporal aspects are poorly taken into account" [121]. The machine learning community has been interested in large-margin and kernel based approaches for quite some time now, and some researchers have successfully used such criteria for learning HMM parameters [122, 123, 124]. Adaptation of domain-mismatched HMMs should be able to benefit from these approaches too.

References

[1] M. Fried-Oken, "Voice recognition device as a computer interface for motor and speech impaired people," *Archives of Physical Medicine and Rehabilitation*, vol. 66, pp. 678–681, 1985.

[2] G. S. Carlson and J. Bernstein, "Speech recognition of impaired speech," in *Proceedings of the 10th Annual Conference on Rehabilitation Technology*, R. D. Steel and W. Gerrey, Eds., 1987, pp. 165–167.

[3] C. L. Coleman and L. S. Meyers, "Computer recognition of the speech of adults with cerebral palsy and dysarthria," *AAC: Augmentative and Alternative Communication*, vol. 7, no. 1, pp. 34–42, 1991.

[4] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, 1976.

[5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice Hall, 1978.

[6] P. Mermelstein and S. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, p. 357, 1980.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[8] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology.* Association for Computational Linguistics, 1994.

[9] L. Baum and J. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bulletin of the American Mathematical Society*, vol. 73, no. 3, pp. 360–363, 1967.

[10] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[11] J. Bilmes, “A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models,” ICSI TR-97-021, Tech. Rep., 1997.

[12] Cambridge University Engineering Department, “HTK speech recognition toolkit,” Dec. 2006. [Online]. Available: http://htk.eng.cam.ac.uk/.

[13] J. Gauvain and C. Lee, “Bayesian learning of Gaussian mixture densities for hidden Markov models,” in *Proceedings of the DARPA Speech and Natural Language Workshop*, 1991, pp. 272–277.

[14] J. Gauvain and C. Lee, “MAP estimation of continuous density HMM: theory and applications,” in *Proceedings of the DARPA Speech and Natural Language Workshop*, 1992, pp. 185–190.

[15] K. Shinoda and C. Lee, “Structural MAP speaker adaptation using hierarchical priors,” in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding.* IEEE, 1997, pp. 381–388.

[16] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[17] C. J. Leggetter and P. C. Woodland, “Flexible speaker adaptation using maximum likelihood linear regression,” in *Proceedings of the ARPA Spoken Language Technology Workshop.* Morgan Kaufmann, 1995, pp. 104–109.

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

[19] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[20] L. Neumeyer, A. Sankar, and V. Digalakis, “A comparative study of speaker adaptation techniques,” in *Proceedings of Eurospeech*, Madrid, 1995, pp. 1127–1130.

[21] V. Digalakis, D. Ritchev, and L. Neumeyer, “Speaker adaptation using constrained estimation of Gaussian mixtures,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, 1995.

[22] C. Chesta, O. Siohan, and C. Lee, “Maximum a posteriori linear regression for hidden Markov model adaptation,” in *Proceedings of Eurospeech*, Budapest, 1999, pp. 211–214.

[23] O. Siohan, T. A. Myrvoll, and C. Lee, “Structural maximum a posteriori linear regression for fast HMM adaptation,” *Computer Speech and Language*, vol. 16, no. 1, pp. 5–24, 2002.

[24] A. Gunawardana and W. Byrne, “Discounted likelihood linear regression for rapid speaker adaptation,” *Computer Speech and Language*, vol. 15, no. 1, pp. 15–38, 2001.

[25] T. Kosaka and S. Sagayama, “Tree-structured speaker clustering for fast speaker adaptation,” in *Proceedings of ICASSP*, Adelaide, Australia, 1994, pp. 245–248.

[26] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, “Rapid speaker adaptation in eigenvoice space,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[27] M. Tipping and C. Bishop, “Probabilistic principal component analysis,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[28] R. Sproat, F. Zheng, L. Gu, D. Jurafsky, I. Shafran, J. Li, S. Tsakalidis, Y. Su, Y. Zheng, and H. Zhou, “Dialectal Chinese speech recognition: Final report,” *Johns Hopkins University, Baltimore, MD, Tech. Rep., October*, 2004.

