# Acoustic Model Adaptation using in-domain Background Models for Dysarthric Speech Recognition

Harsh Vardhan Sharma[1,*], Mark Hasegawa-Johnson[1]

[a] *Department of Electrical and Computer Engineering*
*University of Illinois, Urbana, IL, USA*
[b] *Beckman Institute for Advanced Science and Technology*
*University of Illinois, Urbana, IL, USA*

## 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 it. 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 currently well-studied adaptation algorithms (which make no attempt to address this extent of mismatch in population characteristics).

In this work, we propose an interpolation-based technique for obtaining a prior acoustic model from one trained on unimpaired speech, before adapting it to the dysarthric talker. The method computes a 'background' model of the dysarthric talker's general speech characteristics and uses it to obtain a more

[*] Corresponding Author.
*Email addresses:* `hsharma@illinois.edu` (Harsh Vardhan Sharma), `jhasegaw@illinois.edu` (Mark Hasegawa-Johnson)

suitable prior model for adaptation (compared to the speaker-independent model trained on unimpaired speech). The approach is tested with a corpus of dysarthric speech acquired by our research group, on speech of sixteen talkers with varying levels of dysarthria severity (as quantified by their intelligibility). This interpolation technique is tested in conjunction with the well-known maximum *a posteriori* (MAP) adaptation algorithm, and yields improvements of up to 8% absolute and up to 40% relative, over the standard MAP adapted baseline.

*Keywords:* HMM, dysarthria, acoustic model, adaptation, UBM, MAP

## 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, speaker independent (SI) 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.

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 speaker to speaker, 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 (Kent et al., 1999).

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 (Fried-Oken, 1985; Carlson and Bernstein, 1987; Coleman and Meyers, 1991), instead of a keyboard.

One of the issues with developing ASR systems for dysarthric speakers 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 a speaker dependent (SD) ASR system. Speaker adaptation (SA) 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 speakers with no pathology are of less use to speakers with dysarthria, because the errors characteristic of dysarthria dramatically increase word error rates. The goal of this study 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-it-is. Conventional adaptation techniques such as maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995; Digalakis et al., 1995), maximum *a posteriori* adaptation (MAP) (Gauvain and Lee, 1991, 1992), or structured MAP (SMAP) (Shinoda and Lee, 1997) do not explicitly account for this extent of mismatch.

This article is organized as follows: Section 2 provides an overview of clinical research on motor speech disorders, with a special focus on results that have contributed to motivation for the algorithms developed in this article. Section 3 describes the proposed background interpolation and BI-MAP (background interpolated MAP) algorithms. Section 4 describes experimental methods and numerical results; Section 5 presents a more qualitative analysis of these results. Section 6 concludes the article with a review of the work's key findings.

## 2. Background

This section present the clinical context for the study, and motivates qualitatively the adaptation technique that we propose later.

### *2.1. The Mayo Clinic system of classifying dysarthrias*

The speech correlates of gross motor disorders have been documented since ancient times, but perhaps the first large-scale hypothesis-driven scientific analysis specifically directed at the understanding of speech motor disorders was conducted by Darley, Aronson, and Brown of the Mayo Clinic (Darley et al., 1969a,b). Darley and colleagues proposed specifically the hypothesis that the type of neurological pathology is correlated with the type of speech production deficit ("... speech pathology reflects neuropathology" (Darley et al., 1975, page 229)), and that, in turn, the type of speech production deficit can be measured based on the expert rating of several perceptual attributes of the acoustic speech signal. Based on their extensive clinical experience with dysarthric patients, they chose 38 perceptual dimensions on which to rate each speech signal. Those perceptual dimensions were then combined in various ways to produce unique *clusters* of perceptual scores. Presence of a particular cluster was shown to be mildly predictive of the underlying neuropathology, thus confirming the original hypothesis.

The database used in this study was designed to be homogeneous with respect to the Mayo Clinic classification system: all talkers were diagnosed with either spastic dysarthria, or mixed dysarthria with a component of spasticity. The corpus was so designed in an attempt to limit variability in the types of articulatory substitutions and motor control errors produced by talkers in the database. The attempt was relatively unsuccessful: it was discovered that differences among the intelligibility levels of different talkers, and differences even between individuals with the same intelligibility level, cause tremendous differences in the types of adaptation necessary for successful automatic speech recognition.

The Mayo Clinic system was designed as a method for the labeling of neuropathology. It was not designed to be a comprehensive description of all clinically salient attributes of the pathology exhibited by any particular individual, or even of the most clinically salient attributes, if "clinically salient" is defined to include goals other than identification of the underlying neuropathology. Weismer (2006) 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. The 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. (2011) 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.

The BI-MAP algorithm described in this paper was originally inspired by the observation that the most perceptually salient attribute of dysarthric speech is the degree to which it is outside the normal range of speech variability exhibited within the larger speech community. Kim et al. (2011) index these differences by the intelligibility of the talker, but for the purposes of automatic speech recognition, the most important index variable is not intelligibility, but the acoustic differences themselves: a talker whose speech is not far from the norm and a talker whose speech differs dramatically from the norm require different types of model adaptation for the development of successful automatic speech recognition.

### *2.2. Dysarthria and speaker-adapted ASR*

Very few studies exist on model adaptation for dysarthric speech: Raghavendra et al. (2001) 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 (2007) who compared the performance of SD and "SA" systems on the Nemours database (Menéndez-Pidal et al., 1996) 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 were not enough data in the database to represent intra-speaker variation.

Recently, Sharma and Hasegawa-Johnson (2010) 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 (transition probability matrix is an interpolation between those of a left-to-right HMM and a fully-ergodic HMM; off-diagonal entries are initialized to be comparatively much smaller than the diagonal entries) 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 the additional state-transitions in transition-interpolated HMMs (the off-diagonal entries in the transition probability matrix) 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.

