Rapid Speaker Adaptation With Speaker Clustering

by

Ernest J. Pusateri

B.S., Carnegie Mellon University, 2000

Submitted to
the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Master of Science

at the

Massachusetts Institute of Technology

June 2002

Signature of Author .....
Department of Electrical Engineering and Computer Science
May 10, 2002

Certified by .....
Timothy J. Hazen
Research Scientist
Department of Electrical Engineering and Computer Science

Accepted by .....
Arthur C. Smith
Chair, Department Committee on Graduate Students

Rapid Speaker Adaptation With Speaker Clustering

by

Ernest J. Pusateri

Submitted to the Department of Electrical Engineering and Computer Science
in June, 2002 in partial fulfillment of the requirements for the Degree of
Master of Science

Abstract

This thesis addresses the problem of rapid speaker adaptation. This is the task of altering the parameters of a speaker dependent speech recognition system so as to make that system look more like a speaker dependent system using a very small amount (<10 seconds) of speaker specific data. The approach to speaker adaptation taken in this work is called *speaker cluster weighting* (SCW). SCW extends the ideas of previous speaker cluster techniques by allowing the speaker cluster models (learned from training data) to be adaptively weighted to match the current speaker. This algorithm involves three major steps: speaker clustering, cluster model training, and cluster weighting. We have explored strategies for use in each of these steps. For the first step, we have developed a novel algorithm called *least squares linear regression* clustering for the clustering of speakers for whom only a small amount of data is available. For the second step acoustic models were trained using two strategies. In the first, *model interpolation*, EM trained cluster acoustic models were interpolated with speaker independent models to create more robust cluster models. In the second, *weight adaptation*, the cluster models were trained by adapting the mixture component weights of the SI model. Finally, for the third step, two strategies for cluster weighting, one using the maximum likelihood criterion and one using the minimum classification error criterion, were applied. Experiments were also run where the maximum likelihood cluster model was chosen as the adapted model. Combining some of these strategies, a 10% relative improvement in WER was obtained for child speakers as well as a 9% improvement for female speakers after 1 adaptation utterance.

Keywords: speech recognition, speaker adaptation.

Thesis Supervisor: Timothy J. Hazen
Title: Research Scientist

Acknowledgments

First and foremost, I would like to thank my advisor, T.J. Hazen. Our many discussions have increased my knowledge of speech recognition immensely, and his willingness to listen to my ideas, even the bad ones, was greatly appreciated.

I would also like to thank Jim Glass, Victor Zue and the rest of the staff and students in the Spoken Language Systems group for creating an excellent environment for research and learning. I owe my officemates, Alicia, Han and Laura a special debt of gratitude, as they make it a pleasure to come to the office every day.

My friends and family need also to be acknowledged: my parents for their undying support throughout my life; my little sister for her inspirational fearlessness; my friends back in Pittsburgh, Aubree and Jason, for thinking way too highly of me; and my friends here, Tim, Doug, Mike, Laura, Marc and Matt, for filling my life outside of MIT.

This research was supported by a contract from NTT and by DARPA under Contract N66001-99-1-8904 monitored through the Naval Command, Control, and Ocean Surveillance Center.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Definition

A couple of decades ago a speech recognition system had to be speaker dependent (SD) in order to be usable. In other words, the system's acoustic models had to be trained specifically for the speaker that would be using it. This required a lengthy enrollment process for each specific user of the system, and after the system was trained, any new users would have to retrain in order to obtain acceptable performance.

Through various advances in speech recognition technology, speaker independent (SI) systems can now perform sufficiently well for most tasks. Still, a large performance gap exists between SD and SI systems. For example, SD systems can achieve word error rates that are 50% lower than those achieved by SI recognizers [19]. The goal of speaker adaptation is to help close this gap.

Speaker adaptation techniques involve using utterances from a specific speaker, called adaptation data, to change the parameters of an SI system. The parameters are changed such that the new system more closely resembles an SD system trained for that speaker. Seen in another light, this task is one of using correlations between the acoustic characteristics of the speaker's spoken utterances and the expected acoustic characteristics of future utterances to better model the future utterances. Many techniques have been developed to accomplish this task, the most popular of which will be described in the next section.

A speaker adaptation task is often described according to two parameters. The first describes the amount of adaptation data available. If very little adaptation data is available (less than 10s), adaptation is referred to as *rapid.* The second parameter is whether or not an accurate transcription of the adaptation data is available. If it is, adaptation is referred to as *supervised.* Otherwise, it is *unsupervised.*

This work addresses the problem of rapid, unsupervised adaptation. The fact

that very little training data is available makes the task more difficult, as does the lack of an accurate transcription for the adaptation data. Various algorithms have been developed for the general problem of speaker adaptation, and many of these have been extended specifically to make them applicable to rapid, unsupervised adaptation. These approaches and their extensions are described in the next section.

1.2 Previous Work

1.2.1 MAP-based Approaches

The problem of speaker adaptation is one of finding the most likely set of acoustic model parameters given the adaptation data. That is, given sets of observations for c classes, $X_0, X_1...X_c$ we would like to choose the parameter set λ_i for each class, i, such that:

$$\lambda_i = \arg\max_{\lambda_i} \mathrm{p}(\lambda_i|X_0, X_1...X_c) \tag{1.1}$$

In both maximum-likelihood (ML) and maximum *a posteriori* (MAP) acoustic model training, it is assumed that the parameters for a particular acoustic model, i, depend only on the observations from that class:

$$\mathrm{p}(\lambda_i|X_0, X_1, X_2...X_c) \approx \mathrm{p}(\lambda_i|X_i) \tag{1.2}$$

Using Bayes rule, we can say:

$$\mathrm{p}(\lambda_i|X_i) = \frac{\mathrm{p}(X_i|\lambda_i)\mathrm{p}(\lambda_i)}{\mathrm{p}(X_i)} \tag{1.3}$$

Because $\mathrm{p}(X_i)$ is independent of λ_i, it can be ignored in the maximization. This leaves us with:

$$\mathrm{p}(\lambda_i|X_i) = \arg\max_{\lambda_i} \mathrm{p}(X_i|\lambda_i)\mathrm{p}(\lambda_i) \tag{1.4}$$

While maximum-likelihood approaches assume that $\mathrm{p}(\lambda_i)$ is uniform over all parameter sets, λ_i, MAP estimation does not. This requires estimation of a probability distribution for the model parameters, $p(\lambda_i)$ for each class i. When X_i is a small amount of data, $p(\lambda_i)$ makes the estimated parameters more robust.

As can be seen from the assumption in Equation 1.2, MAP adaptation only allows for adaptation of acoustic models representing classes for which adaptation data has been seen [14]. This makes adaptation with this approach very slow (i.e. many adaptation utterances are required) and, thus, unsuitable for rapid adaptation. However,

MAP does have the desirable property that it will eventually converge to the ML solution.

Various extensions of MAP have been developed for use in rapid adaptation. In one extension, the regression based model prediction (RMP) approach [1], correlations between Gaussian component means are determined from the training data. During adaptation, these correlations are used to adapt multiple component means at once. Ahadi and Woodland reported an 8% improvement in WER after one adaptation utterance on the 1000 word Resource Management task [1].

Another extension, called structural MAP (SMAP) [29], works by organizing Gaussian components into a tree. Adaptation occurs by starting at the root of the tree and adapting each component, using it's parent for the prior parameter distribution. Like RMP, this results in significant WER reduction after only one adaptation utterance.

1.2.2 Transformation-Based Approaches

Maximum Likelihood Linear Regression

In its original form, maximum likelihood linear regression (MLLR) consists of adapting the means of a Gaussian mixture model using an affine transformation, as shown here [26]:

$$\mu_i^{'} = \mathbf{A}_g \mu_i + b_g \tag{1.5}$$

This estimation is generally done using the EM algorithm. The same transformation can be used for all Gaussian components across all acoustic models, i, or different transformations can be used for different groups of Gaussian components, called regression classes. Methods have been proposed for choosing the groups in an optimal manner [25]. MLLR can also be applied to the Gaussian covariance matrices.

In its basic form, MLLR has an advantage over MAP in that adaptation occurs on acoustic models for which data has not been seen. The assumption in Equation 1.2 is replaced with the assumption that a particular model's parameters, λ_i, depend on all the observations within the same regression class. The set of model parameters over which the maximization is done is limited to affine transformations of the speaker independent model parameters. This allows MLLR to achieve about a 15% reduction in WER after about 1 minute of adaptation data [26]. MLLR will converge to the ML solution if the number of regression classes is equal to the total number of Gaussian components. MLLR does not, however, use any speaker correlation information from the training corpus.

Although MLLR requires less adaptation data than MAP, 1 minute of data is still more than is available in most spoken dialogue systems (e.g. Jupiter[17].) Various attempts have been made to make MLLR more robust in the presence of smaller

amounts of adaptation data. These include incorporating prior distributions of MLLR parameters [5, 13], eigenspace based techniques [32], and a combination of the two [4, 31]. Zhou and Hanson [32] achieved relative error rate reductions of 10.5% using supervised adaptation on one adaptation utterance, while smaller reductions were seen by Wang, *et al* [31].

Vocal Tract Length Normalization

Another form of transformation based adaptation is vocal tract length normalization (VTLN). In VTLN, an estimate is made of the length of a speaker's vocal tract. This estimate is used to warp the frequency axis during feature calculation in such a way as to normalize the speech to a standard vocal tract length. The optimal warping is often chosen using the maximum likelihood criterion [24]. Claes, *et al*, instead of normalizing the acoustic features, normalizes the acoustic models to account for different vocal tract lengths [7].

1.2.3 Cluster-Based Approaches

Cluster Adaptive Training

In cluster adaptive training (CAT), the adapted model parameters are assumed to be a linear combination of model parameters from different speaker clusters [12]. In one version, SI model component weights and variances are held constant, and the adapted means are a linear combination of the cluster means. In another, a transformation (specifically MLLR in [12]) is associated with each cluster, and the adapted models are created using a linear combination of the cluster transformations. The cluster weights are computed to maximize the likelihood of the adaptation data. This method has been shown to yield relative improvements of about 7% on a large vocabulary task using one adaptation utterance.

Eigenvoices

Like CAT, the eigenvoice approach is based on creating an adapted model through a linear combination of basis model parameter sets [23]. However, unlike in CAT, when using the eigenvoice technique these parameter sets are determined through principal component analysis (PCA). Supervectors are formed by concatenating speaker mean vectors, and "eigenvoices" are determined by performing PCA on these vectors. The eigenvalues allow the eigenvoices to be ordered in terms of how much of the inter-speaker variance they represent. Thus, one can systematically choose larger and larger sets of eigenvoices as the amount of adaptation data increases.

Unfortunately, in a large vocabulary system with a large number of acoustic models, the supervectors become extremely large and PCA becomes computationally intractable. Nguyen, Wellekens and Junqa developed a less computationally intensive method for finding the eigenvoices using an ML criterion [27]. Still, their work used only 48 context independent acoustic models. Previously mentioned applications of the eigenvoice concept to MLLR have, so far, achieved greater success for large vocabulary rapid adaptation[32, 4, 31].

1.3 Goals and Overview

The goal of this work was to develop an approach for rapid speaker adaptation with the hope of meeting or exceeding the success of other adaptation methods. As the last section revealed, this translates into achieving about a 10% relative reduction in WER.