[29] F. L. Darley, A. E. Aronson, and J. R. Brown, “Differential diagnostic patterns of dysarthria,” *Journal of Speech and Hearing Research*, vol. 12, no. 2, pp. 246–269, 1969.

[30] F. L. Darley, A. E. Aronson, and J. R. Brown, “Clusters of deviant speech dimensions in the dysarthrias,” *Journal of Speech and Hearing Research*, vol. 12, no. 3, pp. 462–496, 1969.

[31] P. J. Zentay, “Motor disorders of the central nervous system and their significance for speech: Part I.Cerebral and cerebellar dysarthrias,” *The Laryngoscope*, vol. 47, no. 3, pp. 147–156, 1937.

[32] P. J. Zentay, “Motor disorders of the central nervous system and their significance for speech: Part II.Clinical forms of motor defects (the spastic child),” *The Laryngoscope*, vol. 47, no. 6, pp. 421–430, 1937.

[33] G. J. Canter, “Speech characteristics of patients with Parkinson’s disease: I. Intensity, pitch, and duration,” *Journal of Speech and Hearing Disorders*, vol. 28, no. 3, pp. 221–229, 1963.

[34] G. J. Canter, "Speech characteristics of patients with Parkinson's disease: II. Physiological support for speech," *Journal of Speech and Hearing Disorders*, vol. 30, no. 1, pp. 44–49, 1965.

[35] G. J. Canter, "Speech characteristics of patients with Parkinson's disease: III. Articulation, diadochokinesis, and over-all speech adequacy," *Journal of Speech and Hearing Disorders*, vol. 30, no. 3, pp. 217–224, 1965.

[36] I. Lehiste, "Some acoustic characteristics of dysarthric speech," *Bibliotheca Phonetica*, 1965.

[37] F. L. Darley, A. E. Aronson, and J. R. Brown, *Motor speech disorders.* Saunders, 1975.

[38] J. F. Werker and J. E. Tees, "Infant speech perception and phonological acquisition," in *Phonological Development: Models, Research, Implications*, C. A. Ferguson, L. Menn, and C. Stoel-Gammon, Eds. Timonium, MD: York Press, 1992, pp. 285–312.

[39] W. Ziegler and D. von Cramon, "Disturbed coarticulation in apraxia of speech: Acoustic evidence," *Brain and Language*, vol. 29, no. 1, pp. 34–47, 1986.

[40] R. D. Kent, J. F. Kent, G. Weismer, R. E. Martin, R. L. Sufit, B. R. Brooks, and J. C. Rosenbek, "Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects," *Clinical Linguistics & Phonetics*, vol. 3, no. 4, pp. 347–358, 1989.

[41] G. S. Turner, K. Tjaden, and G. Weismer, "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," *Journal of Speech and Hearing Research*, vol. 38, no. 5, pp. 1001–1013, 1995.

[42] Y. Kim, G. Weismer, R. D. Kent, and J. R. Duffy, "Statistical models of f2 slope in relation to severity of dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 6, pp. 329–335, 2009.

[43] H. Kim, M. Hasegawa-Johnson, and A. Perlman, "Vowel contrast and speech intelligibility in dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 63, no. 4, pp. 187–194, 2010.

[44] G. Weismer, "Motor speech disorders," *The Handbook of Phonetic Sciences. Cambridge, Blackwell*, pp. 191–219, 1997.

[45] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[46] G. Fant, *Acoustic Theory of Speech Production.* Mouton & Co, The Hague, Netherlands, 1960.

[47] S. Watanabe, K. Arasaki, H. Nagata, and S. Shouji, "[Analysis of dysarthria in amyotrophic lateral sclerosis–MRI of the tongue and formant analysis of vowels]." *Rinshō Shinkeigaku [Clinical Neurology]*, vol. 34, no. 3, pp. 217–223, 1994.

[48] W. Ziegler and D. von Cramon, "Vowel distortion in traumatic dysarthria: A formant study," *Phonetica*, vol. 40, no. 1, pp. 63–78, 1983.

