A Comparison of Novel Techniques for Rapid Speaker Adaptation

This is a preprint version of an article published in
Speech Communication, Vol. 31, No. 1, pp. 15-33, May 2000.

This paper is based on a EUROSPEECH-97 article entitled
" A Comparison of Novel Techniques for Instantaneous Speaker Adaptation"

Timothy J. Hazen

Affiliation:
Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology

Abstract

This paper introduces two novel techniques for rapid speaker adaptation, *reference speaker weighting* and *consistency modeling*. Also presented is an adaptation technique called *speaker cluster weighting* which provides a means for improving upon generic *hierarchical speaker clustering* techniques. Each of these adaptation methods attempts to utilize the underlying within-speaker correlations that are present between the acoustic realizations of different phones. By accounting for these correlations, a limited amount of adaptation data can be used to adapt the models of every phonetic acoustic model, including those for phones which have not been observed in the adaptation data. Results were obtained using the DARPA Resource Management corpus for a set of rapid adaptation experiments where single test utterances were used for adaptation and recognition simultaneously. Using the new adaptation techniques relative word error rate reductions ranging from 4.9% to 8.4% were obtained under various conditions. Using a combination of hierarchical speaker clustering techniques and the novel adaptation techniques, a word error rate reduction of 20% has been achieved from the baseline speaker independent recognition system.

1 Introduction

When developing a speaker independent (SI) automatic speech recognition system, it is important to account for the wide variability that can be present in any speech waveform. This variability can result from changes in the individual speaker, the speaker's environment, the microphone and channel of the recording device, and/or the mechanism which converts the signal into its digital representation. However, it is important not to overlook the fact that the sources of variability often remain fixed throughout any single spoken utterance. In other words, typical speech utterances come from one speaker who stays in the same environment and is recorded using a fixed set of equipment. This knowledge can be used to provide constraints to the recognizer. Thus, by obtaining a little information about the current speaker, environment, microphone, and channel, a speech recognizer should be able improve its performance by adapting to the characteristics particular to the current utterance. In this paper, we will concentrate our efforts on the specific issues surrounding the utilization of speaker constraint within a speaker independent system and will leave the issues regarding environmental, channel and microphone constraint for another time. The goal of the research is to be able to rapidly adapt a speech recognition system to a speaker using only a small amount of adaptation data.

Over the last ten to twenty years, dramatic improvements in the quality of speaker independent speech recognition technology have been made. With the development and refinement of the Hidden Markov Model (HMM) approach [Baker 1975, Bahl 1983, Lee 1988], today's speech recognition systems have been shown to work effectively on various large vocabulary, continuous speech, speaker independent tasks. However, despite the high quality of today's speaker independent systems [Bahl 1995, Gauvain 1995, Kubala 1997], there can still be a significant gap in performance between these systems and their speaker adaptive (SA) or speaker dependent (SD) counterparts. The reduction in a system's error rate between its speaker independent mode and its speaker dependent mode can be more then 50%[Hazen 1998].

The reason for the gap in performance between SI and SD systems can be attributed to flawed assumptions used in the probabilistic framework and training methods employed by typical speech recognizers. One primary problem lies in the fact that almost all speech recognition approaches, including the prevalent HMM approach, assume that all observations extracted from the same speech waveform are statistically independent after being conditioned on the underlying phone string. It has been observed that different acoustic observations extracted from speech from the same speaker can be highly correlated [Hazen 1998]. Thus, assuming independence between observations extracted from the same utterance ignores *speaker correlation*

information which may be useful for decoding the utterance. Speaker correlation information will be defined here as the statistical correlation between different speech events produced by the same speaker.

In SI systems, the independence assumption is particularly troublesome because the SI acoustic models are usually trained from a pool of data which includes all of the available observations from all available training speakers. Using this training technique, SI acoustic models have a much larger variance than a typical SD acoustic model trained on speech from only one speaker. Because of this, SI models do not match any one speaker well despite the fact that they may perform adequately across all speakers. One the other hand, SD models work well because they tightly match the acoustic characteristics of the one speaker on which they are trained and used.

