State-Transition Interpolation and MAP Adaptation for HMM-based Dysarthric Speech Recognition

Harsh Vardhan Sharma
Beckman Institute
405 North Mathews Avenue
Urbana, IL 61801, USA
hsharma@illinois.edu

Mark Hasegawa-Johnson
Beckman Institute
405 North Mathews Avenue
Urbana, IL 61801, USA
jhasegaw@illinois.edu

Abstract

This paper describes the results of our experiments in building speaker-adaptive recognizers for talkers with spastic dysarthria. We study two modifications – (a) MAP adaptation of speaker-independent systems trained on normal speech and, (b) using a transition probability matrix that is a linear interpolation between fully ergodic and (exclusively) left-to-right structures, for both speaker-dependent and speaker-adapted systems. The experiments indicate that (1) for speaker-dependent systems, left-to-right HMMs have lower word error rate than transition-interpolated HMMs, (2) adapting all parameters other than transition probabilities results in the highest recognition accuracy compared to adapting any subset of these parameters or adapting all parameters including transition probabilities, (3) performing both transition-interpolation and adaptation gives higher word error rate than performing adaptation alone and, (4) dysarthria severity is not a sufficient indicator of the relative performance of speaker-dependent and speaker-adapted systems.

1 Introduction

After more than two decades of research, speech recognition 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.

Automatic speech recognition (ASR) systems generally assume that the speech signal is a realisation of some message encoded as a sequence of one or more symbols. To effect the reverse operation of recognising the underlying symbol sequence given a spoken utterance, the continuous speech waveform is first converted to a sequence of equally spaced discrete parameter vectors. The role of the recogniser is to effect a mapping between sequences of speech vectors and the wanted underlying symbol sequences. Most speech recognizers today are based on the *hidden Markov model* (HMM) paradigm: it is assumed that the sequence of observed speech vectors is generated by a Markov model as shown in Fig. 1. A Markov model is a finite state machine which changes state once every time unit and each time t that a state j is entered, a speech vector $\mathbf{o}_t$ is generated from the probability density $b_j(\mathbf{o}_t)$ which

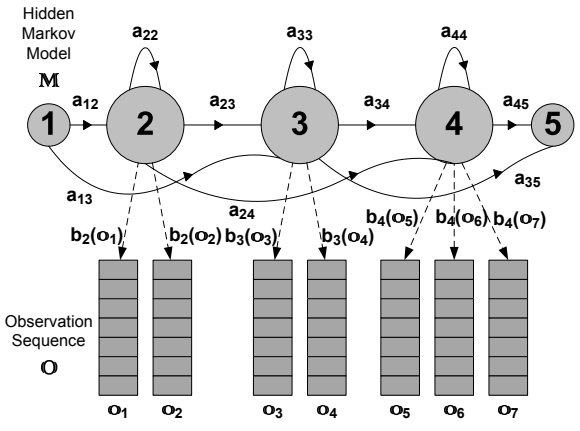

Figure 1: The Markov generation model.

Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies, pages 72–79,
Los Angeles, California, June 2010.

is a mixture-Gaussian density for most standard systems. The transition from state i to state j is also probabilistic and is governed by the discrete probability a_{ij}. Fig. 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 $\mathbf{o}_1$ to $\mathbf{o}_7$. 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.

Fig. 2 shows how HMMs can be used for isolated word recognition. Firstly, an HMM is trained for each vocabulary word using a number of examples of that word – given a set of training examples corresponding to a particular model, the parameters of that model ($\{a_{ij}\}$ and $\{b_j(\mathbf{o}_t)\}$) are determined by a robust and efficient re-estimation procedure. In this example, the vocabulary consists of just three words: "one", "two" and "three". Secondly, to recognise some unknown word, the likelihood (probability) of each model generating that word is calculated and the most likely model identifies the word.