### *2.3. Acoustic variability due to dysarthria characteristics: Impact on ASR*

Generally speaking, acoustic variability has a substantial impact on ASR accuracy. Parker et al. (2006) 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.

Blaney and Wilson (2000) tried to explain the source of intra-speaker variability for dysarthric and normal speakers, and to identify its relationship with ASR accuracy (for the DragonDictate, version 3.00 ASR system).

Significant intra-speaker variability was noted for speakers with dysarthria, with regard 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 distinctions were not preserved (a minimal pair being defined as a pair of words that differ in only one distinctive feature; Blaney and Wilson (2000) document several cases in which dysarthria erased the acoustic distinction between minimal pairs) and timing discrepancies were observed for word stem durations. Finally, some correlations were found between ASR accuracy and variability in Voice Onset Time (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 (2008) 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 Traumatic Brain Injury (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. 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."

### *2.4. Conclusion*

The work described above indicates that the acoustics of dysarthria exhibit certain attributes that are quite different from the attributes of unimpaired speech, and not well modeled by models designed for unimpaired speech. The literature in acoustic phonetics contains a large number of studies describing the acoustic correlates of phonological distinctive features (e.g., (Stevens , 1999)), which can be used to guide the design of automatic speech recognizers (e.g., (Hasegawa-Johnson et al. , 2004)). Conversely, phonetic studies of dysarthria have most often focused on the relationship between acoustics and neuropathology. Studies including (Kim et al., 2011) have demonstrated that the ability of any given talker to create the acoustic phonetic distinctions necessary for intelligible communication may depend on variables other than the talker's neuropathological diagnosis, but few such variables have been defined, and the details of the dependence have been described only in a small number of cases.

This study has chosen to investigate ASR system development research by addressing intra-speaker variability in the acoustic model, on account of the following:

- Differences among speakers with dysarthria are, in many cases, larger than the difference between a speaker with dysarthria and the speech of talkers without dysarthria. 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 (2010) have recently proposed a starting point for the development of that adequate theory in 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.

## 3. Modeling Mismatch with Background Interpolation

Conventional adaptation techniques such as MLLR, MAP, etc. 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. When the mismatch between training population and target population becomes large enough, however, conventional approaches fail. More precisely, each conventional approach generates a speaker-adapted model that represents only a subset of the differences between the training and target speakers. MLLR, for example, creates a speaker-adapted model that is a piece-wise linear transformation of the speaker-independent model; any differences between training and target speakers that can not be represented by a piece-wise linear transformation are not captured by MLLR (Leggetter and Woodland, 1995). The limitations of MAP adaptation are less widely reported, since, as the amount of speaker-dependent training data increases, a MAP-adapted

HMM converges to a maximum of the speaker-dependent likelihood function (Gauvain and Lee, 1991). It is less widely recognized that the likelihood function of an automatic speech recognition system has typically a large finite number of local maxima (Baum and Eagon , 1967), and that if a talker's speech diverges too far from the speaker-independent norm, the result of MAP adaptation for that talker may therefore converge to a significantly suboptimal local maximum of the speaker-dependent likelihood function. Both MAP and MLLR, therefore, are able to learn optimum speaker-adapted speech recognition models only for speakers who deviate from the training population in a limited way, and indeed, absent side information about the target speaker, the same limitation may apply to any other speaker adaptation algorithm. In order to represent a speaker with larger deviations from the norm, it is necessary to use some type of side information to guide adaptation: it is necessary to know something about the way in which the target speaker differs from the norm.

The last section 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 it 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, some notation that will be used in the remainder of this paper. Let $\mathbf{\Lambda}$ denote the AM parameter set for an ASR system in vector notation. 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 $N_{HMM}$ $N$-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) — $i, j \in \{1, \ \dots \ , N\}$; $l \in \{1, \ \dots \ , M\}$; $n \in \{1, \ \dots \ , N_{HMM}\}$. Speaker adaptation is performed in two stages: first, a model that accounts for the mismatch between speaker populations (see Sections 3.1.1 and 3.1.2) is obtained; in the second stage, this model is used as the prior or initial model (see Section 3.1.3) for the actual adaptation.

### *3.1. Background Interpolation: Formulation*

#### *3.1.1. Speaker-Dependent Background models*

The Universal Background Model (UBM) is an effective and widely used framework (Reynolds et al., 2000) 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. The UBM can be viewed as a global evidence function, specifying the range and distribution of speech features when one is given no further information about the phonetic content or speaker identity, and because of this relatively broad interpretation it has been used for many purposes beyond automatic speaker identification, e.g., for automatic language identification (Zhou, Navrátil, Pelecanos, Ramaswamy and Huang, 2008) and to guide the further training of an automatic speech recognizer (Povey, Chu and Varadarajan, 2008); related algorithms have also been used, with different training methods, to rescore non-speech acoustic event detection (Zhuang, Zhou, Hasegawa-Johnson and Huang, 2010) and detect falling object events (Zhuang, Huang, Potamianos, and Hasegawa-Johnson, 2009), to classify visual scenes (Zhou, Zhuang, Tang, Hasegawa-Johnson and Huang, 2010) and locate visible objects (Zhuang, Zhou, Hasegawa-Johnson and Huang, 2009), to estimate the age of a face image (Zhuang, Zhou, Hasegawa-Johnson and Huang, 2008), and to classify emotion of both face (Tang, Hasegawa-Johnson and Huang, 2010) and voice (Tang, Chu, Hasegawa-Johnson and Huang, 2009). Using a phone-independent evidence function like the UBM to guide phone-dependent likelihood functions, like those used in speech recognition, requires substantial additional data and additional computation, e.g., Povey et al. (2008) subdivided the UBM into a hierarchy of phone models using MAP adaptation down the branches of a model tree. The algorithm proposed in this paper takes a very different approach from that of Povey et al., essentially replacing the UBM with a model that guides adaptation toward the target population; in this respect the algorithms proposed in this paper are inspired by the target/non-target interpolation of Stiedl et al. (Steidl et al. , 2003), and by eigenvoice methods more generally (Kuhn et al., 2000), and can be viewed as a special case of these methods tailored to the greater degree of inter-speaker variability evident in dysarthric speech.