In this paper we will discuss the problem of introducing speaker constraint into a speaker independent speech recognition system. This discussion will begin with a presentation of the probabilistic framework of the system we will utilize for our experiments. Next, we will present three methods for introducing speaker constraint into the probabilistic framework. The first method, called *speaker cluster weighting* is a means of applying adaptation to a standard *hierarchical speaker clustering* approach. The second method is a novel adaptation technique called *reference speaker weighting* which can rapidly adapt the parameters of a set of models to match the current speaker, based on the current speaker's similarity to a set of reference speakers from the training data. The final method is a new and unique approach called *consistency modeling* which utilizes speaker correlation information in the acoustic modeling process without performing any explicit speaker adaptation. The paper will conclude with a presentation and discussion of our experimental results when using our techniques to perform rapid speaker adaptation.

2 Probabilistic Framework

In this paper we are concerned with the acoustic modeling problem, i.e., given a sequence of acoustic observations, we must determine the likelihood that these observations were produced by a particular string of phonetic units. To describe the problem mathematically, let U represent a sequence of phonetic units. If U contains a sequence of N units then let it be expressed as:

$$U = \{u_1, u_2, \ldots, u_N\} \tag{1}$$

Here each u_n represents the identity of one phonetic unit in the sequence. Next, let X be a sequence of feature vectors which represent the acoustic information of an utterance. In standard HMM systems, each feature vector would represent one short frame of speech where each phonetic segment may span multiple frames. However, in the system used in this work, segment-based feature vectors are used. These vectors contain acoustic information spanning multiple frames and are mapped one-to-one with hypothesized phonetic segments. If X contains one feature vector for each unit in U then X can be expressed as:

$$X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\} \tag{2}$$

Given the above definitions, the likelihood of observing the feature vectors in X given the string of phonetic units U is represented as $\mathrm{p}(X|U)$. This expression is referred to as the *acoustic model*.

In order to develop effective and efficient methods for estimating the acoustic model likelihood, typical recognition systems use a variety of simplifying assumptions. To begin, the general expression can be expanded as follows:

$$\mathrm{p}(X|U) = \mathrm{p}(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N|U) = \prod_{n=1}^{N} \mathrm{p}(\vec{x}_n|\vec{x}_{n-1}, \ldots, \vec{x}_1, U) \tag{3}$$

At this point, speech recognition systems almost universally assume that the acoustic feature vectors are independent. With this assumption the acoustic model is expressed as follows:

$$\mathrm{p}(X|U) = \prod_{n=1}^{N} \mathrm{p}(\vec{x}_n|U) \tag{4}$$

Because this is a standard assumption in most recognition systems, the term $\mathrm{p}(\vec{x}_n|U)$ will be referred to as the *standard acoustic model*.

Speech recognition systems often simplify the problem further by utilizing only a portion of the context available in U when scoring any given feature vector $\vec{x}_n$.

The most extreme simplification is the assumption of context independence. In this case the output feature vector is dependent only on the identity of its corresponding phone. Thus, a context independent acoustic model is represented as:

$$\mathrm{p}(X|U) = \prod_{n=1}^{N} \mathrm{p}(\vec{x}_n|u_n) \tag{5}$$

All of the experiments presented in this paper will be performed using a context independent system. However, the probabilistic framework that will be developed in this section does not make any assumptions about the amount of context that will be used during modeling. Thus, the full phonetic string U is used in all of the probabilistic expressions that will be presented in the remainder of this section even though only a small amount of phonetic context is likely to be used by typical recognition systems.