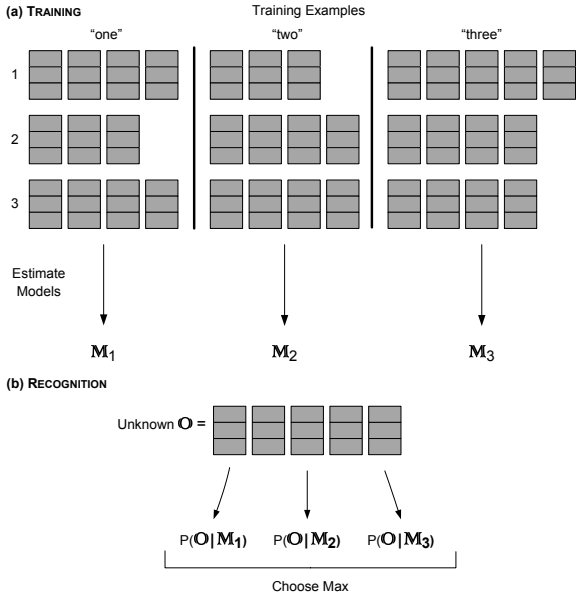

Figure 2: Using HMMs for isolated word recognition.

For creating a speech recognizer for a particular speaker, there are two approaches: one is to create a speaker-dependent (SD) system by utlizing speech of that speaker alone to train the HMMs; the other is to create a speaker-adapted (SA) system by first training the HMMs in a speaker-independent fashion by utlizing speech of several speakers, and then customising the HMMs to the characteristics of the particular speaker by using training examples of their speech to modify the HMM parameters. The parameter values do not get overwritten; they are adjusted using a regularized or constrained machine learning algorithm. Regularization (e.g., using Maximum A Posteriori learning) or constraints (e.g., using linear transformations) 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. Dysarthria is often a symptom of a gross motor disorder, whose other symptoms usually 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 instead of a keyboard (Carlson and Bernstein, 1987; Coleman and Meyers, 1991; Deller et al., 1988; Deller et al., 1991; Fried-Oken, 1985). 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 (Polur and Miller, 2005a; Polur and Miller, 2005b). They found that an ergodic HMM with a slight left-to-right character (called a *transition-interpolated* HMM from hereon) provides higher word recognition accuracy (WRA) than a standard left-to-right HMM, apparently because the transition-interpolated HMM is able to capture outlier events 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 (Deller et al., 1991). Section 2.1.2 explains

in more detail the difference between these HMM topologies.

Speaking for long periods of time is tiring, especially for a person with dysarthria, therefore it is difficult for a person with dysarthria to train a speaker-dependent ASR. Speaker adaptation then seems a useful method to overcome this obstacle in developing dysarthric speech recognizers. Raghavendra et al. (Raghavendra et al., 2001) have compared recognition accuracies of an SA system and an SD system. They found that the SA system adapted well to the speech of talkers with mild or moderate dysarthria, but the recognition scores were lower than 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 (Rudzicz, 2007) who compared the performance of SD and "SA" systems on the Nemours database (Menendez-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 speaker-independent 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.

The study described in this paper 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 SA 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 an SA 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 talker's dysarthria severity.

2 Experimental Setup

2.1 Modifications investigated

The following modifications to the HMM structure were studied in our experiments:

2.1.1 Adaptation

All SA systems were developed by adapting a speaker-independent system in a *Maximum A Posteriori* (MAP) manner, as outlined by Gauvain and Lee (Gauvain and Lee, 1991; Gauvain and Lee, 1992). MAP adaptation involves the use of prior knowledge about the model parameter distribution. Hence, if we know what the parameters of the model are likely to be (before observing any adaptation data) using the prior knowledge, we might well be able to make good use of the limited adaptation data, to obtain a decent MAP estimate. For MAP adaptation purposes, the informative priors that are generally used are the speaker independent model parameters (empirical Bayes approach). In (Gauvain and Lee, 1991), they derive expressions of MAP estimates for all HMM parameters except the transition probabilities (Gaussian mixture-component means, diagonal Gaussian mixture-component covariance matrices and, mixture-component weights) and also provide an initialization scheme for the prior density of these parameters. In (Gauvain and Lee, 1992), they derive expressions for MAP estimates of transition probabilities in addition to those for full-covariance Gaussian mixture-component parameters, and provide a MAP variant of the Expectation-Maximization (EM) re-estimation algorithm. All systems developed in our study modeled the observations as mixture of Gaussians with diagonal covariance matrices.