The most successful approaches to rapid speaker adaptation use intra-speaker inter-phone correlation information from the corpus. In the case of RMP, this information is embedded in the correlations between Gaussian component means, while in SMAP it is embodied in the component tree. In CAT and eigenvoices, the speaker clusters contain the speaker correlation information.

In this work, too, speaker correlation information was obtained from the training corpus in order to aid in rapid speaker adaptation. As in CAT and Eigenvoices, this information was embedded in speaker clusters. The approach taken in this work was speaker cluster weighting, first presented in [19].

This thesis is organized as follows. In Chapter 2, relevant details of the SUMMIT recognizer and Jupiter corpus are presented. Chapter 3 describes and evaluates the clustering algorithms explored. The speaker cluster weighting algorithm is presented in Chapter 4 as well as details of the strategies utilized for each stage in the algorithm and results. Finally, a summary and a description of possible future work are provided in Chapter 5.

Chapter 2

The SUMMIT Recognizer and Jupiter Corpus

2.1 SUMMIT Recognizer

This work utilized the SUMMIT segment-based speech recognizer[16]. As SUMMIT is a segment-based speech recognition system, the speech signal is first divided into segments of relative acoustic regularity. These segments are then classified phonetically using probabilistic acoustic models. A lexicon and language model are applied to the resulting labelling in order to obtain sentence hypotheses. The details of this process along with the specific recognizer configuration used for this work follow.

2.1.1 Segmentation

Segmentation in SUMMIT is accomplished either by looking for points of large acoustic change [15, 24] or through a probabilistic approach [3, 24]. The former is used in this work. The result of segmentation is a set of *landmarks* positioned throughout the utterance. These landmarks are placed both on the boundaries between phones as well as in places internal to phones.

One can model the signal characteristics between these landmarks as well as the characteristics of the acoustic change at these landmarks. In the first case, the models are referred to as *segment* models, while in the second case they are referred to as *boundary* models. Although SUMMIT is capable of using both of these types of models simultaneously, in this work only boundary models are utilized.

2.1.2 Acoustic Modeling

Features

As opposed to frame-based approaches where features are extracted at evenly spaced frames, in SUMMIT a set of features is extracted for each hypothesized segment or boundary. In this work, initially 112 features were extracted from each hypothesized boundary. These features consisted of 14 Mel-frequency cepstral coefficients averaged over 8 different regions near the boundary.

Principal component analysis was used to make the global covariance of the data equal to the identity matrix [2]. The 50 dimensions with the largest eigenvalues were used as the final acoustic features.

Acoustic Classes

In general, SUMMIT is capable of using any predefined set of acoustic classes. For this work, the classes were determined through an automatic, decision tree clustering process. The process began with a set of 2971 boundary models created from a 58 phone set. These models consisted of two types: those representing acoustic boundaries between two phones (denoted $\mathrm{t}(p_1|p_2)$) and those representing acoustic boundaries within a single phone (denoted $\mathrm{i}(p_1)$). These were clustered into 1357 classes used for recognition.

Model Structure and Training

The acoustic models used in this work were mixtures of Gaussian components. Thus, the probability of an observation, $\vec{x}$, given an acoustic class, p, can be expressed as:

$$\mathrm{p}(\vec{x}|p) = \vec{c} \cdot \vec{p}(\vec{x}|p) \tag{2.1}$$

Where $\vec{c}$ denotes the vector of Gaussian component weights:

$$\vec{c} = \begin{bmatrix} c_0 \\ \vdots \\ c_N \end{bmatrix} \tag{2.2}$$

The vector of component distributions is represented by $\vec{p}(x|p)$:

$$\vec{p}(\vec{x}|p) = \begin{bmatrix} \mathrm{p}_0(\vec{x}|p) \\ \vdots \\ \mathrm{p}_N(\vec{x}|p) \end{bmatrix} \tag{2.3}$$

Each component has a Gaussian distribution:

$$\mathrm{p}_i(\vec{x}|p) \sim \mathrm{N}(\mu_i, \sigma_i^2) \tag{2.4}$$

The number of Gaussian components used for a particular model, N was dependent on the amount of training data available for that class. Specifically, it was determined using the following formula, where T is the number of training tokens:

$$N = min(50, \frac{T}{50}) \tag{2.5}$$

Models were trained to maximize the likelihood of the data. This was completed using the Expectation-Maximization (EM) algorithm [9].

2.1.3 Lexical and Language Modeling

In SUMMIT, lexical modeling is accomplished through the assignment of multiple pronunciations for each word in the lexicon. First, each word is given a small number of canonical phonemic representations. Then, a set of phonological rules is applied that further expands the number of acceptable pronunciations. The result of applying the lexicon to the labeled boundaries is a word network. To this, a language model is applied, yielding the best (N-best) hypothesis (hypotheses). In this work, a lexicon of 1893 words was used with a class-based, tri-gram language model.

2.2 Jupiter Corpus

The experiments utilize data collected from the Jupiter domain [17]. Jupiter is a telephone-based spoken dialogue system for weather information, where users are can ask multiple questions during one call. Utterances average about 3 seconds in length. The Jupiter system does not keep track of callers' identities, so each call is treated as if it came from a different speaker, although many people call the system more than once. The collected data used in this thesis was divided into 3 mutually exclusive sets: the training set, the development set and the test set.

2.2.1 Training Set

The training set used in this work consisted of 17,116 calls. Each call had an average of 4.7 utterances per call giving the training corpus a total of 80,487 utterances. 13,030 calls contained male speakers, 3,297 calls involved female speakers and 1206 calls involved child speakers. (Note that one call could contain more than one speaker.)

The gender distribution of the utterances in the training set was 73% male, 18% female, and 9% child.

2.2.2 Development Set

The development set contained 275 male calls consisting of 1299 utterances, 84 female calls consisting of 274 utterances and 103 child calls consisting of 244 utterances.

2.2.3 Test Set

The test set was comprised of only calls with at least 6 utterances. It contained 105 male calls consisting of 1282 utterances, 24 female calls consisting of 285 utterances, and 31 child calls consisting of 367 utterances.

Chapter 3

Speaker Clustering

A key step in the adaptation algorithm used in this work is the clustering of the training speakers. In this chapter, the general clustering approach is described, as well as the distance metrics used with this approach. The quality of the clusters produced using different distance metrics is evaluated.

3.1 Cluster Requirements

The goal of all speaker adaptation approaches is to learn something about the speaker, then use that knowledge to improve recognition accuracy. In adaptation approaches based on speaker clusters, this knowledge takes the form of intra-speaker, inter-class correlation information, and is captured in the models built from the speaker clusters.

This means that speakers within each cluster should be acoustically similar across a large number of phones. If this is the case, the acoustic models built from the cluster data will be more focussed than the SI model. This is important if adaptation is going to increase the discriminative power of the acoustic models.

In order for a cluster's acoustic models to be robust, it is also required that the cluster have a sufficient number of speakers. This means that clusters must be reasonably balanced in terms of numbers of speakers.

3.2 Clustering Process

Ideally, all of the speakers in the corpus would have been clustered using a bottom-up approach. However, the size of the corpus makes this impractical. Instead a combination of bottom-up and k-means clustering was utilized.

The process consisted of three steps. First, a distance matrix was generated for all speakers. This was done using one of the distance metrics described in the next

section. The distance between each speaker and every other speaker was calculated and stored in this matrix.

Second, bottom up clustering was performed on the first 5000 speakers using the information in the distance matrix. At each iteration of the bottom up algorithm, the two clusters with the smallest distance between them were merged. The distance between two speakers was defined as follows: The distances for all pairs of points such that the points do not belong to the same cluster were calculated. The distance between the most distant pair was used as the initial distance between the two clusters. This distance is represented as $\mathrm{b}(c_i, c_j)$.

Scaling was done in order to decrease the likelihood that very small clusters would result. The final distance between two clusters was computed as follows:

$$\mathrm{d}(c_i, c_j) = \frac{N_i + N_j}{N_i N_j} \mathrm{b}(c_i, c_j) \tag{3.1}$$

Here N_i and N_j represent the number of speakers in clusters i and j respectively.

In the third step of the clustering process, the remaining speakers are put into the nearest cluster using the same distance measure as was described above. The fourth step was completed only when using the feature mean distance and LSLR characteristic distance metrics. In this step, k-means clustering was completed using the clusters resulting from the previous step as seeds.

3.3 Distance Metrics

The distance metric used in the clustering algorithm plays a large part in determining the characteristics of the resulting clusters. The distance metrics described here are designed such that clusters built using them will meet the cluster requirements described in Section 3.1. As it has been reported that successful clustering procedures separate male and female speakers, these metrics are evaluated based on their ability to separate speakers of different genders [22]. The Jupiter training corpus, as described in Section 2.2, was used for these evaluations.

3.3.1 Gender

Description

Using this metric, speakers are manually clustered according to their gender. For this work, three genders are defined: male, female, and child. This simple metric has proven very useful for speaker adaptation [19]. Its usefulness can be attributed to the presence of strong within-gender acoustic correlation, largely due to the fact that speakers of the same gender often share similar vocal tract lengths [11, 21].

Gender-based metrics are especially useful on corpra such as the one to be used in this research, where very little data is available for each speaker. Metrics based on an acoustic similarity measure often fail under these conditions. However, using gender as the distance metric limits the number of clusters to three.

3.3.2 Feature Mean Distance

Description

One straightforward measure of acoustic similarity is the feature mean distance. For each speaker a mean feature vector is computed by taking the average value of the PCA normalized features across all acoustic observations for a particular speaker. This calculation is shown in equation 3.12, where $\vec{v}^{(s)}$ is the mean feature vector for speaker s, $\vec{o}_{ij}^{(s)}$is the jth feature vector for the ith class for speaker s, $n_i^{(s)}$ is the number of feature vectors for class i for speaker s, and C is the total number of classes.

$$\vec{v}^{(s)} = \frac{\sum_{i=1}^{C} \sum_{j=1}^{n_i^{(s)}} \vec{o}_{ij}^{(s)}}{\sum_{i=1}^{C} n_i^{(s)}} \tag{3.2}$$

Distances between speakers are calculated by computing the Euclidean distance between these averages:

$$d_{st} = \sqrt{\vec{v}^{(s)} \cdot \vec{v}^{(t)}} \tag{3.3}$$

This method may work well when a large amount of data is present for each speaker. However, in the corpus used in this work, speakers are represented by only a small number of utterances, resulting relatively little phonetic diversity. This lack of diversity may result in poor estimates for the speaker feature means and hurt the performance of this clustering approach.

Evaluation

Cluster	Male (%)	Female (%)	Child (%)
1	38	45	17
2	76	18	6
3	89	9	6
4	76	14	10
5	81	14	5

Table 3.1: Gender distribution for 5 clusters created using the feature mean distance.

Table 3.1 shows the gender distribution for 5 clusters created from the data in the training set using the feature mean distance metric. We see that some separation of the genders is present. Cluster 1 contains a large number of females and children compared to the others, while cluster 3 contains relatively few. The other clusters, however, are fairly mixed, indicating that this distance metric may not be capturing the necessary speaker correlation information.

The failure of this clustering method in this context can probably be explained by the limited data available for each speaker. The speaker feature means may have been more affected by the phonetic content of each speaker's data than by the general acoustic characteristics of each speaker.

3.3.3 Average Class Feature Mean Distance

Description

The average phoneme feature mean distance is based on the feature mean distance. However, instead of one mean feature vector being computed for each speaker, a set of mean feature vectors corresponding to each class is computed. While in the case of the feature mean distance, a global principal component transformation was performed, here the transformation is class specific. The transformation is also constructed such that the variance for each dimension after transformation is 1. This ensures that inter-speaker Euclidean distances between averages from the same class will be on the same scale, regardless of the class chosen.