To develop a speaker-specific model of population mismatch, a speaker-dependent background (SDB) model, $\mathbf{\Lambda}_{SDB}$ is first created: a Gaussian Mixture Model (GMM) with the same number of components as any state-specific distribution in the to-be-adapted SI system is trained using all speech (the transcriptions are not used; speech vs. silence at beginning/end of utterance is the only distinction made) from the dysarthric speaker. This GMM is then replicated for each state of an HMM to obtain an SDB HMM. Finally this SDB HMM is replicated for each HMM in the SI system, to obtain $\mathbf{\Lambda}_{SDB}$. The SDB model 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 (and not on what was spoken by them). Since all speech (excluding test-set speech) from the target speaker is used for creating the SDB, it can have a high parameter count (as many parameters as a GMM from the SI system) like the UBM.

### *3.1.2. 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 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 $\mathbf{\Lambda}^*_{SI}$.

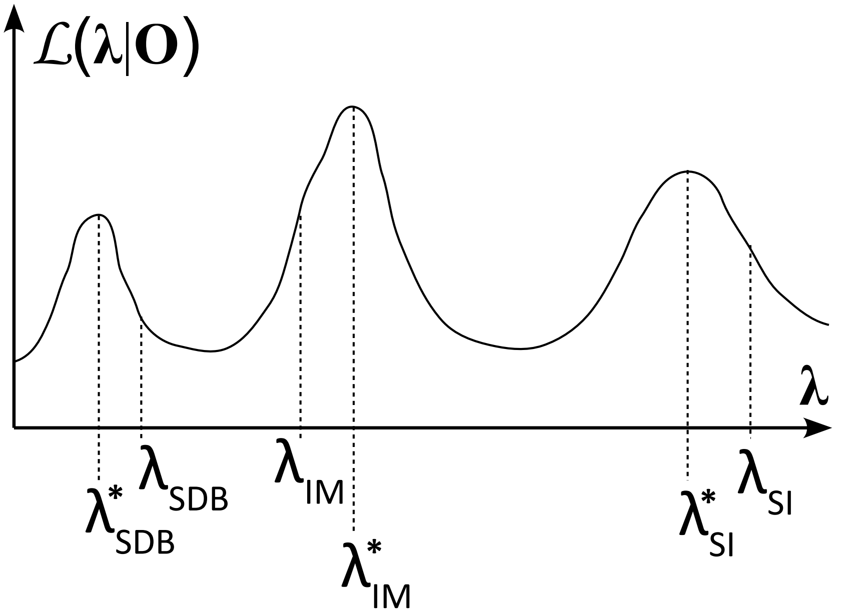

Figure 1: *Schematic description of a situation in which the BI-MAP algorithm would result in a better model than standard SI-MAP. Schematic depicts a fictitious likelihood curve in which MAP adaptation beginning from the SDB results in a different final model than does MAP adaptation beginning from the SI model. By interpolating between these two initial points, it may be possible to find a third initial model that yields recognition accuracy higher than that achieved from either the SDB or SI starting points. Although the likelihood function in the figure is fictitious, experimental results presented in this paper confirm that background interpolation beats SI-MAP in most cases.*

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. This large separation means that it is possible (though certainly not guaranteed) that there may be 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 $\mathbf{\Lambda}_{SI}$:

$$\mathbf{\Lambda}_{IM} = \mathbf{\Delta} \cdot \mathbf{\Lambda}_{SI} + (\mathbf{I} - \mathbf{\Delta}) \cdot \mathbf{\Lambda}_{SDB} \tag{1}$$

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 $\mathbf{\Lambda}_{SDB}$.

### *3.1.3. 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.

## *3.2. Background Interpolated MAP adaptation (BI-MAP)*

Conventional/classical MAP adaptation (Gauvain and Lee, 1991) utilizes Dirichlet distribution priors for $\{\pi_i\}_i$, $\{a_{ij}\}_{i,j}$, $\{c_{il}\}_{i,l}$ and a Gamma-Normal distribution prior for each $\left\{\mu_{il_d}, r_{il_d} = \sigma_{il_d}^{-2}\right\}$ pair. Ignoring constant terms, the overall log-prior for an HMM is ($\lambda$ denoting the parameter set for a single HMM):

$$\begin{aligned}
\log G(\lambda) = &\sum_{i=1}^{N} \pi_{i_0} M\tau \cdot \log(\pi_i) \\
&+ \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij_0} M\tau \cdot \log(a_{ij}) \\
&+ \sum_{i=1}^{N}\sum_{l=1}^{M} c_{il_0} M\tau \cdot \log(c_{il}) \\
&+ \sum_{i=1}^{N}\sum_{l=1}^{M}\sum_{d=1}^{p} \left[\frac{\tau}{2} \cdot \log(r_{il_d}) - \frac{\tau\sigma_{il_{d_0}}^2}{2} r_{il_d}\right] \\
&- \sum_{i=1}^{N}\sum_{l=1}^{M}\sum_{d=1}^{p} \frac{\tau}{2} r_{il_d} (\mu_{il_d} - \mu_{il_{d_0}})^2
\end{aligned} \tag{2}$$

where $\{\pi_{i_0}\}_i$ , $\{a_{ij_0}\}_{i,j}$ , etc. constitute the starting/initial parameter set $\lambda_0$.