In Equation (3), the likelihood of a particular feature vector is deemed dependent on the observation of all of the feature vectors which have preceded it. In Equation (4), each feature vector $\vec{x}_n$ is treated as an independently drawn observation which is not dependent on any other observations, thus implying that no statistical correlation exists between the observations. What these two equations do not show is the net effect of making the independence assumption. Consider applying Bayes rule to the likelihood expression $\vec{x}_n$ as expressed in Equation (3). In this case the likelihood expression for $\vec{x}_n$ can be rewritten as:

$$\mathrm{p}(\vec{x}_n|\vec{x}_{n-1},\ldots,\vec{x}_1,U) = \mathrm{p}(\vec{x}_n|U)\frac{\mathrm{p}(\vec{x}_{n-1},\ldots,\vec{x}_1|\vec{x}_n,U)}{\mathrm{p}(\vec{x}_{n-1},\ldots,\vec{x}_1|U)} \tag{6}$$

After applying Bayes rule, the conditional probability expression contained in (3) is rewritten as a product of the standard acoustic model $\mathrm{p}(\vec{x}_n|U)$ and a probability ratio which we refer to as the *consistency ratio*. The *consistency ratio* is a multiplicative factor which is ignored when the feature vectors are considered independent. It represents the contribution of the correlations which exist between the feature vectors.

To understand what information is conveyed by the consistency ratio, it is important to understand the difference between the numerator and denominator. Both the numerator and denominator provide a likelihood score for all of the feature vectors preceding the current feature vector $\vec{x}_n$. In the numerator, this likelihood score is conditioned on $\vec{x}_n$ while in the denominator it is not. In essence, this ratio is determining if all of the previous observed feature vectors are more likely or less likely given the currently observed feature vector $\vec{x}_n$ and the given phonetic sequence U.

Consider what this ratio represents during recognition when the phonetic string U is merely a hypothesis which may contain errors. When scoring a hypothesis, the

standard acoustic model would be responsible for scoring each $\vec{x}_n$ as an independent element. The consistency ratio would then be responsible for determining if the current feature vector and its phone hypothesis is *consistent* with the previous feature vectors and their phone hypotheses under the assumption that the entire utterance was spoken by the same speaker. If the hypotheses for all of the previous feature vectors are *consistent* with the hypothesis for the current feature vector then it is expected that the value of the numerator will be greater than that of the denominator. However, if the current feature vector's hypothesis is *inconsistent* with the hypotheses of the previous feature vectors then it is expected that the numerator would be less than the denominator.

Given the above description, it is easy to see that the consistency ratio can be used to account for the within-speaker correlations which exist between phonetic events. As such the consistency ratio provides a measure of speaker constraint which is lacking in the standard SI acoustic model. Hypotheses whose aggregate consistency ratio is greater than one are deemed consistent with the assumption that all of the phones were spoken by the same person. These hypotheses thus have their standard acoustic model likelihoods boosted by the application of the consistency ratio. Likewise, hypotheses deemed to be inconsistent by the consistency ratio have their standard acoustic model likelihoods reduced.

If an accurate estimate of the consistency ratio can be obtained then all of the speaker correlation information which is ignored in the standard acoustic model will be accounted for in the estimate for $\mathrm{p}(X|U)$. However, this ratio requires an estimate for the likelihood of a large joint feature vector $(\vec{x}_{n-1}, \ldots, \vec{x}_1)$ under two different conditions. This is a very difficult modeling problem which will be discussed in Section 5.

The independence assumption is a major weakness of typical SI systems. By ignoring the correlations which exist between different observations, these systems are unable to provide any speaker constraint. On the other hand, SD systems provide full speaker constraint. Because SD systems have been trained with a large amount of speech from the one speaker of interest, there is relatively nothing new to be learned about the speaker's models from newly observed speech from that speaker. Because of this, if we assume that the only significant source of correlation between acoustic observations is the individual speaker, the consistency ratio for a speaker dependent system can be approximated as follows:

$$\frac{\mathrm{p}_{sd}(\vec{x}_{n-1}, \ldots, \vec{x}_1|\vec{x}_n, U)}{\mathrm{p}_{sd}(\vec{x}_{n-1}, \ldots, \vec{x}_1|U)} \approx 1 \tag{7}$$

Taking this into account, the acoustic model can utilize the following approximation

when the recognition is performed in speaker dependent mode:

$$\mathrm{p}_{sd}(\vec{x}_n|\vec{x}_{n-1},\ldots,\vec{x}_1,U) \approx \mathrm{p}_{sd}(\vec{x}_n|U) \tag{8}$$

In short, this states that the independence assumption is relatively sound for SD systems.