Distances between speakers are calculated by finding the average Euclidean distance between the mean normalized feature vectors corresponding to the same classes in each speaker. Calculation of the mean normalized feature vectors is shown here:

$$\vec{\mu}_i^{(s)} = \sum_{j=1}^{n_i^{(s)}} \vec{o}_{ij}^{(s)} \tag{3.4}$$

To calculate the distance between two speakers, the set of classes, P, for which both speakers have data is determined:

$$\mathcal{P}_{st} = \{i | \exists \vec{\mu}_i^{(s)}\} \cap \{i | \exists \vec{\mu}_i^{(t)}\} \tag{3.5}$$

The average distance between these classes is then calculated:

$$d_{st} = \frac{1}{num(\mathcal{P}_{st})} \sum_{\forall i \epsilon \mathcal{P}_{st}} \vec{\mu}_i^{(s)} \cdot \mu_i^{(t)} \tag{3.6}$$

This distance measure has the potential to overcome the phonetic diversity problem encountered when using the feature mean distance. However, this distance measure may not prove robust for speakers whose data has very few common classes.

Evaluation

Cluster	Male (%)	Female (%)	Child (%)
1	74	21	5
2	77	18	5
3	72	23	5
4	77	16	7
5	74	19	8

Table 3.2: Gender distribution for 5 clusters created using the average class feature mean distance.

Table 3.2 shows the gender distribution obtained when clustering using the average class feature mean distance. We see that all of the clusters have very similar gender distributions. This is probably, in part, because, with the small amount of data available for each speaker, the speakers share very few phones. The nature of this distance metric is such that if two speakers do not both have examples of a significant number of classes, the distance measurement between those speakers will not be robust.

Another possible reason for the failure of this algorithm deals with the variability of the within speaker variance across different classes. The feature vectors are normalized to account for differences in feature variance across all feature vectors for a particular class. However, they are not normalized to account for within-speaker variance. Figure 3.1 illustrates this. First notice that, if we look at all of the data points across both speakers, the variance across each dimension could feasibly be the same for classes A and B. If we looked at a random sample of feature vectors from class A for each speaker, it is very possible the speakers would look acoustically similar. However, looking at class B, this is clearly not the case. Thus, differences in feature means for class A are not as meaningful as the differences in feature means for class B. The average feature mean distance does not take this into account.

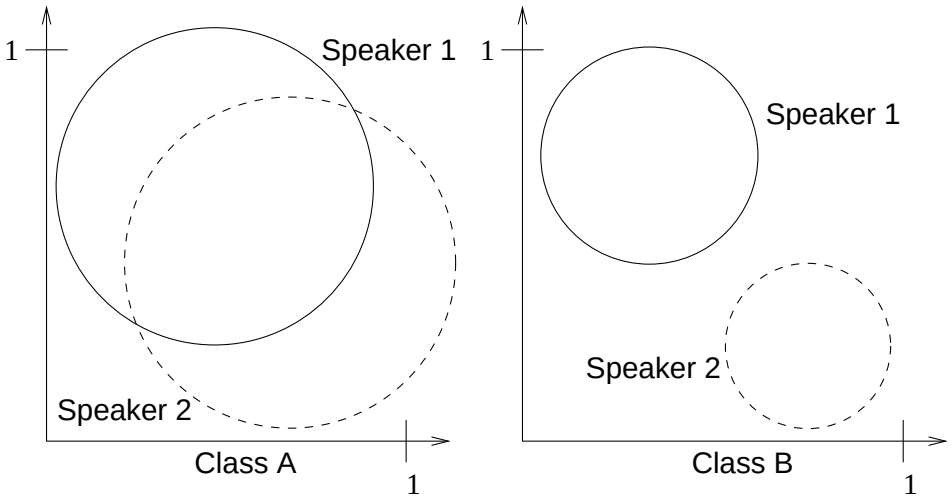

Figure 3.1: Illustration of one possible deficiency of the average feature mean distance.

3.3.4 LSLR Characteristic Distance

Description

The LSLR characteristic was designed to overcome the shortcomings of the feature mean distance and average class feature mean distance. Unlike the former, it is reasonably insensitive to the phonetic content of a particular speaker's data, and, unlike the latter, it does not require that two speakers share a large number of phones.

Using this approach, a characteristic vector is computed for each speaker. The vector is computed in two steps. First, an affine transformation is applied to the observed feature vectors of a selected set of classes to transform them into a generic class space. Second, the transformed features from all observations from the selected set of classes from one speaker are averaged to create that speaker's characteristic vector. This computation for speaker s is shown here:

$$\vec{v}^{(s)} = \frac{1}{(\sum_{\forall k \epsilon \mathcal{C}} n_k^{(s)})} \sum_{\forall i \epsilon \mathcal{C}} \sum_{j=1}^{n_i^{(s)}} (\mathbf{A}_i \vec{o}_{ij}^{(s)} + \vec{b}_i) \tag{3.7}$$

Here $n_i^{(s)}$ is the number of observations for speaker s for class i. $\mathbf{A}_i$ and $\vec{b}_i$ constitute the affine transformation to the generic class and $\mathcal{C}$ is the set of classes being transformed.

The transformation is computed in order to minimize the intra-speaker variation of observations from different phones. Our work is partially influenced by Doh and Stern who demonstrated the utility of an inter-class affine transformation [8].

To compute the transformation for each phone, we begin by computing a mean vector $\vec{\mu}_i^{(s)}$ for phone i for each speaker s from all training observations of that phone from that speaker. For practical purposes, one phone is chosen to be the destination phone into which all other phones are transformed. The mean vector for the predetermined destination class for each speaker s is defined as $\vec{\mu}_g^{(s)}$. The optimal LSLR transformation for phone i is thus expressed as follows, where N_s is the number of speakers in the training corpus:

$$\arg\min_{\mathbf{A}_i, b_i} \sum_{s=1}^{N_s} (\mathbf{A}_i \vec{\mu}_i^{(s)} + \vec{b}_i - \vec{\mu}_g^{(s)})^2 \tag{3.8}$$

The transform for phone i is shared over all speakers s and is computed to minimize the average intra-speaker distance between the transformed mean vector for phone i and the destination mean vector. An LSLR transformation is computed independently for each phone.

The minimization was performed with a standard least-squares approach. To estimate the transformation from phone i to the selected destination phone g, the average feature vectors for all of the speakers for phone i are collected in a matrix, $\mathbf{W}_i$. The last row of this matrix consists of 1's, to allow for computation of the shift:

$$\mathbf{W}_i = \begin{bmatrix} \vec{\mu}_i^{(0)} \\ \vdots \\ \vec{\mu}_i^{(N)} \\ 1 \ldots 1 \end{bmatrix} \tag{3.9}$$

The average feature vectors for phone g are collected into another matrix, $\mathbf{X}$, shown below:

$$\mathbf{X} = \begin{bmatrix} \vec{\mu}_g^{(0)} \\ \vdots \\ \vec{\mu}_g^{(N)} \end{bmatrix} \tag{3.10}$$

The rotation matrix and shift term are determined in the standard least-squares manner:

$$\left[\mathbf{A}_i \vec{b}_i\right] = ((\mathbf{W}_i^T \mathbf{W}_i)^{-1} \mathbf{W}_i^T \mathbf{X})^T \tag{3.11}$$

Once the transforms for each phone are computed, the characteristic feature vector for each speaker is computed by averaging the transformed feature vectors from all training observations of a selected subset of phones for that speaker. The distance between two speakers is defined as the Euclidean distance between their characteristic vectors:

$$d_{st} = \sqrt{\vec{v}^{(s)} \cdot \vec{v}^{(t)}} \tag{3.12}$$

Evaluation

When using the LSLR characteristic distance, a set of classes must be chosen. Only the data from this class set will be transformed and averaged to obtain the final LSLR characteristic for a particular speaker.

For many reasons, using the entire set of acoustic classes is not ideal. First, not all of the classes convey important information about the speaker. Fricatives and stops provide very little information about how a particular speaker's vowels and nasals will be realized acoustically. Nasals, although good for speaker identification [10], have little within speaker correlation with other phones. Computational considerations also motivate the selection of a limited set of phonetic classes. Estimating the transformation matrices involved in computing the LSLR characteristic takes a great deal of computation as well as storage.

Various class sets were tried for this purpose, shown in Table 3.3. All sets involved internal boundary models only. These models are labeled i(p), where p is the phone for which the internal landmark is being modeled. The cluster gender distributions for the 4 different sets are shown in Tables 3.4, 3.5, 3.6, and 3.7.

Class Set	Classes
Fricatives	i(f), i(s), i(sh), i(th), i(v), i(z), i(jh), i(zh)
Nasals	i(m), i(n), i(ng)
Vowels	i(aa), i(ae), i(ah), i(ah_fp), i(ao), i(aw), i(ax), i(axr), i(ay), i(eh), i(el), i(em), i(en), i(epi), i(er), i(ey), i(ih), i(ix), i(iy), i(ow), i(oy), i(uh), i(uw), i(ux)
Selected Vowels	i(aa), i(ae), i(ey), i(ih), i(iy), i(ow), i(uw)

Table 3.3: Class sets used for LSLR characteristic computation.

Cluster	Male (%)	Female (%)	Child (%)
1	81	14	5
2	72	22	5
3	65	24	11
4	69	24	7
5	93	5	2

Table 3.4: Gender distribution for 5 clusters created using the LSLR characteristic distance with the fricative class set.

Cluster	Male (%)	Female (%)	Child (%)
1	99	1	0
2	74	22	4
3	82	13	5
4	15	57	28
5	98	2	0

Table 3.5: Gender distribution for 5 clusters created using the LSLR characteristic distance with the nasal class set.

Cluster	Male (%)	Female (%)	Child (%)
1	100	0	0
2	100	0	0
3	98	2	0
4	1	64	34
5	37	55	9

Table 3.6: Gender distribution for 5 clusters created using the LSLR characteristic distance with the vowel class set.

We see that the fricatives worked very poorly for clustering speakers. Although the cluster containing the largest percentage of children, cluster 3, also contains the smallest percentage of males, all of the clusters have a male majority. These results were expected. Except for the lower cutoff frequencies of s and sh, the acoustic realizations of most fricatives show very little dependence on vocal tract length. Thus, an LSLR characteristic computed using fricative data will have very little correlation with the gender of the speaker.

The nasal set resulted in better clusters than the fricative set. We see that cluster 4 contains very few males and a large percentage of children, while clusters 1 and 5 are almost entirely male. While a speaker's realization of the nasals has little dependence on the oral cavity, it is highly correlated with the length of his vocal tract and the shape of his nasal passage. Both of these physical characteristics are correlated with a speaker's gender, which explains the gender separation that occurred in the resulting clusters.

The best clustering results were achieved with the two vowel sets. As shown in Table 3.6, using all of the internal vowel boundary classes results in three clusters that are almost entirely male, one cluster that is almost entirely children and females and one mixed cluster. This is also true of the clustering result obtained using the

Cluster	Male (%)	Female (%)	Child (%)
1	99	1	0
2	100	0	0
3	1	64	35
4	97	2	0
5	39	52	9

Table 3.7: Gender distribution for 5 clusters created using the LSLR characteristic distance with the selected vowel class set.

selected vowel set. This is explained by the fact that the acoustic realization of vowels is affected by both the vocal tract length and the oral cavity, both of which are highly correlated with the gender of the speaker.