Combining $G(\lambda)$ with the maximum-likelihood (ML) auxiliary function gives us the MAP auxiliary function (for iteration $u+1$ of Expectation Maximization (EM), as a function of $\lambda^{(u)}$). Maximizing the MAP auxiliary function results in the following parameter re-estimates:

$$
\begin{aligned}
\pi_i^{(u+1)} &= \frac{\pi_{i_0} M\tau + \sum_{k=1}^{K} \gamma_i^{uk}(1)}{M\tau + \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_j^{uk}(1)} \\
a_{ij}^{(u+1)} &= \frac{a_{ij_0} M\tau + \sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \xi_{ij}^{uk}(t)}{M\tau + \sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \gamma_i^{uk}(t)} \\
c_{il}^{(u+1)} &= \frac{c_{il_0} M\tau + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)}{M\tau + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_i^{uk}(t)} \\
\mu_{il_d}^{(u+1)} &= \frac{\tau \mu_{il_{d_0}} + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t) o_{t_d}^{k}}{\tau + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)} \\
\sigma^2_{il_d}{}^{(u+1)} &= \frac{\tau \sigma^2_{il_{d_0}} + \tau(\mu_{il_d}^{(u+1)} - \mu_{il_{d_0}})^2}{\tau + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)} \\
&\quad + \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t) \cdot (o_{t_d}^{k} - \mu_{il_d}^{(u+1)})^2}{\tau + \sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)}
\end{aligned}
\tag{3}
$$

where $\vec{o_t}^k$ is the observation vector at time $t$ from the $k^{\text{th}}$ observation sequence $\mathcal{O}_k = \{\vec{o_1}^k \dots \vec{o_t}^k \dots \vec{o_{T_k}}^k\}$; 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 $\lambda^{(u)}$).

The $\tau$ hyperparameter in the above equations 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 (2) is still valid. The difference is that the starting/initial parameter set $\lambda_0$ is now an interpolation between the SI parameter set $\lambda_0^{\mathrm{I}}$, and the SDB parameter set $\lambda_0^{\mathrm{D}}$:

$$
\begin{aligned}
\pi_{i_0} &= \delta_i \cdot \pi_{i_0}^{\mathrm{I}} + (1-\delta_i) \cdot \pi_{i_0}^{\mathrm{D}} \\
a_{ij_0} &= \delta_{ij} \cdot a_{ij_0}^{\mathrm{I}} + (1-\delta_{ij}) \cdot a_{ij_0}^{\mathrm{D}} \\
c_{il_0} &= \delta_{il}^{w} \cdot c_{il_0}^{\mathrm{I}} + (1-\delta_{il}^{w}) \cdot c_{il_0}^{\mathrm{D}} \\
\vec{\mu}_{il_0} &= \delta_{il}^{m} \cdot \vec{\mu}_{il_0}^{\mathrm{I}} + (1-\delta_{il}^{m}) \cdot \vec{\mu}_{il_0}^{\mathrm{D}} \\
\sigma_{il_{d_0}}^{2} &= \delta_{il}^{s} \cdot \sigma_{il_{d_0}}^{2^{\mathrm{I}}} + (1-\delta_{il}^{s}) \cdot \sigma_{il_{d_0}}^{2^{\mathrm{D}}}
\end{aligned}
\tag{4}
$$

where the $\delta$s 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 $\delta$s to be the same $\delta$. This is the same as setting $\delta_i = \delta \; \forall \; i$ in Equation 1.

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

$$
\begin{aligned}
\vec{\mu}_{il}^{(u+1)} = {} & \frac{\tau_{\mathrm{I}}}{\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)} \cdot \vec{\mu}_{il_0}^{\mathrm{I}} \\
& + \frac{\tau_{\mathrm{D}}}{\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)} \cdot \vec{\mu}_{il_0}^{\mathrm{D}} \\
& + \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)\vec{o}_t^{k}}{\tau + \sum_{k=1}^{K}\sum_{t=1}^{T_k} \gamma_{il}^{uk}(t)}
\end{aligned}
\tag{5}
$$

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 4 in Equation 3.

## 4. Experiments

### *4.1. The UA-Speech corpus of dysarthric speech*

All experiments in this study have made use of the UA-Speech corpus (Kim et al., 2008). It contains recordings of 16 subjects informally diagnosed with dysarthria. All speakers have the same clinical diagnosis (spastic dysarthria as a symptom of spastic CP): there is no variability in the corpus in terms of clinical or physiological variables, only in terms of intelligibility of the speaker.

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"), the 26 letters of the international radio alphabet ("alpha, bravo, charlie,...."), 19 computer commands ("command, enter, paragraph,..."), and the 100 most common words ("is, it,...") in the Brown corpus of written English (Kucera and Francis, 1967). The uncommon words were selected from 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 were usually polysyllabic (more than half of the uncommon words were trisyllabic or longer). Each speaker recorded a total of 765 words, including 455 distinct words. Intelligibility assessment is described in (Kim et al., 2008). Two hundred distinct words were selected from the recording of the second block: 10 digits, 25 radio alphabet letters, 19 computer commands and, 73 words randomly selected from each of the 'common words' and 'uncommon words' categories. Five naive listeners were recruited for each speaker and were instructed to provide orthographic transcriptions of each word that they thought the speaker said. The percentage of correct responses was then averaged across five listeners to obtain each speaker's intelligibility. Based on their intelligibility rating, each speaker was classified as belonging to one of these categories: *Very Low* (0–25%), *Low* (25–50%), *Mid* (50–75%), and *High* (75–100%).

### *4.2. 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. All steps in the experiments (with the exception of model interpolation) were performed using the HTK toolkit.