2.1.2 Transition-Interpolation

Fig. 3 illustrates the topologies of strictly left-to-right (LR) and transition-interpolated (TI) HMMs with 3 emitting states. If $\mathbf{A} = \{a_{ij}\}$ be the N $\times$ 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.

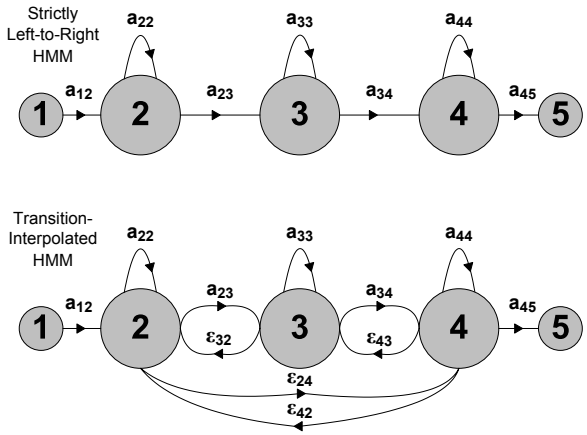

Figure 3: Difference between strictly left-to-right and transition-interpolated HMM topologies.

The TI model is an LR model which has non-zero transition probabilties for jumps and transitions to preceding states from a particular state (for emitting states). These probabilties are however small compared to self-transition and next-state–transition probabilties. 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 $\sum_{j=1}^{N} a_{ij} = 1$. After this initialization, the transition probability matrix is re-estimated for speaker-dependent systems using the standard Maximum Likelihood EM algorithm, and for speaker-adapted systems using the MAP variant of the EM algorithm.

2.2 Data used

The experiments described in this paper utilized speech of 7 speakers from the UA-Speech database (Kim et al., 2008). This corpus was constructed with the aim of developing large-vocabulary dysarthric ASR systems which would allow users to enter unlimited text into a computer. All speakers exhibited symptoms of spastic dysarthria, according to an informal evaluation by a certified speech-language pathologist. Each speaker recorded 765 isolated words in 3 blocks of 255 words each; (a) common to all blocks: 10 digits (D), 19 computer commands (C), 26 radio alphabet letters (L), and 100 common words (CW) selected from the Brown corpus of written English; and (b) unique to each block: 100 uncommon words (UW) selected from children's novels digitized by Project Gutenberg. Vocabularies D and CW were primarily composed of monosyllables, C and L of bisyllables, and UW of polysyllabic words. The speakers' speech was affected by dysarthria associated with cerebral palsy. Data acquisition and intelligibility assessment is described in more detail 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 CW and UW 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 speakers intelligibility. Table 1 lists the speakers whose speech materials from the UA-Speech database were used, along with their human listener intelligibility ratings. The first letter of the speaker code ('M' or 'F') indicates their gender.

Speaker	**Age**	**Speech Intelligbility** (%)
M09	18	high (86%)
M05	21	mid (58%)
M06	18	low (39%)
F02	30	low (29%)
M07	58	low (28%)
F03	51	very low (6%)
M04	>18	very low (2%)

Table 1: Summary of Speaker Information (in decreasing order of human listener intelligibility rating).

For building the "MAP prior" speaker-independent system, the unadapted HMMs were trained on speech from the TIMIT corpus (Garofolo et al., 1993).

2.3 System Configurations

Table 2 lists the characteristics of the various system configurations that were studied: SD stands for speaker-dependent, SA for speaker-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. These systems were developed for each of the seven

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

Table 2: Summary of ASR System Configurations

speakers listed in Table 1, and employed word-internal, context-dependent triphone HMMs, with three hidden states and observations modeled as mixture-of-Gaussians. Configuration C00 was developed by Sharma and Hasegawa-Johnson (2009) and is the baseline configuration for the present experiments. For configurations C11 through C15, the speaker-independent systems trained on TIMIT employed left-to-right HMMs. For systems C15, the transition-interpolation was performed after obtaining the speaker-independent 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 speaker-independent 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 comprised of 12 Perceptual Linear Prediction coefficients (Hermansky, 1990) 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 measure used for assessing the performance of the developed recognizers is the fraction of task–vocabulary words correctly recognized (in percent), defined in Equation 1.