In this research, because we focus on modeling acoustic correlation due to the speaker only, we are are implicitly assuming that correlations from other effects do not exist. Strictly speaking, the independence assumption is not completely validated when the system is a well trained SD system. Other factors could contribute to the existence of correlations between different observations. Some additional sources of constraint which may also affect the speech signal are the speaker's physiological state (healthy or sick), the speaker's emotional state (happy or sad), and the speaking style (read speech or spontaneous speech).

If it is assumed that the independence assumption is valid for SD systems, then it is reasonable to believe that the invalidity of the independence assumption in SI mode is a major factor in the severe drop in performance when a system is moved from SD mode to SI mode. This being said there are two ways of addressing the problem. The first way is to try to adjust the set of standard acoustic models used during recognition to match, as closely as possible, the characteristics of the current speaker (even if the current speaker is a stranger in the system's eyes). This is the approach taken by systems which utilize *speaker adaptation*. The most common approaches to speaker adaptation include *maximum a posteriori probability* (MAP) adaptation [Gauvain 1994], *extended maximum a posteriori probability* (EMAP) adaptation [Lasry 1984], and *maximum likelihood linear regression* (MLLR) adaptation [Leggetter 1995]. The second possible way to attack the problem is to utilize speaker correlation information directly within the probabilistic framework of the SI system. One way to accomplish this is to create models which can be used to estimate the contribution of the consistency ratio. This approach will be called *consistency modeling*. Both approaches are examined in this paper.

3 Speaker Clustering Techniques

3.1 Hierarchical Speaker Clustering

One method of providing speaker constraint to speech recognition systems that has proven successful is hierarchical speaker clustering [Furui 1989, Kosaka 1994a, Kosaka 1994b, Mathan 1990]. Hierarchical speaker clustering allows similar training speakers to be clustered to create models which represent specific speaker types. In this approach, similar reference speakers are grouped together into a speaker cluster for which one model is trained.

There are a variety of ways in which a hierarchical speaker cluster tree can be constructed. The construction can be performed using unsupervised bottom-up clustering based on an acoustic similarity measure [Kosaka 1994a, Kosaka 1994b], unsupervised top-down clustering based on an acoustic similarity measure [Furui 1989, Mathan 1990], or some supervised method. In our case, a very simple cluster tree is created in a supervised fashion. This tree first clusters speakers by gender and then into three classes of speaking rate, fast, medium and slow. This yields a total of six different models at the leaves of the tree. Figure 1 illustrates the hierarchical speaker clustering that we utilized.

When using speaker clustering, there is a trade-off between robustness and specificity. Large clusters are more general but can be trained more robustly. Smaller clusters can represent more specific speaker types but may lack a sufficient amount of training data required for accurate density function estimation (i.e., the sparse data problem). To increase the robustness of the models in the tree, model interpolation is utilized. For example, the final interpolated acoustic models used for each gender dependent phone model, $\mathrm{p}_{igd}(\vec{x}|u)$, are an interpolation of the maximum likelihood trained gender dependent model, $\mathrm{p}_{gd}(\vec{x}|u)$ and the speaker independent model, $\mathrm{p}_{si}(\vec{x}|u)$. The form of this interpolation is:

$$\mathrm{p}_{igd}(\vec{x}|u) = \lambda \mathrm{p}_{gd}(\vec{x}|u) + (1-\lambda)\mathrm{p}_{si}(\vec{x}|u) \tag{9}$$