#### *4.2.1. AM-space Gaussian alignment*

When performing BI-MAP interpolation, 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 $\mathbf{\Lambda}_{SDB}$. Optimal correspondence is difficult to define, e.g., should Gaussians be considered comparable if they have similar mean vectors, similar covariance matrices, or if they are minimally divergent according to an information theoretic measure? In the method proposed here, most Gaussians will be not be comparable by any of these definitions: the SDB represents speech of all phones, whereas each SI model represents speech corresponding to only one phone, therefore any indexed correspondence among the Gaussian components is somewhat arbitrary. An approximate solution was therefore adopted, according to which correspondence among the Gaussians is established prior to training, and ignores any divergence among corresponding components caused by further training (Fig. 2).

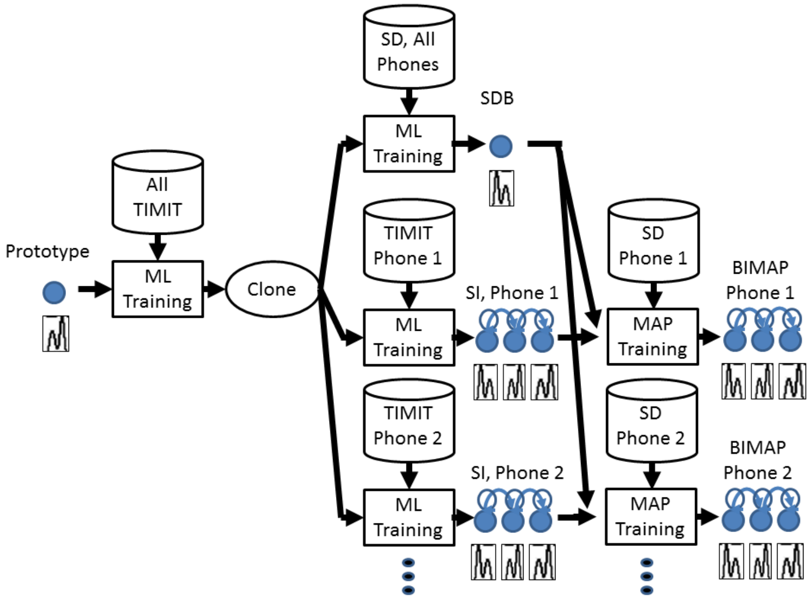

Figure 2: *The speaker-dependent background model (SDB) and speaker-independent phone models (SI) must be trained in a manner that preserves, as well as possible, the correspondence among identically indexed Gaussian components. Perfect correspondence is impossible, but the training algorithm depicted in this figure maintains a useful approximate correspondence. Cylinders denote speech data, circles denote HMM states parameterized by GMM mean, variance, and mixture weights, and rectangles denote GMM or HMM training.*

Fig. 2 describes the training procedure used to define a correspondence among the Gaussian components of different HMMs. Here, a single prototype HMM with the final number of Gaussians was first created, and trained using all of the TIMIT training set (all data from 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 $\mathbf{\Lambda}_{SI}$, the prototype 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) and this cloned AM was re-estimated completely using maximum likelihood (ML) training. To obtain $\mathbf{\Lambda}_{SDB}$, the 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).

### *4.2.2. Interpolation parameters – independent or dependent?*

In principle, one can have as many interpolation factors (the $\delta$s) 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 $\delta$ for all AM parameters" scenario. Fixing $\delta$ 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 came from $\mathbf{\Lambda}_{SDB}$ ($\delta = 0$). The interpolation $\delta$ 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.

### *4.3. Evaluation*

*Architecture, speech features, and corpus for mismatched-population*: These were identical to those for the experiments described in (Sharma and Hasegawa-Johnson, 2010). All acoustic models in this study were context-dependent, tied-state, word-internal triphone HMMs (three hidden states). The features extracted from the speech waveform comprised of 12 Perceptual Linear Prediction (PLP) coefficients 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 resulted in a 39-dimensional acoustic feature vector. 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. Blocks 1 and 3 for each dysarthric talker were used as adaptation data (and for computing their SDB model), and Block 2 was used as test data.

*Baseline*: The performance of BI-MAP adaptation was compared to that of the standard MAP adaptation (referred to as SI-MAP henceforth).

*Recognition performance evaluation*: Word Recognition Accuracy (WRA) was used for these experiments. In the test setup described in (Sharma and Hasegawa-Johnson, 2010), the test corpus contains one isolated utterance of each word in the vocabulary; therefore, WRA equals the fraction of vocabulary items correctly recognized.

*Significance Testing*: Statistical significance of the difference in ASR recognition accuracies between two ASR systems was compared at two levels. Gillick and Cox (Gillick and Cox, 1989) proposed two tests for the comparison of isolated word recognition accuracies: for cases in which the two systems recognize different waveform tokens, they proposed using McNemar's test, while for cases in which the two systems both assign labels to the same set of waveform tokens, they proposed using a more sensitive matched-pairs test. The Gillick-Cox matched-pairs test was used for each speaker with dysarthria to determine if the difference in the recognition accuracies using BI-MAP versus SI-MAP was statistically significant. In the event that SI-MAP has higher recognition accuracy for some speakers and BI-MAP for other speakers, a Wilcoxon signed-rank test (Wilcoxon, 1945) is then used to determine whether or not there is any overall group preference.

### *4.4. 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 1. 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 $\delta$ 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 $\tau$ set to 5.0 for all configurations.

Table 1: *BI-MAP system configurations studied. Gaussian means and variances were always interpolated.*

| **BI-MAP** | **Prior** | **Prior** |
|---|---|---|
| **config.** | **Mixt. Weights** | **Trans. Probs.** |
| C00 | interpolated | interpolated |
| C01 | | SI |
| C02 | | SDB |
| C10 | SI | interpolated |
| C11 | | SI |
| C12 | | SDB |

### *4.5. Results*

Columns 3 and 4 in Table 2 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 $\delta$ and the system configuration that gave the best WRA. For each speaker, the better of the two scores is 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.