$$PWC = \frac{\#\ words\ correctly\ recognized}{\#\ words\ attempted} \times 100 \quad (1)$$

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): standard methods for choosing this number (using development test data) could not be employed on account of insufficient data. The results reported in the next section should therefore be interpreted as development test results. In order to avoid over-tuning, the number of Gaussian components was constrained to be the same across all speakers. For the speaker-dependent systems (C00 and C01), results are for HMMs with 2 Gaussian components per probability density. For the speaker-adapted systems (C11-C15), results are for HMMs with 32 Gaussian components per probability density: while training the speaker-independent 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.

3 Results

Tables 3, 4 list the PWC scores for the various system configurations developed. The speakers are listed in decreasing order of intelligibility rating. The scores for systems C00 are restated here from Sharma and Hasegawa-Johnson (2009) (Table 6, under the column 'T10').

We see that speaker-dependent systems with left-to-right HMMs (C00) have higher recognition accuracy than the speaker-dependent 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 speaker-dependent systems, for all speakers except F02 and M07 (worse than both speaker-dependent systems). System C13 with adaptation of all parameters ex-

	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 3: PWC scores for each speaker's configurations C00-C12.

	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

Table 4: PWC scores for each speaker's configurations C00,C13-C15.

cept transition-probabilities has the highest recognition accuracy for all subjects except F02 and M07 (highest among speaker-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 speaker-dependent 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 speaker-dependent systems whenever the corresponding C13 (and C14) system also performs better than them.

These results are plotted in Fig. 4 along with the human listeners' intelligibility ratings of these speakers (the black circles). For speakers M09 and M05, system C13 with the best overall PWC score is still far from doing as well as human listeners. For

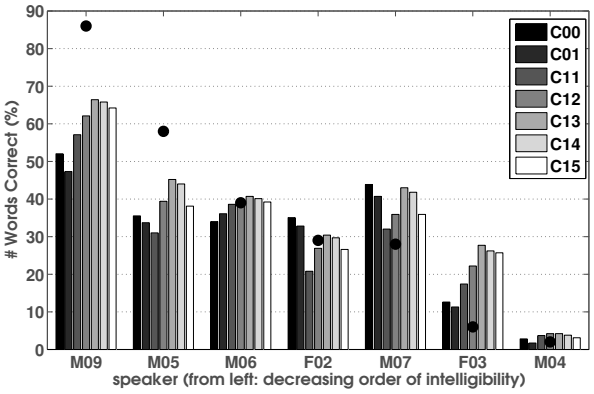

Figure 4: PWC scores for various system configurations (the black circles indicate speakers' human listener intelligibility ratings).

the remaining subjects, it has however been able to do as well or better than human listeners even when it performed worse than the corresponding speaker-dependent systems (C00,C01): in fact, for speaker M06, it does better than human listeners when the speaker-dependent systems don't.

Fig. 5 plots, for all speakers, the percentage difference PWC(x)/PWC(C00)-1 between the PWC of system x ($x \in \{C01 - C15\}$) and the PWC of system C00.

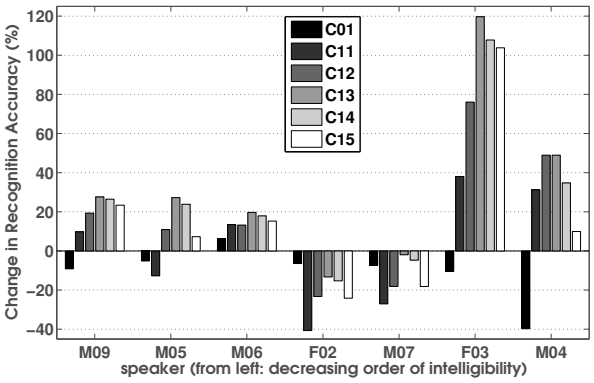

Figure 5: Percentage change in PWC scores for various system configurations relative to configuration C00's PWC score.