[49] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 1, pp. 114–125, 2010.

[50] K. Tjaden, D. Rivera, G. Wilding, and G. S. Turner, "Characteristics of the lax vowel space in dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 48, no. 3, pp. 554–566, 2005.

[51] S. Sapir, J. Spielman, L. O. Ramig, S. L. Hinds, S. Countryman, C. Fox, and B. Story, "Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on ataxic dysarthria: A case study," *American Journal of Speech-Language Pathology*, vol. 12, no. 4, pp. 387–399, 2003.

[52] K. Tjaden and G. E. Wilding, "Rate and loudness manipulations in dysarthria: acoustic and perceptual findings," *Journal of Speech, Language, and Hearing Research*, vol. 47, no. 4, pp. 766–783, 2004.

[53] P. A. McRae, K. Tjaden, and B. Schoonings, "Acoustic and perceptual consequences of articulatory rate change in Parkinson disease," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 1, pp. 35–50, 2002.

[54] G. Weismer, N. Y. Jeng, J. S. Laures, R. D. Kent, and J. F. Kent, "Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders," *Folia Phoniatrica et Logopaedica*, vol. 53, pp. 1–18, 2001.

[55] G. Weismer, J. S. Laures, J. Y. Jeng, R. D. Kent, and J. F. Kent, "Effect of speaking rate manipulations on acoustic and perceptual aspects of the dysarthria in amyotrophic lateral sclerosis," *Folia Phoniatrica et Logopaedica*, vol. 52, no. 5, pp. 201–219, 2000.

[56] G. Weismer, R. D. Kent, M. Hodge, and R. Martin, “The acoustic signature for intelligibility test words,” *Journal of the Acoustical Society of America*, vol. 84, pp. 1281–1291, 1988.

[57] G. Weismer, R. Martin, R. D. Kent, and J. F. Kent, “Formant trajectory characteristics of males with amyotrophic lateral sclerosis,” *Journal of the Acoustical Society of America*, vol. 91, pp. 1085–1098, 1992.

[58] R. D. Kent, J. F. Kent, G. Weismer, R. L. Sufit, J. C. Rosenbek, R. E. Martin, and B. R. Brooks, “Impairment of speech intelligibility in men with amyotrophic lateral sclerosis,” *Journal of Speech and Hearing Disorders*, vol. 55, no. 4, pp. 721–729, 1990.

[59] R. D. Kent, R. L. Sufit, J. C. Rosenbek, J. F. Kent, G. Weismer, R. E. Martin, and B. R. Brooks, “Speech deterioration in amyotrophic lateral sclerosis: a case study,” *Journal of Speech and Hearing Research*, vol. 34, no. 6, pp. 1269–1275, 1991.

[60] J. F. Kent, R. D. Kent, J. C. Rosenbek, G. Weismer, R. E. Martin, R. Sufit, and B. R. Brooks, “Quantitative description of the dysarthria in women with amyotrophic lateral sclerosis,” *Journal of Speech and Hearing Research*, vol. 35, no. 4, pp. 723–733, 1992.

[61] M. Mulligan, J. Carpenter, J. Riddel, M. K. Delaney, G. Badger, P. Krusinski, and R. Tandan, “Intelligibility and the acoustic characteristics of speech in amyotrophic lateral sclerosis (ALS),” *Journal of Speech and Hearing Research*, vol. 37, no. 3, pp. 496–503, 1994.

[62] R. D. Kent, R. Netsell, and J. H. Abbs, “Acoustic characteristics of dysarthria associated with cerebellar disease,” *Journal of Speech and Hearing Research*, vol. 22, no. 3, pp. 627–648, 1979.

[63] A. J. Caruso and E. K. Burton, “Temporal acoustic measures of dysarthria associated with amyotrophic lateral sclerosis,” *Journal of Speech and Hearing Research*, vol. 30, no. 1, pp. 80–87, 1987.

[64] R. J. Morris, “VOT and dysarthria: a descriptive study,” *Journal of Communication Disorders*, vol. 22, no. 1, pp. 23–33, 1989.