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.

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 $\alpha$ levels of the Gillick-Cox test.

Figure 3 plots the WRA of the six BI-MAP configurations, as a function of $\delta$, for 4 speakers – one from each intelligibility category, obtained at steps of size 0.05 between 0 and 1. For all speakers, we see gradual improvement as one moves away from $\mathbf{\Lambda}_{SDB}$ ($\delta = 0$) and towards $\mathbf{\Lambda}_{SI}$ ($\delta = 1$). $WRA\,(\delta)$ peaks at an intermediate value of $\delta$, with the optimal $\delta$ occurring between 0.5 and 0.9.

## 5. Discussion

### *5.1. Evaluating what the Acoustic Model learnt*

It is clear that BI-MAP was able to obtain higher recognition accuracy compared to the conventional SI-MAP technique. It would be interesting to see if the acoustic models (HMMs for the sub–word units) converge to significantly different final models from different prior models.

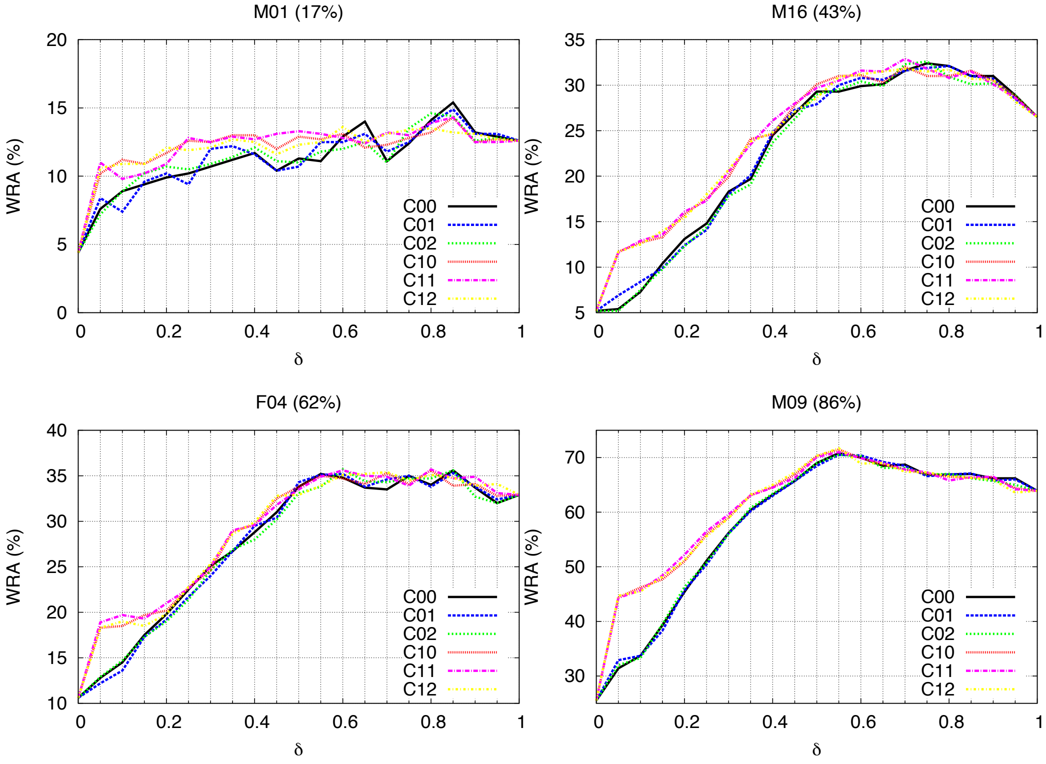

Figure 3: *Recognition Accuracies for representative speakers of each intelligibility category.*

After data-driven clustering of HMM states, the acoustic models generated in this study's experiments ended up having 3041 unique states, called senones. 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 henceforth. In order to explore effects of the prior model on the converged final model, we want to efficiently select those senones whose spectra are significantly different in SI-MAP and BI-MAP adapted acoustic models, and for which BI-MAP had better WRA.

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

Step 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 senone labels for each test-set utterance. For BI-MAP the hypothesis transcription came from the configuration (C00, etc.) with the best WRA.

Step 2. String alignment was performed for senone-label sequences, for a.) reference aligned with SI-MAP hypothesis; and b.) reference aligned with BI-MAP hypothesis.

Step 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 generated test-set transcriptions (when comparing reference transcription to the SI-MAP and BI-MAP hypothesis transcriptions): the idea here is to compare for each test-set utterance, the location of a particular senone label in the reference transcription and its location in the transcription generated by the ASR with the adapted acoustic model. If they occur at the same location for both SI-MAP and BI-MAP generated transcriptions, relative to the location in the reference transcription, then the two adaptation techniques have not learnt that senone differently (at least for that particular senone's location in that particular test-set utterance). This is determined for every occurrence of that senone-label in the entire test-set reference transcription. Finally, this process is repeated 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 of the test-set utterances.

Step 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 transformed to obtain the log magnitude Fourier 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).

Step 5. From these 228 spectral representations, the ones for which SI-MAP and BI-MAP spectra were visually similar were discarded: we are interested in senone spectra that are different for the SI-MAP and BI-MAP models. 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 position within the phone segment (2 = phone-initial, 3 = phone-medial, 4 = phone-final). The remaining digits index a particular set of triphone contexts: e.g., `t_s320` is the 20th set of triphone contexts for which a distinct senone was trained to represent the medial state of `/t/`.