For speakers who have an intelligibility rating above 35% or below 25%, the speaker-adapted systems generally do better than their speaker-dependent counterparts. System C01, with transition interpolation, performs worse than system

C00 for all speakers except M06. The surprising result though is that for speakers with highly severe dysarthria (F03 and M04), speaker-adapted systems have substantially better recognition accuracies than their speaker-dependent counterparts, when previous studies have indicated that for such subjects, speaker-dependent systems perform better than speaker-adapted systems.

4 Conclusions

This study investigated adaptation and state-transition interpolation techniques for medium vocabulary HMM-based speech recognition of talkers with spastic dysarthria. It was found that performing transition-interpolation generally worsens recognition performance when compared to left-to-right HMMs. Performing both adaptation and transition-interpolation results in higher recognition accuracy compared to the speaker-dependent 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, speaker-adaptation was able to achieve substantial improvement in recognition accuracy, compared to the speaker-dependent systems. This finding is significant in that it is contrary to the conclusions of previously published studies. The results reported in this paper 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.

References

Gloria S. Carlson and Jared Bernstein. 1987. Speech Recognition of Impaired Speech. *Proceedings of RESNA 10th Annual Conference on Rehabilitation Technology*, 165–167.

Colette L. Coleman and Lawrence S. Meyers. 1991. Computer Recognition of the Speech of Adults with Cerebral Palsy and Dysarthria. *AAC: Augmentative and Alternative Communication*, 7(1):34–42.

John R. Deller, D. Frank Hsu and Linda J. Ferrier. 1988. Encouraging Results in the Automated Recognition of Cerebral Palsy Speech. *IEEE Transactions on Biomedical Engineering*, 35(3):218–220.

John R. Deller, D. Frank Hsu and Linda J. Ferrier. 1991. On the use of Hidden Markov modelling for Recognition of Dysarthric Speech. *Computer Methods and Programs in Biomedicine*, 35(2):125–139.

Melanie Fried-Oken. 1985. Voice Recognition Device as a Computer Interface for Motor and Speech Impaired People. *Archives of Physical Medicine and Rehabilitation*, 66:678–681.

Jean-luc Gauvain and Chin-hui Lee. 1991. Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models. *Proceedings of DARPA Speech and Natural Language Workshop*, 272–277.

Jean-luc Gauvain and Chin-hui Lee. 1992. MAP Estimation of Continuous Density HMM: Theory and Applications. *Proceedings of DARPA Speech and Natural Language Workshop*, 185–190.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren and Victor Zue. 1993. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. http://www.ldc.upenn.edu/Catalog/LDC93S1.html.

Hynek Hermansky. 1990. Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752.

Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin and Simone Frame. 2008. Dysarthric Speech Database for Universal Access Research. *Proceedings of Interspeech, Brisbane, Australia*, 22–26.

Xavier Menendez-Pidal, James B. Polikoff, Shirley M. Peters, Jennie E. Leonzio, H. T. Bunnell. 1996. The Nemours Database of Dysarthric Speech. *Proceedings of the Fourth International Conference on Spoken Language Processing, Philadelphia, PA, USA*.

Prasad D. Polur and Gerald E. Miller. 2005a. Effect of High-Frequency Spectral Components in Computer Recognition of Dysarthric Speech based on a Mel-Cepstral Stochastic Model. *Journal of Rehabilitation Research & Development*, 42(3):363–372.

Prasad D. Polur and Gerald E. Miller. 2005b. 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*, 13(4):558–561.

Parimala Raghavendra, Elisabet Rosengren and Sheri Hunnicutt. 2001. An Investigation of Different Degrees of Dysarthric Speech as Input to Speaker-Adaptive and Speaker-Dependent Recognition Sys-

tems. *AAC: Augmentative and Alternative Communication*, 17(4):265–275.

Frank Rudzicz. 2007. Comparing Speaker-Dependent and Speaker-Adaptive Acoustic Models for Recognizing Dysarthric Speech. *Proceedings of ASSETS'07, Tempe, AZ, USA*.

Harsh Vardhan Sharma and Mark Hasegawa-Johnson. 2009. Universal Access: Speech Recognition for Talkers with Spastic Dysarthria. *Proceedings of Interspeech, Brighton, UK*, 1451–1454.