In order to ascertain the impact of the rotation matrix on cluster quality, clusters were made using no transformation. The selected vowel set described in the previous section was used. Table 3.8, shows the clustering results for this case. This is equivalent to using the feature mean distance, but only averaging the features for the selected vowel set. We see that speakers of different genders are separated reasonably well. This can be attributed to two factors. First, the classes chosen for the selected vowel set are acoustically similar insofar as they are all vowels. Second, the vowels in the selected vowel set are used often. These two factors serve to mitigate the phonetic diversity problem encountered when using the feature mean distance. Still, comparing these results to those shown in Table 3.7, it is clear that using the transformation improves the quality of the clusters.

Cluster	Male (%)	Female (%)	Child (%)
1	96	4	0
2	92	7	2
3	13	58	29
4	95	5	0
5	11	77	13

Table 3.8: Gender distribution for 5 clusters created without transforming the data.

3.4 Conclusions

The results presented in this section have shown that both the feature mean distance and average class feature mean distance metrics perform poorly on our corpus. The failure of these two approaches is due largely to the small amount of data available for each speaker. When using the feature mean distance, this leads to feature mean vectors that are affected more by the phonetic content of a speaker's training data than by her general acoustic characteristics. Using the average class feature mean distance, this leads to pairs of speakers who share data for only a small number of classes, resulting in unreliable distance estimates.

The LSLR distance overcomes the shortcomings of the feature mean distance and average class feature mean distance. By transforming the speaker data from a selected set of classes into a generic, global class, the LSLR distance results in clusters that show separation of speakers of different genders. Use of a full matrix and shift and selection of an appropriate set of classes is important to the success of LSLR clustering.

Chapter 4

Speaker Cluster Weighting

While Chapter 3 explored robust methods for creating speaker clusters, this chapter will describe how those clusters were used as part of an adaptation approach called speaker cluster weighting (SCW).

4.1 Overview

Speaker cluster weighting consists of two phases. The first phase occurs off-line (i.e. before the recognition process begins), while the second phase occurs during recognition. These phases are illustrated in Figure 4.1.

The first phase of SCW consists of two steps. In the first step, speakers in the training data are clustered according to acoustic similarity. The preceding chapter presented various approaches to this task. In the second step, acoustic models are trained for each speaker cluster. These models need to be both focused and robust. Section 4.3 describes the approaches taken in this work for this second step.

During SCW's second phase, the acoustic models for each cluster are combined or selected in an optimal way using the adaptation data. Section 4.2 describes the algorithms explored for this process.

4.2 Cluster Model Combination or Selection

After clustering has been completed and acoustic models are trained for these clusters, adaptation is accomplished either by interpolating the cluster models or choosing the set that best represents the current speaker. This process must allow enough flexibility so that the resulting model can accurately represent those speakers poorly represented by the SI model. However, if the approach is too flexible, adaptation will not be robust in the presence of a small amount of adaptation data.

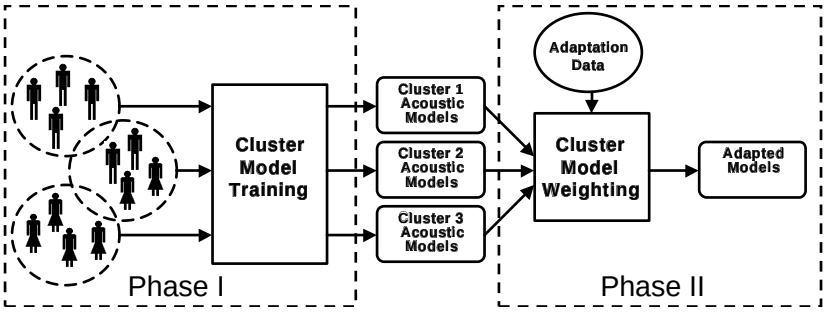

Figure 4.1: Schematic description of speaker cluster weighting.

4.2.1 Best Cluster

In the best cluster approach, the set of cluster acoustic models is chosen that maximizes the adaptation data probability. While this approach does not allow as much flexibility in the adaptation process as cluster weighting, it may be more robust in the presence of small amounts of adaptation data. The acoustic model resulting from choosing the best cluster from L clusters for a given phone p can be represented as:

$$\mathrm{p}_{bc}(\vec{x}|p) = \mathrm{p}_{l}(\vec{x}|p) \tag{4.1}$$

$\mathrm{p}_{l}(\vec{x}|p)$ is the probability distribution for class p for cluster l. l is chosen to satisfy the following equation:

$$l = \arg\max_{l} \mathrm{p}_{bc}(X|U,l) \tag{4.2}$$

If u_n represents the class label for adaptation token x_n, and assuming the observations are independent:

$$\mathrm{p}_{bc}(X|U,l) = \prod_{\forall \vec{x}_n} \mathrm{p}_{l}(\vec{x}_n|u_n) \tag{4.3}$$

4.2.2 Cluster Weighting

Whereas in the best cluster approach a "hard" decision is made about which cluster model to use for recognition, a "soft" decision is made when using the cluster weighting approach. In the two cluster weighting approaches described here, adapted acoustic models are weighted combinations of the cluster models. The acoustic model resulting from cluster weighting for L clusters and a given phone p can be represented

as:

$$\mathrm{p}_{scw}(\vec{x}|p) = \sum_{l=1}^{L} w_l \mathrm{p}_l(\vec{x}|p) \tag{4.4}$$

We examined two criteria for choosing the weights, $\vec{w}$. In the first case the likelihood of the adaptation data is maximized. In the second, a minimum classification error criterion is used.

Maximum Likelihood

Using maximum likelihood cluster weighting, the optimal weights, $\vec{w}$ will satisfy the equation:

$$\vec{w} = \arg\max_{\vec{w}} \mathrm{p}_{scw}(X|U, \vec{w}) \tag{4.5}$$

If u_n represents the class label for adaptation token x_n, and assuming the observations are independent:

$$\mathrm{p}_{scw}(X|U, \vec{w}) = \prod_{\forall \vec{x}_n} \mathrm{p}(\vec{x}_n|u_n, \vec{w}) \tag{4.6}$$

This maximization is performed using the EM algorithm.

Minimum Classification Error

When training acoustic models for recognition, the goal is to make the models as discriminative as possible. Training acoustic models using maximum-likelihood methods does not directly optimize for this criterion. This can be especially problematic when a limited amount of training data is available. To address this issue, various discriminative training methods have been developed [28].

When performing cluster weighting, the goal is also to make the resulting model as discriminative as possible in the face of limited training data. Thus, it seems reasonable to apply discriminative criteria to the cluster weighting problem.

The approach taken is based on that described in Chou, *et al* for training acoustic models [6]. First, a misclassification measure is defined based on the N-best utterance hypothesis $\{U_1 ... U_N\}$, with an empirically tuned parameter ν:

$$d(X, \vec{w}) = -\log\{\mathrm{p}(X, U_1|\vec{w})\} + \log\{\frac{1}{N-1}\sum_{k=2}^{N} e^{-\log\{\mathrm{p}(X,U_k|\vec{w})\}\eta}\}^{\frac{1}{\eta}} \tag{4.7}$$

The loss function is a sigmoid function of this misclassification error measure with an empirically tuned parameter η:

$$l(X, \vec{w}) = \frac{1}{1 + e^{-\gamma d(X, \vec{w})}} \tag{4.8}$$

The goal of minimum classification error (MCE) acoustic model training is to minimize the expected value of this loss function across all training utterance. In our work, MCE has only been applied to the case when we are adapting using one adaptation utterance, so we need only to minimize $l(X, \vec{w})$ for that particular utterance.

$l(X, \vec{w})$ is minimized through a simple gradient descent approach. At each iteration, $2L$ sets of possible cluster weights are computed. Each set of weights results from either adding or subtracting a small value ϵ from one of the weights in the current weight vector and then re-normalizing the weights such that they add to 1. The set of weights giving the lowest value for l is then adopted as the new set. This process continues until a minimal change in l results.

4.3 Cluster Model Training

The method in which the acoustic models are trained for each speaker cluster has a significant impact on both the recognition accuracy of the adapted models as well as the computation required to perform recognition. It is also intimately related to the cluster combination or selection algorithm being used. By altering the way the cluster models are trained, we are changing $\mathrm{p}_l(\vec{x}|p)$ in Equations 4.1 and 4.4.

4.3.1 Standard ML Training

In standard ML training, the Gaussian component means and variances are trained to maximize the likelihood of the cluster data. This is accomplished using the EM algorithm. The number of components in each model is determined using Equation 2.5, except that the maximum number of components is sometimes reduced.

4.3.2 Weight Adaptation

When using cluster weighting, the size of the resulting adapted acoustic models is of particular concern. If each of the cluster models is as large as the SI model, the resulting adapted acoustic models will require approximately L times as much computation as the SI models (where L is the number of clusters.) However, if the cluster models share the same Gaussian components, this problem is avoided. Training cluster models using weight adaptation consists of adapting only the weights of the Gaussian components of the SI model for each cluster. That is, for cluster l, for each class p, a set of component weights $\vec{c}_l^{(p)}$ is chosen such that:

$$\vec{c}_l^{(p)} = argmax_{\vec{c}_l^{(p)}} \vec{c}_l^{(p)} \cdot \vec{p}_{si}(X_l^{(p)}|p) \tag{4.9}$$

Here $\vec{p}_{si}(X_l^{(p)}|p)$ represents the vector of Gaussian component distributions from the speaker independent model for class p and $X_l^{(p)}$ is the set of observations for class p from cluster l. This maximization can be completed using the EM algorithm [9]. The resulting probability distribution for cluster l and class p can be expressed as follows:

$$\mathrm{p}_l(\vec{x}|p) = \vec{c}_l^{(p)} \cdot \vec{p}_{si}(\vec{x}|p) \tag{4.10}$$

Substituting this into Equation 4.4:

$$\mathrm{p}_{scw}(\vec{x}|p) = \sum_{l=1}^{L} w_l(\vec{c}_l^{(p)} \cdot \vec{p}_{si}(\vec{x}|p)) \tag{4.11}$$

Grouping terms, we obtain:

$$\mathrm{p}_{scw}(\vec{x}|p) = (\sum_{l=1}^{L} w_l \vec{c}_l^{(p)}) \cdot \vec{p}_{si}(x|p) \tag{4.12}$$

From this final equation, we see that creating the adapted acoustic model is simply a matter of assigning new weights to the speaker independent model components. These new component weights, $\vec{c}_{si}^{(p)}$, are computed by linearly combining the cluster component weights according to the adaptation weights:

$$\vec{c}_{si}^{(p)} = \sum_{l=1}^{L} w_l \vec{c}_l^{(p)} \tag{4.13}$$

4.3.3 Model Interpolation

The best cluster approach allows each cluster model to be the same size as the SI model without creating a computational problem. However, using the best cluster requires that each of the models be very robust. If not, an error in the model selection process or the selection of a model corresponding to a cluster with very little data will result in suboptimal performance, potentially even worse than that of the SI model. In another light, while cluster weighting achieves robustness by making a "soft" decision, the best cluster approach relies on the robustness of the cluster models themselves.