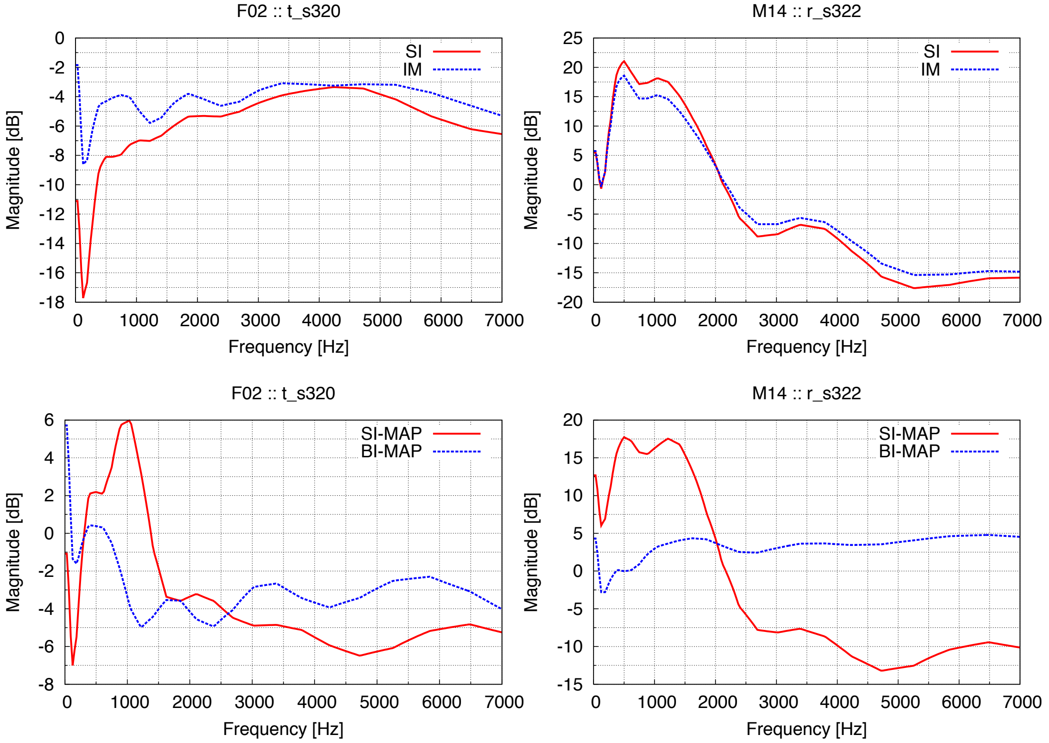

Figure 4: *Senone spectra before and after adaptation, for standard MAP and BI-MAP.*

Figure 4 shows the spectra for a senone each from speakers F02 and M14: the twentieth version of the middle emitting state for the unvoiced stop '`t`' for F02, and the twenty-second version of the middle emitting state for the liquid '`r`' for M14. 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 F02, and C02 with $\delta = 0.85$ for M14.

For F02's '`t_s320`', the SI-MAP spectrum exhibits a peak near F1, suggesting that the Baum-Welch algorithm incorrectly aligned this state with acoustic spectra from neighboring vowels; the BI-MAP spectrum shows no such spurious peak. Background interpolation with F02's SDB model has helped here because it allowed better state alignment during MAP adaptation.

Conversely for M14, SI-MAP has learned an /r/-like spectrum, but BI-MAP has not. Possibly in this case the BI-MAP model has suffered state misalignment, or possibly M14's 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 $\delta$), 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).

### *5.2. Differences in the behavior of transition probabilities vs. mixture weights*

The experiments described in the previous sections indicate that the mixture weights affected WRA differently from the 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 Sharma and Hasegawa-Johnson (2010), 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 (in both cases, the means and variances were adapted).

2. Looking at the WRA curves for BI-MAP adapted systems again (Figure 3), we see that the C1 subgroup of configurations had higher recognition accuracies compared to the C0 subgroup of configurations, for low values of $\delta$. In that range, the interpolated prior model is not too far from the SDB prior model in the AM-parameter space. Forcing the mixture weights to 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).

## 6. Conclusion

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. Inspired by the hypothesis of Weismer and Kim (2010), this study has developed a model in which the distance between training population and target speaker is explicitly modeled. This distance is so great that even a relatively crude speaker-dependent model (the SDB), trained with very little speaker-dependent data, serves to anchor the far end of a BI-MAP interpolation.

For a parameter count almost identical to the baseline approach (BI-MAP has used only one additional hyperparameter: the interpolation factor $\delta$), background interpolation has been able to achieve significantly better recognition accuracy, for the UA-Speech corpus.

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. However, it has been hitherto unexplored for speaker-adaptation of HMMs, 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.

## Acknowledgment

This study was funded by NSF grant IIS 05-34106. All opinions expressed here are those of the authors. We would also like to thank Gary Weismer and Torrey Loucks for their suggestions and comments.

## References

Baum, L.E., Eagon, J.A., 1967. An Inequality With Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology. Bull. Am. Math. Soc. 73, 360–363.

Blaney, B., Wilson, J., 2000. Acoustic variability in dysarthria and computer speech recognition. Clinical Linguistics & Phonetics 14, 307–327.

Bunton, K., Kent, R.D., Duffy, J.R., Rosenbek, J.C., Kent, J.F., 2007. Listener agreement for auditory-perceptual ratings of dysarthria. Journal of Speech, Language, and Hearing Research 50, 1481–1495.

Carlson, G.S., Bernstein, J., 1987. Speech recognition of impaired speech, in: Steel, R.D., Gerrey, W. (Eds.), Proceedings of the 10th Annual Conference on Rehabilitation Technology, pp. 165–167.

Coleman, C.L., Meyers, L.S., 1991. Computer recognition of the speech of adults with cerebral palsy and dysarthria. AAC: Augmentative and Alternative Communication 7, 34–42.