[65] W. J. Hardcastle, R. A. Morgan Barry, and C. J. Clark, “Articulatory and voicing characteristics of adult dysarthric and verbal dyspraxic speakers: an instrumental study,” *British Journal of Disorders of Communication*, vol. 20, no. 3, pp. 249–270, 1985.

[66] G. Weismer, “Articulatory characteristics of Parkinsonian dysarthria: Segmental and phrase-level timing, spirantization, and glottal-supraglottal coordination,” in *The Dysarthrias: Physiology, Acoustics,*

Perception, Management, M. R. McNeil, J. C. Rosenbek, and A. E. Aronson, Eds. Timonium, MD: San Diego, CA: College Hill Press, 1984, pp. 101–130.

[67] R. D. Kent and J. C. Rosenbek, "Acoustic patterns of apraxia of speech," *Journal of Speech and Hearing Research*, vol. 26, no. 2, pp. 231–249, 1983.

[68] H. Ackermann and I. Hertrich, "Dysarthria in Friedreich's ataxia: timing of speech segments," *Clinical Linguistics & Phonetics*, vol. 7, no. 1, pp. 75–91, 1993.

[69] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *Journal of the Acoustical Society of America*, vol. 64, pp. 1358–1368, 1978.

[70] D. Kewley-Port, "Time-varying features as correlates of place of articulation in stop consonants," *Journal of the Acoustical Society of America*, vol. 73, no. 1, pp. 322–335, 1983.

[71] P. Shinn and S. E. Blumstein, "Phonetic disintegration in aphasia: Acoustic analysis of spectral characteristics for place of articulation," *Brain and Language*, vol. 20, no. 1, pp. 90–114, 1983.

[72] K. Forrest, G. Weismer, P. Milenkovic, and R. N. Dougall, "Statistical analysis of word-initial voiceless obstruents: preliminary data," *Journal of the Acoustical Society of America*, vol. 84, pp. 115–123, 1988.

[73] W. Ziegler and D. von Cramon, "Spastic dysarthria after acquired brain injury: An acoustic study," *International Journal of Language & Communication Disorders*, vol. 21, no. 2, pp. 173–187, 1986.

[74] K. Tjaden and G. S. Turner, "Spectral properties of fricatives in amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 6, pp. 1358–1372, 1997.

[75] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *Journal of the Acoustical Society of America*, vol. 108, pp. 1252–1263, 2000.

[76] H. T. Bunnell, J. Polikoff, and J. McNicholas, "Spectral moment vs. Bark cepstral analysis of childrens word-initial voiceless stops," in *Proceedings of the 8th International Conference on Spoken Language Processing.* Jeju, Korea, 2004.

[77] R. D. Kent, "Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders," *American Journal of Speech-Language Pathology*, vol. 5, no. 3, pp. 7–23, 1996.

[78] P. D. Neilson and N. J. O'Dwyer, "Reproducibility and variability of speech muscle activity in athetoid dysarthria of cerebral palsy," *Journal of Speech and Hearing Research*, vol. 27, no. 4, pp. 502–517, 1984.

[79] N. J. O'Dwyer and P. D. Neilson, "Voluntary muscle control in normal and athetoid dysarthric speakers," *Brain*, vol. 111, no. 4, pp. 877–899, 1988.

[80] G. Weismer, "Philosophy of research in motor speech disorders," *Clinical Linguistics & Phonetics*, vol. 20, no. 5, pp. 315–349, 2006.

[81] B. J. Zyski and B. E. Weisinger, "Identification of dysarthria types based on perceptual analysis," *Journal of Communication Disorders*, vol. 20, pp. 367–378, 1987.

[82] J. Zeplin and R. D. Kent, "Reliability of auditory-perceptual scaling of dysarthria," in *Disorders of Motor Speech: Assessment, Treatment, and Clinical Characterization*, D. A. Robin, K. M. Yorkston, and D. R. Beukelman, Eds. Baltimore, MD: Paul H. Brookes Publishing Co., 1996, pp. 145–154.