This robustness is achieved through model interpolation. With model interpolation, a set of small acoustic models is trained for each cluster. These models are then interpolated with the SI model. For those classes for which a large amount of training

data is present in the cluster, the cluster model is weighted more heavily. For those with a smaller amount of training data present in the cluster, the cluster model gets less weight. The exact formula for these weights is presented in [20] and shown here:

$$\mathrm{p}_{int}(\vec{x}|p) = \frac{N_p^{(l)}}{N_p^{(l)} + K}\mathrm{p}_l(\vec{x}|p) + \frac{K}{N_p^{(l)} + K}\mathrm{p}_{si}(\vec{x}|p) \tag{4.14}$$

Here $\mathrm{p}_{int}(\vec{x}|p)$ is the final model density for class p, $\mathrm{p}_l(\vec{x}|p)$ is the PDF of the model for class p trained on cluster l, and $\mathrm{p}_{si}(\vec{x}|p)$ is the PDF of the SI model for class p. $N_p^{(l)}$ is the number of examples of the class p found in cluster l and K is an empirically determined interpolation factor.

4.4 Experiments and Results

4.4.1 Experimental Setup

Recognition experiments were conducted using both gender and LSLR clustering. For gender clustering, three clusters were used, one containing all the males in the training data (13,030 calls, 59,076 utterances), one containing all the females (3,297 calls, 14,386 utterances) and one containing all the children (1206 calls, 7034 utterances.) The five clusters created using the selected vowel set, shown in Table 3.6, were used in the experiments using LSLR clusters.

Unless otherwise noted, recognition was performed on each utterance in the test set, using up to 5 of the same speaker's utterances to adapt. The utterance being recognized was not included in the adaptation data. The best path obtained using the SI model was used as the transcription for the adaptation data, making the task unsupervised.

4.4.2 Best Cluster

Model Training

For the experiments using the best cluster approach, models were trained using both standard ML training (described in Section 4.3.1) and model interpolation (described in Section 4.3.3.) The standard ML models had a maximum of 50 Gaussian components per class. In the case of model interpolation, a set of SI models were trained with a maximum of 40 components per class. These were interpolated with cluster models with a maximum of 10 components per class.

Results

The results of the experiments using the best cluster approach are shown using gender clusters and LSLR clusters in Figures 4.2 and 4.3. Looking at the results obtained using standard ML training, we see that, in some cases, recognition actually gets worse after one adaptation utterance. This can probably be attributed to the fact that, when adapting with only one utterance, it is more likely that a suboptimal cluster will be chosen.

This problem seems to be mitigated by using interpolated models. The use of interpolated models improves the recognition accuracy for speakers outside of a particular cluster. This means that recognition results are not as affected by an incorrect cluster choice. Still, the interpolated models are focussed enough to produce WER improvements after adaptation.

Tables 4.1 and 4.2 show WER's for the standard ML and interpolated cluster models on the test data. We see that the interpolated models perform better not only on speakers of the gender corresponding to the cluster, but also on speakers of different genders. This supports the hypothesis that, when using interpolated models, WER is not as affected by an incorrect cluster choice.

	Speaker Gender		
Model	Male	Female	Child
Male	13.2%	35.0%	66.1%
Female	30.1%	17.0%	30.8%
Child	42.6%	25.3%	29.2%

Table 4.1: WER for speakers of different genders on standard ML cluster models.

	Speaker Gender		
Model	Male	Female	Child
Male	12.1%	23%	41.9%
Female	14.3%	17.1%	29.2%
Child	14.1%	18.5%	27.3%

Table 4.2: WER for speakers of different genders on interpolated cluster models.

4.4.3 Maximum Likelihood Cluster Weighting

Model Training

Experiments using maximum likelihood cluster weighting were conducted using models trained using standard ML training as well as models trained using weight adap-

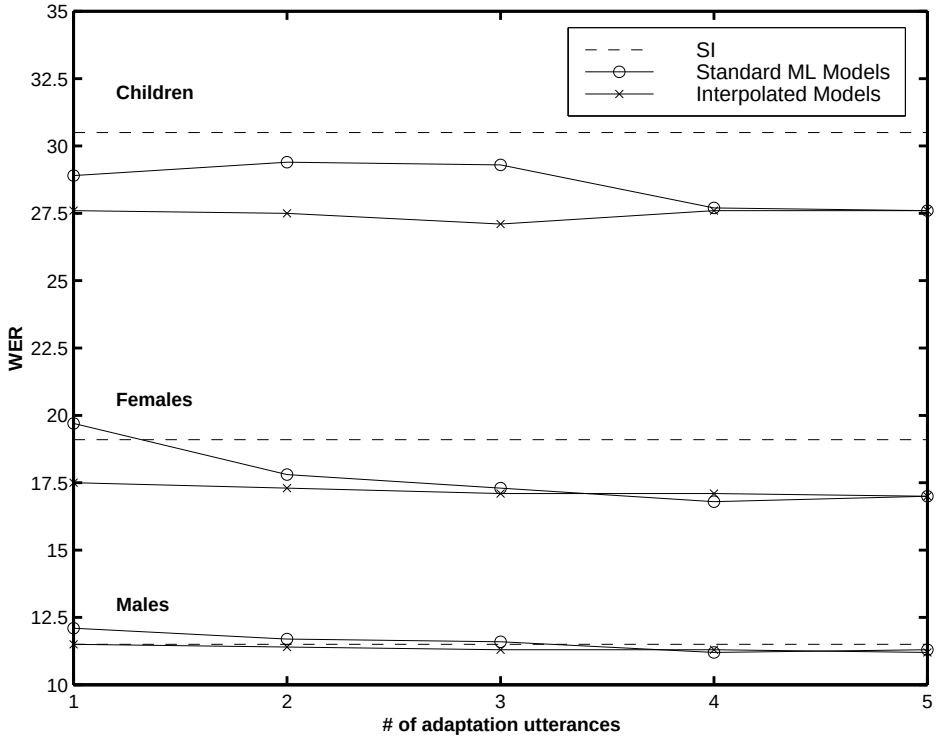

Figure 4.2: Recognition results using best cluster with gender clusters. Cluster models were trained in the standard ML manner and using model interpolation.

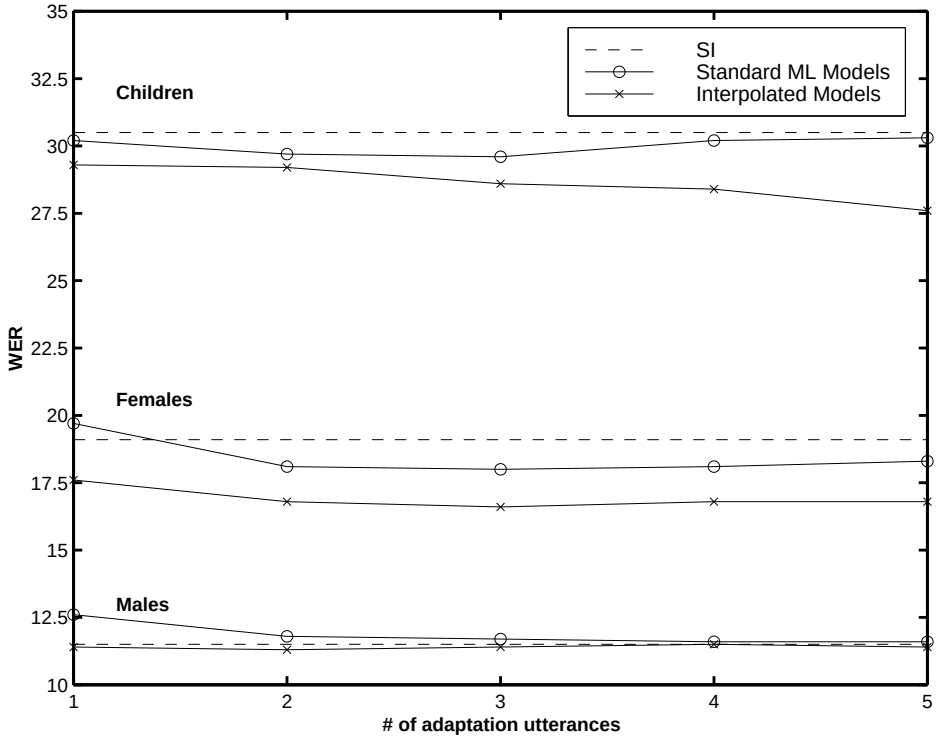

Figure 4.3: Recognition results using best cluster with LSLR clusters. Cluster models were trained in the standard ML manner and using model interpolation.

tation (described in Section 4.3.2.) When using standard ML training, each set of cluster models had a maximum of 12 Gaussian components per class when using the 3 gender clusters and a maximum of 9 components per class when using the LSLR clusters. Two SI models, one with a maximum of 12 components per class and one with a maximum of 9 components per class were also trained. Smaller standard ML models were necessary so that the final adapted models would be comparable in size to the SI models. When using weight adaptation, SI models with a maximum of 50 components per cluster were trained. Weight adaptation was then used to create each of the cluster models.

Results

Figures 4.4 and 4.5 show results using maximum likelihood cluster weighting with gender and LSLR clusters. We see that, as was the case when using BC with models trained in the standard ML manner, a degradation in performance sometimes occurs after the first adaptation utterance. This is probably because one adaptation utterance is not enough data to robustly estimate the cluster weights.

Using the weight adapted models results in better performance than when using models trained in the standard ML manner, except on child speakers using gender clusters. There are two possible reasons for this. First, while the same number of cluster weights are being determined when using the weight adapted models, the adapted model's Gaussian component weights are more restricted. When using the standard ML models, any Gaussian component in any model in a particular cluster is tied to all the Gaussian components *in that cluster.* However, when using weight adapted models, a components weight is tied to all the Gaussian components *in all the clusters.* This added restriction allows for more robust adaptation of the weights. Second, because only the component weights are being adapted for each cluster model, training involves fewer free parameters, and the cluster models can be more robustly trained.

The fact that better performance is achieved on children using the standard ML models when using gender clusters is notable. One possible explanation is that, in this case, many more Gaussian components are dedicated to child speakers than in the SI model. While only about 9% of the training data consist of children, one third of the Gaussian component means in the adapted model will have been determined by the child speakers.

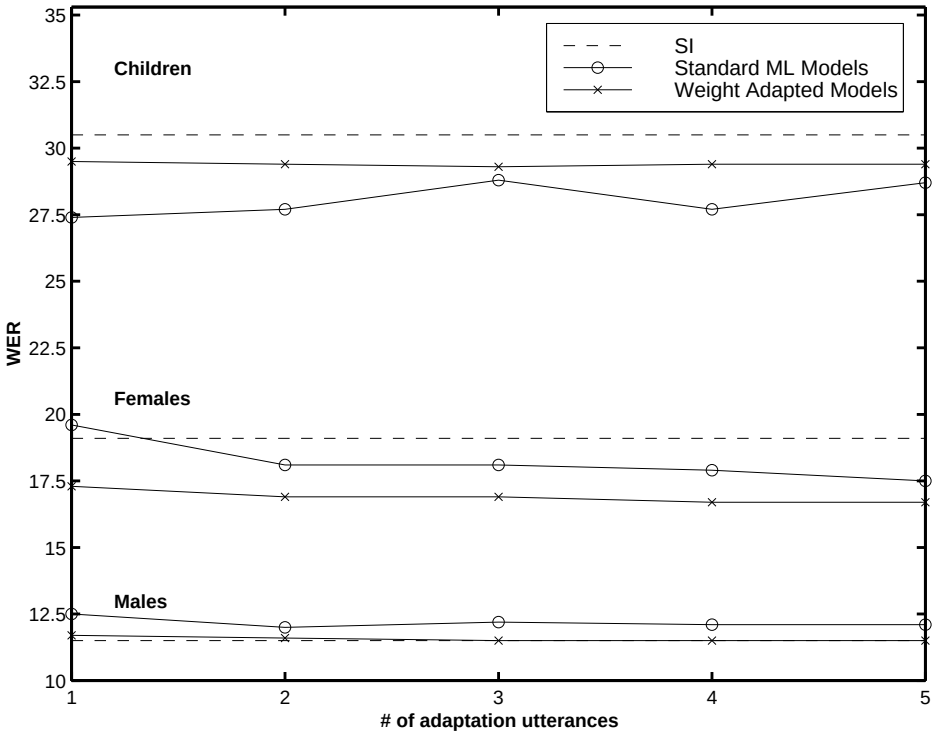

Figure 4.4: Recognition results using maximum likelihood cluster weighting with gender clusters. Cluster models were trained in the standard ML manner and using weight adaptation.