Darley, F.L., Aronson, A.E., Brown, J.R., 1969a. Clusters of deviant speech dimensions in the dysarthrias. Journal of Speech and Hearing Research 12, 462–496.

Darley, F.L., Aronson, A.E., Brown, J.R., 1969b. Differential diagnostic patterns of dysarthria. Journal of Speech and Hearing Research 12, 246–269.

Darley, F.L., Aronson, A.E., Brown, J.R., 1975. Motor speech disorders. Saunders.

Digalakis, V., Ritchev, D., Neumeyer, L., 1995. Speaker adaptation using constrained estimation of Gaussian mixtures. IEEE Transactions on Speech and Audio Processing 3, 357–366.

Fager, S.K., 2008. Duration and variability in dysarthric speakers with traumatic brain injury. Ph.D. thesis. University of Nebraska - Lincoln.

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

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

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

Gillick, L., Cox, S., 1989. Some statistical issues in the comparison of speech recognition algorithms, in: Proceedings of ICASSP, Glasgow, UK. pp. 532–535.

Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., Juneja, A., Kirchhoff, K., Livescu, K., Mohan, S., Muller, J., Sonmez, K., Wang, T., 2004. Landmark-Based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop, in: Proceedings of ICASSP, pp. 1213–1216.

Kent, R.D., Weismer, G., Kent, J.F., Vorperian, H.K., Duffy, J.R.., 1999. Acoustic Studies of Dysarthric Speech: Methods, Progress, and Potential J. Commun. Disord. 32, 141–186.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., Frame, S., 2008. Dysarthric speech database for universal access research, in: Proceedings of Interspeech, Brisbane, Australia.

Kim, Y., Kent, R.D., Weismer, G., 2011. An Acoustic Study of the Relationships Among Neurologic Disease, Dysarthria Type, and Severity of Dysarthria. Journal of Speech, Language, and Hearing Research 54, 417–429.

Kucera, H., Francis, W.N., 1967. Computational Analysis of Present-Day American English. Brown University, Providence, RI.

Kuhn, R., Junqua, J-C., Nguyen, P., Niedzelski, N., 2000. Rapid speaker adaptation in eigenvoice space. IEEE Transactions on Speech and Audio Processing 8, 695–707.

Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language 9, 171–185.

Menéndez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzio, J.E., Bunnell, H.T., 1996. The Nemours database of Dysarthric Speech, in: Proceedings of the Fourth International Conference on Spoken Language Processing, pp. 1962–1965.

Parker, M., Cunningham, S., Enderby, P., Hawley, M., Green, P., 2006. Automatic speech recognition and training for severely dysarthric users of assistive technology: the STARDUST project. Clinical Linguistics & Phonetics 20, 149–156.

Povey, D., Chu, S.M., Varadarajan, B., 2008. Universal Background Model Based Speech Recognition, in: Proceedings ICASSP

Raghavendra, P., Rosengren, E., Hunnicutt, S., 2001. An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. AAC: Augmentative and Alternative Communication 17, 265–275.

Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19–41.

Rudzicz, F., 2007. 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. p. 256.

Sharma, H., Hasegawa-Johnson, M., 2010. 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, Association for Computational Linguistics, Los Angeles, CA. pp. 72–79.

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

Stevens, K.N., 1999. Acoustic Phonetics. MIT Press, Cambridge, MA.

Steidl, S., Stemmer, G., Hacker, C., Noth, E., and Niemann, H., 2003. Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer, in: Proceedings of DAGM-Symposium 2003.

Tang, H., Chu, S., Hasegawa-Johnson, M., Huang, T., 2009. Emotion Recognition from Speech via Boosted Gaussian Mixture Models, in: Proceedings International Conference on Multimedia and Expo (ICME).

Tang, H., Hasegawa-Johnson, M., Huang, T., 2010. Non-Frontal View Facial Expression Recognition, in: Proceedings International Conference on Multimedia and Expo (ICME), 1202–1207.

Weismer, G., 2006. Philosophy of research in motor speech disorders. Clinical Linguistics & Phonetics 20, 315–349.

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

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1, 80–83.

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

Zhou, X., Navrátil, J., Pelecanos, J.W., Ramaswamy, G.N., Huang, T.S., 2008. Intersession Variability Compensation for Language Detection, in: Proceedings ICASSP.

Zhou, X., Zhuang, X., Tang, H., Hasegawa-Johnson, M., Huang, T., 2010. Novel Gaussianized Vector Representation for Improved Natural Scene Categorization. Pattern Recognition Letters 31, 702–708.

Zhuang, X., Zhou, X., Hasegawa-Johnson, M., Huang, T., 2008. Face Age Estimation Using Patch-based Hidden Markov Model Supervectors, in: Proceedings International Conference on Pattern Recognition (ICPR), 10.1.1.139.846:1-4.

Zhuang, X., Zhou, X., Hasegawa-Johnson, M., Huang, T., 2009. Efficient Object Localization with Gaussianized Vector Representation, in: International Multimedia for Consumer Electronics (IMCE), 83–96.

Zhuang, X., Huang, J., Potamianos, G., Hasegawa-Johnson, M., 2009. Acoustic Fall Detection Using Gaussian Mixture Models and GMM Supervectors, in: Proceedings of ICASSP, 69–72.

Zhuang, X., Zhou, X., Hasegawa-Johnson, M., Huang, T., 2010. Real-World Acoustic Event Detection. Pattern Recognition Letters 31, 1543–1551.

Table 2: *Speaker intelligibility and Recognition Accuracies.*

| **Speaker** | **Average** | **SI-MAP** | **BI-MAP** | $\alpha = 0.05$ | $\alpha = 0.10$ |
|---|---|---|---|---|---|
| | **Intell. (%)** | **WRA (%)** | **WRA (%)** | | |
| 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** | 🟩 | 🟧 |