[83] K. Bunton, R. D. Kent, J. R. Duffy, J. C. Rosenbek, and J. F. Kent, "Listener agreement for auditory-perceptual ratings of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 6, pp. 1481–1495, 2007.

[84] S. Fonville, H. B. van der Worp, P. Maat, M. Aldenhoven, A. Algra, and J. van Gijn, "Accuracy and inter-observer variation in the classification of dysarthria from speech recordings," *Journal of Neurology*, vol. 255, no. 10, pp. 1545–1548, 2008.

[85] M. Van der Graaff, T. Kuiper, A. Zwinderman, B. Van de Warrenburg, P. Poels, A. Offeringa, A. Van der Kooi, H. Speelman, and M. De Visser, "Clinical identification of dysarthria types among neurologists, residents in neurology and speech therapists," *European Neurology*, vol. 61, no. 5, pp. 295–300, 2009.

[86] Y. Kim, R. D. Kent, and G. Weismer, "An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 2, pp. 417–429, 2011.

[87] G. Weismer and Y. Kim, "Classification and taxonomy of motor speech disorders: what are the issues?" in *Speech Motor Control: New developments in basic and applied research*, B. Maassen and P. van Lieshout, Eds. New York: Oxford University Press, 2010, pp. 229–242.

[88] L. J. Ferrier, H. C. Shane, H. F. Ballard, T. Carpenter, and A. Benoit, "Dysarthric speakers intelligibility and speech characteristics in relation to computer speech recognition," *Augmentative and Alternative Communication*, vol. 11, no. 3, pp. 165–174, 1995.

[89] F. Chen and A. Kostov, "Optimization of dysarthric speech recognition," in *Proceedings of International Conference of the IEEE Engineering in Medicine and Biology Society*, 1997, pp. 1436–1439.

[90] J. Deller, D. Hsu, and L. Ferrier, "Encouraging results in the automated recognition of cerebral palsy speech," *IEEE Transactions on Biomedical Engineering*, vol. 35, no. 3, pp. 218–220, 1988.

[91] J. Deller, D. Hsu, and L. Ferrier, "On the use of hidden Markov modelling for recognition of dysarthric speech," *Computer Methods and Programs in Biomedicine*, vol. 35, no. 2, pp. 125–139, 1991.

[92] P. D. Polur and G. E. Miller, "Effect of high-frequency spectral components in computer recognition of dysarthric speech based on a mel-cepstral stochastic model," *Journal of Rehabilitation Research & Development*, vol. 42, no. 3, pp. 363–372, 2005.

[93] P. D. Polur and G. E. Miller, "Experiments with fast Fourier transform, linear predictive and cepstral coefficients in dysarthric speech recognition algorithms using hidden Markov model," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 4, pp. 558–61, 2005.

[94] G. Jayaraman and K. Abdelhamad, "Experiments in dysarthric speech recognition using artificial neural networks," *Journal of Rehabilitation Research & Development*, vol. 32, no. 2, pp. 162–9, 1995.

[95] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, "HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria," in *Proceedings of ICASSP*, Toulouse, France, 2006.

[96] P. D. Polur and G. E. Miller, "Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals," *Medical Engineering and Physics*, vol. 21, no. 8, pp. 741–748, 2006.

[97] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," *AAC: Augmentative and Alternative Communication*, vol. 17, no. 4, pp. 265–275, 2001.

[98] F. Rudzicz, “Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech,” in *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility.* ACM, 2007, p. 256.

[99] X. Menéndez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, “The Nemours database of dysarthric speech,” in *Proceedings of the Fourth International Conference on Spoken Language Processing*, 1996, pp. 1962–1965.

[100] M. Hasegawa-Johnson and H. Sharma, “State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition,” in *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies.* Los Angeles, CA: Association for Computational Linguistics, June 2010. [Online]. Available: http://www.aclweb.org/anthology/W10-1310 pp. 72–79.