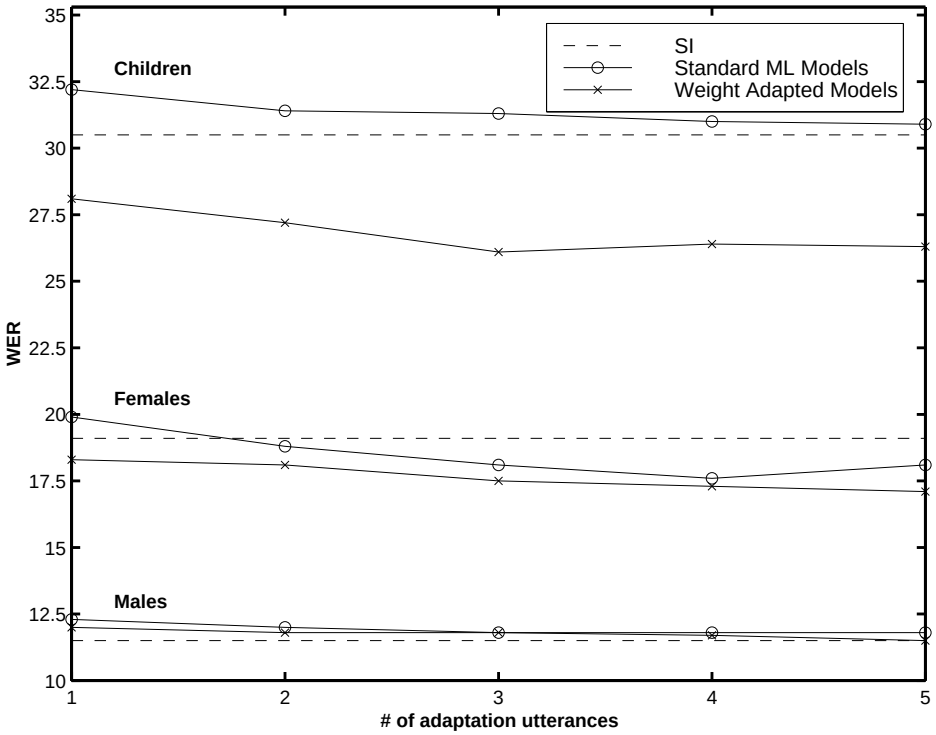

Figure 4.5: Recognition results using maximum likelihood cluster weighting with LSLR clusters. Cluster models were trained in the standard ML manner and using weight adaptation.

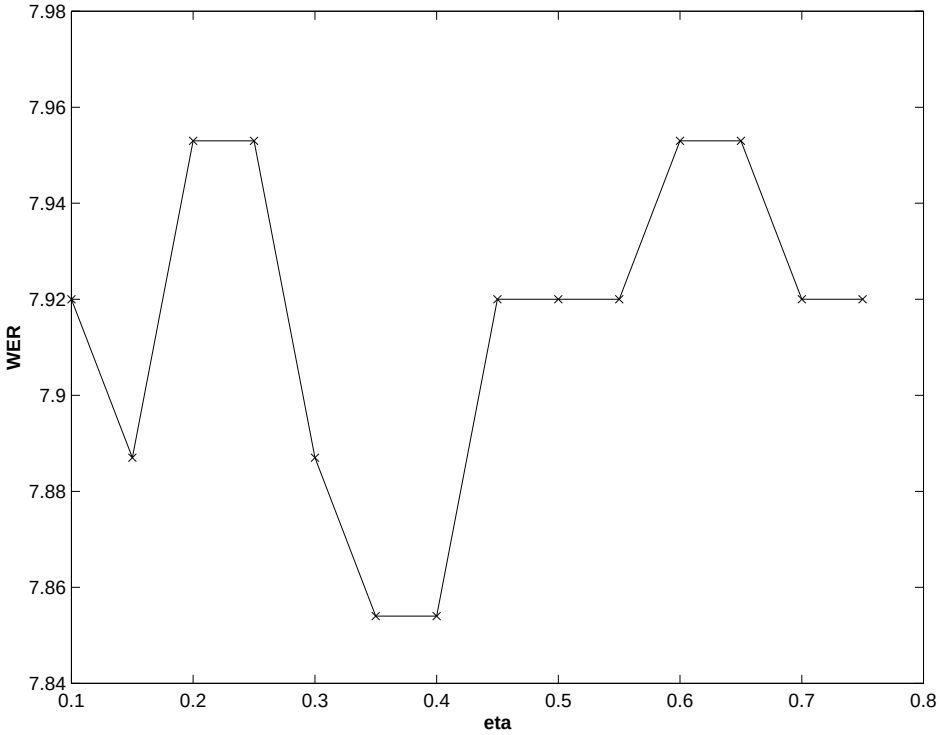

Figure 4.6: Average WER for various values of η.

4.4.4 Minimum Classification Error Cluster Weighting

Model Training

Minimum classification error (MCE) cluster weighting was applied only to the case when the cluster models were trained using weight adaptation and gender clusters. The same weight adapted models were utilized as were used in the ML case.

Tuning γ and η

As is described in Section 4.2.2 and shown in Equations 4.7 and 4.8, using the MCE criterion for cluster weighting requires that two parameters be estimated, γ and η. γ was chosen to be .001. A small fractional value was necessary due to the very small log probabilities obtained for each N-best hypothesis. It was found that performance on the development set remained the same for values of η greater than .8, so values of η between .1 and .8 were then considered. The WER's averaged over males, females and children for these values are shown in Figure 4.6. From these results, .35 was chosen for use in the experiments.

Results

Experiments were run under four different test conditions: supervised instantaneous, unsupervised instantaneous, supervised rapid and unsupervised rapid. Here, instantaneous refers to the case when the same utterance was used for adaptation as for recognition. Rapid refers to the case when a different utterance was used for adaptation. All the MCE experiments used only one adaptation utterance.

Figure 4.7 shows the relative reductions in WER obtained under the 4 different test conditions. We notice that, in all cases except rapid adaptation on the children speakers, performance degrades when going from supervised to unsupervised adaptation. This is to be expected due to mistakes in the best path used for adaptation in the unsupervised case.

We also notice that the difference in WER reduction between the supervised and unsupervised cases is larger for the instantaneous adaptation experiments. This may be because, in the instantaneous case, adapting using an incorrect transcription works *directly* to increase the likelihood of the incorrect transcription given the adaptation data. However, the negative impact this has on recognizing utterances different from the adaptation data comes from the *indirect* effect of class models poorly adapted due to the incorrect transcription.

We notice that, in the case of males and children, moving from supervised instantaneous adaptation to supervised rapid adaptation results in worse performance. This makes sense given that adaptation in the instantaneous case is focussed exactly on those acoustic models that are used to recognize the utterance. It is unclear why a decrease does not occur for the female speakers.

Using Figure 4.8, we can also compare the results of unsupervised rapid adaptation using ML cluster weighting with the results obtained using MCE cluster weighting. We see that using the MCE criterion leads to lower WER's for females and children, while performance for males remains the same. These results are reasonably consistent with work done using discriminative criteria with MLLR [30, 18]. In Wallhoff, *et al*, a variant of MLLR using the MMI criterion resulted in an improvement using supervised adaptation [30]. Gunawardana and Byrne also used the MMI criterion with MLLR, achieving improvements in the supervised case [18].

Gunawardana and Byrne ran experiments in the unsupervised case as well, and, as was the finding in this work, this resulted in a degradation in comparison to supervised adaptation. However, the degradation seen in that work lead to WER's higher than when using standard MLLR, while the unsupervised results obtained here are still superior to those obtained with ML. This may be attributed to the much smaller number of parameters being adapted in this work. This allows the parameter space to be searched more thoroughly and also allows for more robust estimation in the presence of a small amount of data.

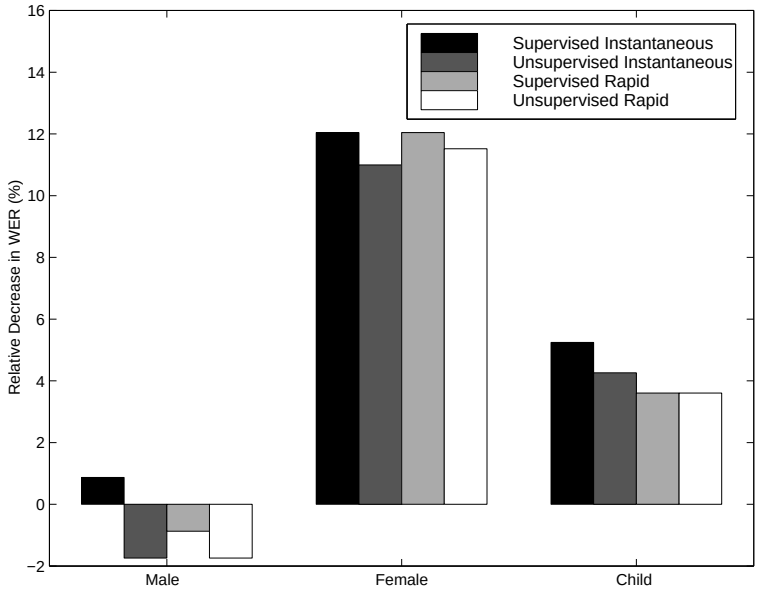

Figure 4.7: Relative reduction in WER for instantaneous and rapid adaptation in the supervised and unsupervised cases using MCE criterion.

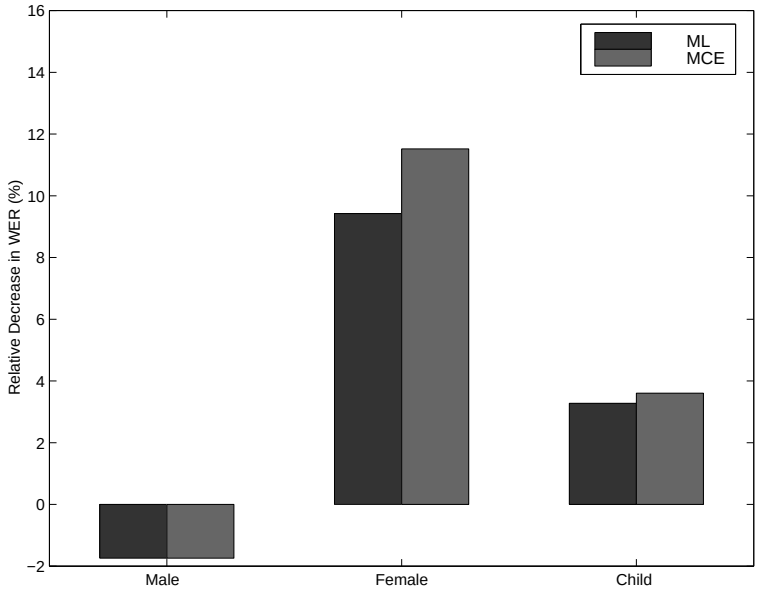

Figure 4.8: Relative reduction in WER using unsupervised rapid adaptation with ML and MCE criteria, both using one adaptation utterance and weight adapted gender cluster models.

4.4.5 Analysis

Best Cluster vs. Cluster Weighting

Figures 4.9 and 4.10 allow us to compare the best cluster approach to EM cluster weighting. Both figures show results using the best cluster approach with interpolated models and maximum likelihood cluster weighting with weight adapted models.

As a first observation, we notice that neither the maximum likelihood cluster weighting approach nor the best cluster approach is consistently superior. This is not even the case if the clustering approach is held constant.

Looking at the children speakers, we see that, when using gender clusters, the best cluster approach outperforms maximum likelihood cluster weighting. However, when using LSLR clusters, the opposite is true. This is most likely because none of the LSLR clusters consist entirely of children. This means that, when using best cluster, a single cluster model may not represent a child speaker well. However, maximum likelihood cluster weighting allows for the interpolation of different models. It's possible that, for instance, interpolating a set of models from a cluster consisting largely of females with a set of models from a cluster with a large number of children will produce a better set of models for a particular child speaker than the set of models corresponding to a cluster with a large number of children alone.

In the case of gender clustering, one of the clusters consists entirely of children. Although this cluster may not be optimal for all children speakers, the negative impact of the robustness problem encountered when estimating interpolation weights outweighs the negative impact of being forced to choose only one cluster.

For females, the performance of best cluster and cluster weighting are comparable when using gender clustering. When using LSLR clusters, however, best cluster decreases WER by up to 1.3% more than maximum likelihood cluster weighting.

Looking at the male speakers, again the performance of best cluster and cluster weighting are comparable when gender clusters are used. However, using LSLR clusters, cluster weighting outperforms best model for male speakers when less than 5 adaptation utterances are present.

The speed at which adaptation occurs is also important. We observe that, when using 5 LSLR clusters, adaptation occurs more quickly using the best cluster approach. Using gender clusters, adaptation occurs at about the same rate for both sets of clusters.

Gender vs. LSLR Clustering

In Figures 4.11 and 4.12 we see that neither gender nor LSLR clustering proved consistently superior to the other. This suggests that LSLR clustering is capturing meaningful acoustic correlations between speakers.

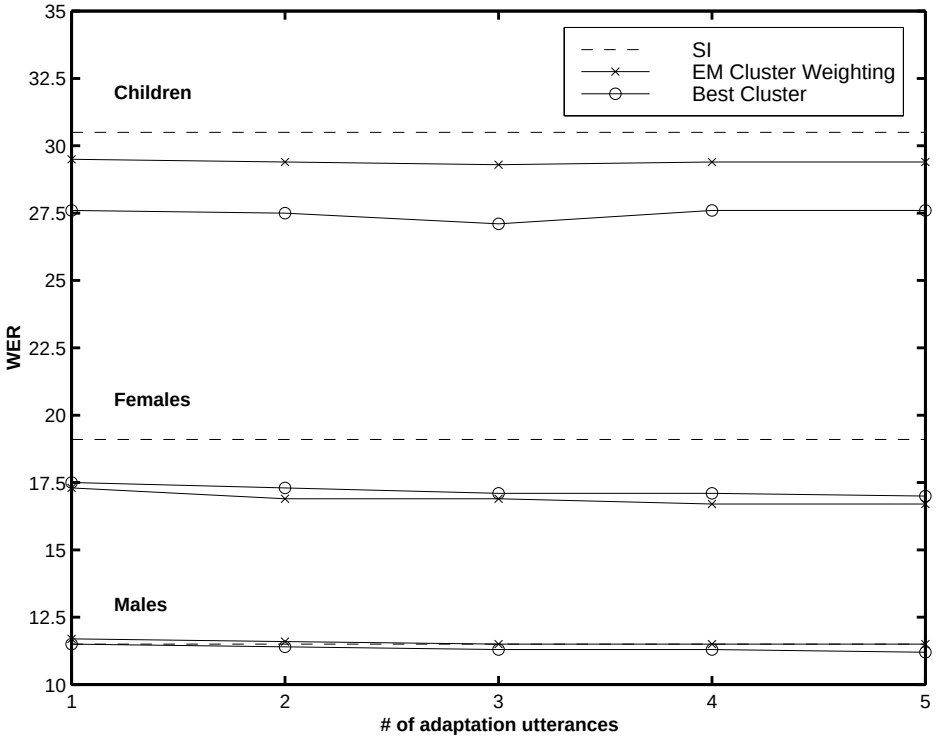

Figure 4.9: Recognition results using maximum likelihood cluster weighting with weight adapted models and best model with interpolated models. Gender clusters were used in these experiments.

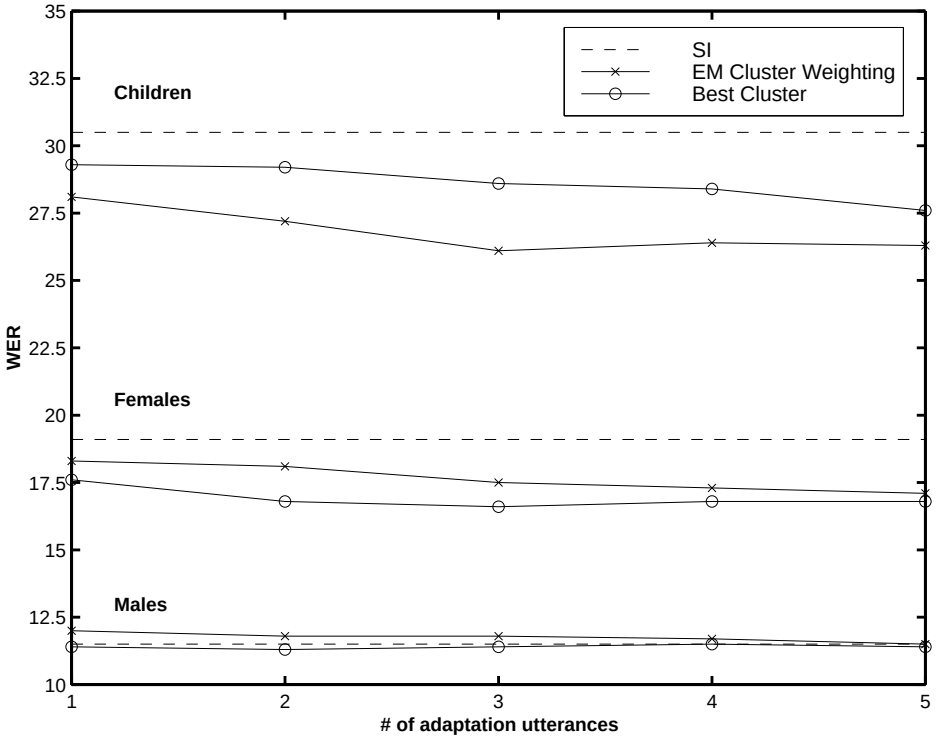

Figure 4.10: Recognition results using maximum likelihood cluster weighting with weight adapted models and best model with interpolated models. Clusters were created using the LSLR distance metric.

Examining Figure 4.11, we see that after 3 adaptation utterances, using 5 LSLR clusters results in 3.2% more relative reduction in WER than using gender clusters. Children comprise only about 9% of the training data. It may be that the flexibility provided by having more clusters allows better adaptation to those speakers least represented in the training data.

Again using Figure 4.11, we can compare the speed of adaptation when using 3 gender clusters to that achieved when using 5 LSLR clusters. In both cases ML cluster weighting is used. We see that, ignoring which set of clusters results in better overall performance, it takes longer to reach the minimum WER when using the 5 LSLR clusters than it does when using the 3 gender clusters. This makes sense, given that using 5 clusters requires the estimation of 6 parameters (a weight for each cluster model and the SI model), while using 3 clusters requires the estimation of only 4 parameters. Robustly estimating fewer parameters requires less adaptation data.

4.4.6 Conclusions

From the recognition experiments performed in this section, we can draw five main conclusions. First, when using the best model approach, using interpolated models results in significant improvement over using standard ML trained models. Interpolated models provide robustness to compensate for the hard decision made when the best cluster is chosen.

Second, when using the cluster weighting approach, weight adapted cluster models are superior to standard ML cluster models. The weight adapted models lead to more constraint during the weighting procedure, while weight adaptation also allows for more robust training of the individual cluster models.

Third, using an MCE criterion for cluster weighting results in small improvements over the ML criterion. Although improvements are greatest using supervised instantaneous adaptation, gains are still seen in the unsupervised rapid case. This is consistent with other work done using the MCE criterion for adaptation [30, 18].

Fourth, neither best cluster nor cluster weighting performs consistently better than the other. While cluster weighting achieves higher reductions in WER on children using gender clusters and on females using LSLR clusters, best cluster achieves higher reductions on children using LSLR clusters and on females using gender clusters. Because of its simplicity, it would seem that, in many applications, best cluster would be preferable.

Finally, LSLR and gender clustering perform similarly overall. However, results on child speakers, shown in Figure 4.11 suggest that, by allowing for more than 3 clusters, LSLR clustering provides added flexibility that results in better adaptation for children speakers. On the other hand, Figure 4.11 suggests that increasing the number of clusters decreases the speed of adaptation.

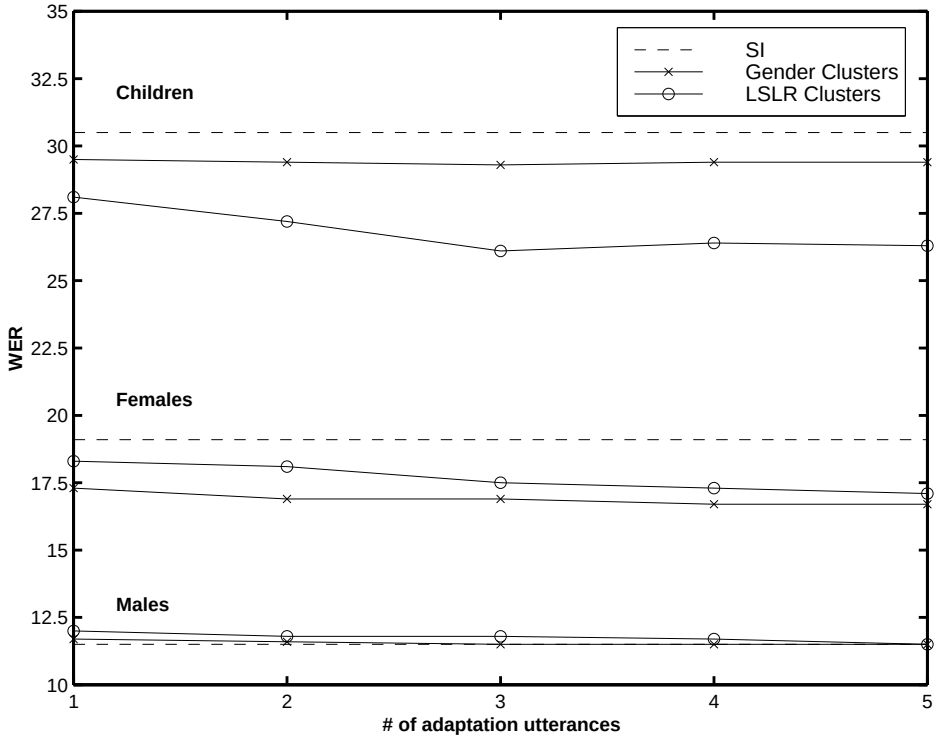

Figure 4.11: Recognition results using 3 gender clusters and 5 LSLR clusters. Both cases used EM cluster weighting with weight adapted models.