[101] M. Parker, S. Cunningham, P. Enderby, M. Hawley, and P. Green, “Automatic speech recognition and training for severely dysarthric users of assistive technology: the STARDUST project,” *Clinical Linguistics & Phonetics*, vol. 20, no. 2-3, pp. 149–156, 2006.

[102] R. D. Kent, J. F. Kent, J. R. Duffy, J. E. Thomas, G. Weismer, and S. Stuntebeck, “Ataxic dysarthria,” *Journal of Speech, Language, and Hearing research*, vol. 43, no. 5, pp. 1275–1289, 2000.

[103] W. Ziegler, E. Hartmann, and P. Hoole, “Syllabic timing in dysarthria,” *Journal of Speech and Hearing Research*, vol. 36, no. 4, pp. 683–693, 1993.

[104] B. Blaney and J. Wilson, “Acoustic variability in dysarthria and computer speech recognition,” *Clinical Linguistics & Phonetics*, vol. 14, no. 4, pp. 307–327, 2000.

[105] S. K. Fager, “Duration and variability in dysarthric speakers with traumatic brain injury,” Ph.D. dissertation, University of Nebraska - Lincoln, 2008.

[106] K. Tjaden and J. Sussman, “Perception of coarticulatory information in normal speech and dysarthria,” *Journal of Speech, Language, and Hearing Research*, vol. 49, no. 4, pp. 888–902, 2006.

[107] G. Weismer and R. E. Martin, “Acoustic and perceptual approaches to the study of intelligibility,” in *Intelligibility in speech disorders: Theory, Measurement and Management*, R. D. Kent, Ed. Amsterdam: John Benjamins Publishing Company, 1992, pp. 67–118.

[108] J. M. Liss, S. Spitzer, J. N. Caviness, C. Adler, and B. Edwards, "Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech," *Journal of the Acoustical Society of America*, vol. 104, pp. 2457–2466, 1998.

[109] J. M. Liss, S. M. Spitzer, J. N. Caviness, C. Adler, and B. W. Edwards, "Lexical boundary error analysis in hypokinetic and ataxic dysarthria," *Journal of the Acoustical Society of America*, vol. 107, pp. 3415–3424, 2000.

[110] J. M. Liss, S. M. Spitzer, J. N. Caviness, and C. Adler, "The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria," *Journal of the Acoustical Society of America*, vol. 112, pp. 3022–3030, 2002.

[111] S. L. Mattys and J. M. Liss, "On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as artificial normality," *Perception & Psychophysics*, vol. 70, no. 7, pp. 1235–1242, 2008.

[112] J. R. Deller, M. S. Liu, L. J. Ferrier, and P. Robichaud, "The Whitaker database of Dysarthric (Cerebral Palsy) Speech," *Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3516–3518, Sep. 1993.

[113] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proceedings of Interspeech*, Brisbane, Australia, Sep. 22–26, 2008.

[114] H. Kucera and W. N. Francis, *Computational Analysis of Present-Day American English.* Providence, RI: Brown University, 1967.

[115] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, pp. 1–19, 2011.

[116] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[117] M. Hasegawa-Johnson, H. Sharma, J. Huang, X. Zhuang, X. Zhou, and C. Hu, "Statistical Speech Technology Group: Beckman Institute, University of Illinois," 2011, talk given at University of Toronto, Canada.

[118] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings of ICASSP*, Glasgow, UK, 1989, pp. 532–535.

[119] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[120] N. M. I. Group, "NIST Multimodal Information Group Website," Feb. 2010. [Online]. Available: http://www.itl.nist.gov/iad/mig/tools/.

[121] J. Keshet and S. Bengio, *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods.* John Wiley & Sons, Ltd., 2009.

[122] X. Li, H. Jiang, and C. Liu, "Large margin HMMs for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*, vol. 5. IEEE, 2005, pp. 513–516.

[123] H. Jiang, X. Li, and C. Liu, "Large margin hidden Markov models for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1584–1595, 2006.

[124] D. Yu, L. Deng, X. He, and A. Acero, "Large-margin minimum classification error training for large-scale speech recognition tasks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2007*, vol. 4. IEEE, 2007, pp. 1137–1140.