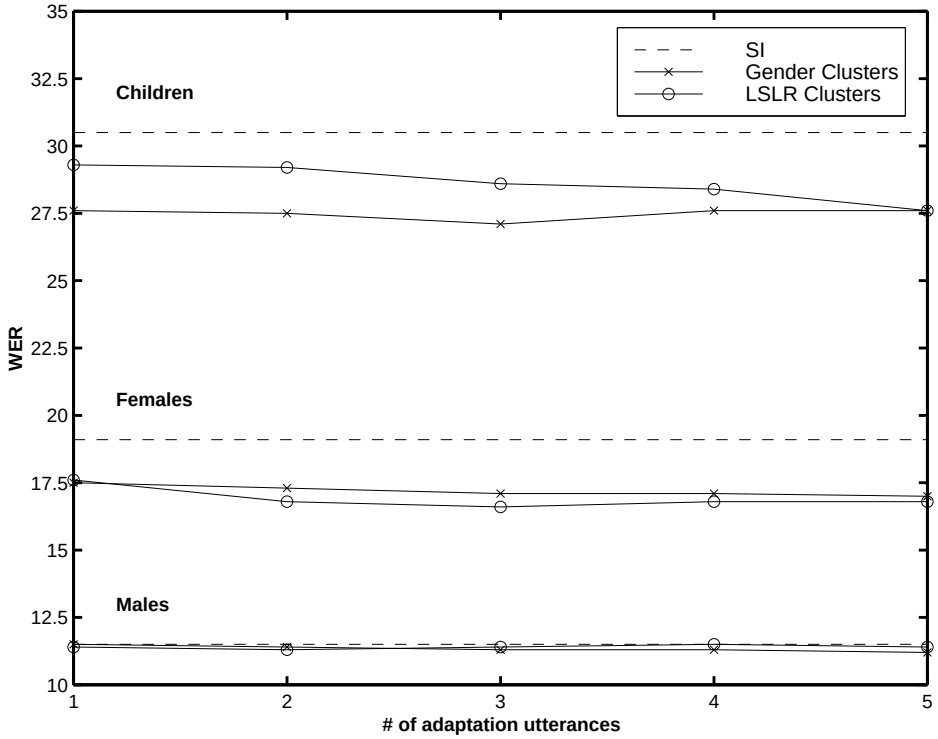

Figure 4.12: Recognition results using 3 gender clusters and 5 LSLR clusters. Both cases used best model with interpolated models.

Chapter 5

Conclusion

5.1 Summary

In Chapter 3, we explored three different distance metrics for automatic speaker clustering. We showed that two of these metrics, the feature mean distance and average class feature mean distance, performed poorly on the Jupiter training corpus. This was due to the small amount of data available for each speaker. In the case of the feature mean distance, this led to feature means that were skewed by the phonetic content of each speaker's data. In the case of the average feature mean distance, the lack of per-speaker data led to pairs of speaker's who did not have enough data for the same classes to allow for a reliable measure of distance between them. By transforming a small set of vowels from each speaker to a generic phonetic space, then averaging the transformed feature vectors, the LSLR characteristic distance was able to overcome this problem. With this metric, clusters were created that showed separation of speakers of different genders.

In Chapter 4, various strategies applicable to SCW were described and evaluated. This included the evaluation of gender and LSLR clustering in the context of SCW, as well as approaches to cluster combination and selection and cluster model training. Using the best cluster approach to model selection with model interpolation and gender clustering resulted in a 9% relative improvement for female speakers, and a 10% relative improvement for children speakers after 1 adaptation utterance. A 16% improvement was obtained for child speakers and 10% improvement for females using 5 LSLR clusters with maximum likelihood cluster weighting and 5 adaptation utterances.

Overall, LSLR clustering yielded results comparable to those obtained using gender clusters. The superior performance of LSLR clustering on children after 5 adaptation utterances suggested that, by allowing the creation of a larger number of clusters,

those speakers most different from the majority in the corpus could be better adapted to. However, the larger number of clusters also lead to longer adaptation times.

The important relationship between the cluster model training algorithm and the cluster combination or selection algorithm was also demonstrated in Chapter 4. Interpolated cluster models, by more robustly modeling the speaker clusters as well as softening the best cluster hard decision, performed consistently better with the best cluster approach than standard ML trained models. Weight adaptation proved superior to standard ML training when using the cluster weighting approach, by allowing the creation of larger cluster acoustic models without increasing the size of the final adapted model,

Finally, it was shown that using an MCE criterion for cluster weighting yielded slight improvements over using an ML criterion. Specifically, an additional 2% relative improvement for females after 1 adaptation utterance was obtained using the MCE criterion with gender clusters. Performance for males and children remained the same.

5.2 Future Extensions

One straightforward direction for future work involves improving the major steps in SCW. LSLR clustering uses only vowels, and does not weight the vowels according to their speaker-discriminative power. An approach where more of a speaker's data is used for clustering and where different acoustic classes are weighted according to how well they characterizes a speaker could result in better speaker clusters.

Different cluster model training procedures could also be tried. MLLR or MAP could be used to train acoustic models for the best cluster approach. Using the cluster weighting approach, cluster model means could be adapted instead of component weights. If the model weights and variances were kept constant in the adapted acoustic models, this would make SCW look more like CAT or eigenvoices.

Another possible direction for future work would be to extend SCW for longer term adaptation. This could be done by systematically increasing the number of clusters as the amount of adaptation data increases. It could also be accomplished by combining SCW with MLLR or MAP. One possibility is to adapt the individual cluster models using one of these approaches before performing cluster weighting.

The LSLR characteristic distance metric could also be used as a point of departure for future work. The feature normalization technique used to create LSLR characteristic vectors could be used in other contexts. One possible application is in speaker identification. By transforming all speaker data into a generic class (or generic classes), it may be possible to reduce both the amount of enrollment data as well as the amount of data needed for speaker identification.

Bibliography

[1] S. Ahadi and P. Woodland. Combined bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 11:187–206, 1997.

[2] C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon press, Oxford, 1995.

[3] J. Chang. *Near-Miss Modeling: A Segment-Based Approach to Speech Recognition*. PhD thesis, MIT Department of Electrical Engineering and Computer Science, 1998.

[4] K. Chen and H. Wang. Eigenspace-based maximum *a posteriori* linear regression for rapid speaker adaptation. In *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 917–920, Salt Lake City, 2001.

[5] C. Chesta, O. Siohan, and C. Lee. Maximum *a posteriori* linear regression for hidden markov model adaptation. In *Proc. European Conf. on Speech Communication and Technology*, pages 211–214, Budapest, 1999.

[6] W. Chou, C. Lee, and B. Juang. Minimum error rate training based on n-best string models. In *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages 625–655, Minneapolis, Minnesota, 1993.

[7] T. Claes, I. Dologlou, L. Bosch, and D. Compernolle. A novel feature transformation for vocal tract length normalization in automatic speech recognition. *IEEE Transactions on Signal and Audio Processing*, 6(6):549–557, 1998.

[8] S. Doh and R. Stern. Inter-class MLLR for speaker adaptation. In *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 1775–1778, Istanbul, Turkey, 2000.

[9] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, NY, 1973.

[10] J.P. Eatock and J.S. Mason. A quantitative assessment of the relative speaker discriminant properties of phonemes. In *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 133–136, Adelaide, April 1994.

[11] G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, The Netherlands, 1970.

[12] M. Gales. Cluster adaptive training of hidden markov models. *IEEE Transactions on Signal and Audio Processing*, 8(4):417–428, 2000.

[13] T. Gao, M. Padmanabhan, and M. Pichney. Speaker adaptation based on pre-clustering training speakers. In *Proc. European Conf. on Speech Communication and Technology*, volume 4, pages 2091–2094, Rhodes, Greece, 1997.

[14] J. Gauvain and C. Lee. Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of markov chains. *IEEE Transactions on Signal and Audio Processing*, 2:291–198, 1994.

[15] J. Glass. *Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition*. PhD thesis, MIT Department of Electrical Engineering and Computer Science, 1998.

[16] J. Glass, J. Chang, and M. McCandless. A probabilistic framework for feature-based speech recognition. In *Proc. Intl. Conf. on Spoken Language Processing*, pages 2277–2280, Philadelphia, October 1996.

[17] J. R. Glass, T.J. Hazen, and I. L. Hetherington. Real-time telephone-based speech recognition in the Jupiter domain. In *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 61–64, Phoenix, 1999.

[18] A. Gunawardana and W. Byrne. Discriminative speaker adaptation with conditional maximum likelihood linear regression. In *Proc. European Conf. on Speech Communication and Technology*, Aalborg, Denmark, 2001.

[19] T. J. Hazen. A comparison of novel techniques for rapid speaker adaptation. *Speech Communication*, 31:15–33, 2000.

[20] T.J. Hazen. *The Use of Speaker Correlation Information for Automatic Speech Recognition*. PhD thesis, MIT Department of Electrical Engineering and Computer Science, January 1998.

[21] D. Klatt and L. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[22] T. Kosaka and S. Sagayama. Tree-structured speaker clustering for fast speaker adaptation. In *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 245–248, Adelaide, Australia, 1994.

[23] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707, 2000.

[24] L. Lee and R. Rose. A frequency warping approach to speaker normalization. *IEEE Transactions on Signal and Audio Processing*, 6(1):49–60, 1998.

[25] C. Leggetter and P. Woodland. Flexible speaker adapatation using maximum likelihood linear regression. In *Proc. of the ARPA Spoken Langauge Technology Workshop*, pages 104–109, 1995.

[26] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adpatation of continuous density hidden markov models. *Computer Speech and Language*, 9(2):171–185, 1995.

[27] P. Nguyen, C. Wellekens, and J. Junqua. Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments. In *Proc. European Conf. on Speech Communication and Technology*, Budapest, 1999.

[28] R. Schluter, W. Macherey, B. Muller, and H. Ney. Comparison of discriminative training criteria and optimization methods for speech recognition. *Speech Communication*, 34:287–310, 2001.

[29] K. Shinoda and C. Lee. A structural bayes approach to speaker adaptation. *IEEE Transactions on Signal and Audio Processing*, 9(3):276–287, 2001.

[30] F. Wallhoff, D. Willett, and G. Rigoll. Speaker adaptation based on pre-clustering training speakers. In *Proc. European Conf. on Speech Communication and Technology*, Aalborg, Denmark, 2001.

[31] N. Wang, S. Lee, F. Seide, and L. Lee. Rapid speaker adaptation using *a priori* knowledge by eigenspace analysis of MLLR parameters. In *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, Salt Lake City, 2001.

[32] B. Zhou and J. Hansen. A novel algorithm for rapid speaker adaptation based on structural maximum likelihood eigenspace mapping. In *Proc. European Conf. on Speech Communication and Technology*, pages 1215–1218, Aalborg, Denmark, September 2001.