Similarly, an interpolated gender and speaking rate dependent model, $\mathrm{p}_{igrd}(\vec{x}|u)$, can be created using the expression:

$$\mathrm{p}_{igrd}(\vec{x}|u) = \lambda_1 \mathrm{p}_{grd}(\vec{x}|u) + \lambda_2 \mathrm{p}_{gd}(\vec{x}|u) + (1-\lambda_1-\lambda_2)\mathrm{p}_{si}(\vec{x}|u) \tag{10}$$

The λ values are determined from the training data using deleted interpolation [Bahl 1991, Huang 1996]. Deleted interpolation optimizes the λ values by maximizing the likelihood of data jack-knifed from the training set using the Expectation

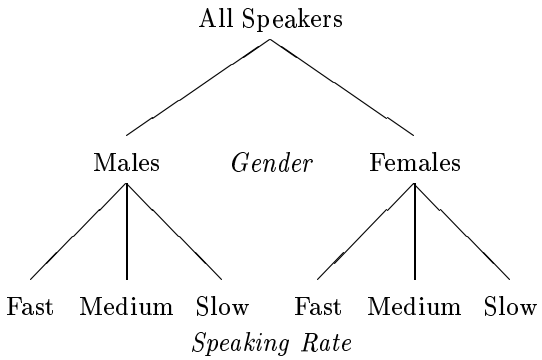

Figure 1: Hierarchical cluster tree utilized by our system.

Maximization (EM) algorithm. Using the deleted interpolation algorithm, each phone model receives a different set of interpolation weights. If a particular speaker cluster has plenty of data to reliably estimate the density function for a particular phone, then the interpolation weights typically favor the more specific cluster model. On the other hand, if a cluster contains only a small amount of data for a particular phone, then the interpolation weights typically place emphasis on the more general model.

There are a variety of ways in which the cluster models can be used during recognition. One potential system could run all of the cluster models in parallel and choose the model which produces the best scoring path. In our experiments a serial approach is utilized, i.e., recognition is performed with a two-pass strategy. First, the test utterance is passed through the speaker independent (SI) recognizer. The best path using the SI models is then rescored by gender specific models to determine the gender of the speaker. The best path is also utilized to estimate the speaking rate. The appropriate gender and speaking rate specific model is then used for a second recognition pass. In our experiments different recognition experiments were conducted using either gender dependent (GD) clusters or six gender and speaking rate dependent (GRD) clusters.

3.2 Speaker Cluster Weighting

When using hierarchical speaker clustering, recognition is performed using a model set selected from a finite set of predetermined model sets. The individual models in each predetermined model set are themselves interpolations of various general and specific models. The weightings used to perform the interpolation are precomputed using the deleted interpolation algorithm. An alternative approach is an interpolation scheme

which determines the weighting factors on the fly to match the current speaker. This is the basic idea behind *speaker cluster weighting* (SCW) adaptation.

In speaker cluster weighting, a predetermined set of L different model sets exists. The final SCW model set used for recognition uses weighted combinations of the models from the predetermined set of L different models sets. Let $p_l(\vec{x}|u)$ represent the acoustic model from model set l for phonetic unit u. The final SCW model for phonetic unit u is a weighted combination of the L different models and can be represented as:

$$\mathrm{p}_{scw}(\vec{x}|u) = \sum_{l=1}^{L} w_l \mathrm{p}_l(\vec{x}|u) \tag{11}$$

The difficult part of the problem is to determine the values for each w_l weight. These weights can be different for each phonetic unit or they can be shared amongst phonetic units belonging to a common class. Sharing the weighting factors across all phonetic units within a predetermined phonetic class helps provide weighting factor estimates which are more robust in the face of limited or sparse adaptation data. For each class of phones the goal is to find the set of weights which maximizes the likelihood of the adaptation data for that phonetic class for the current speaker. To illustrate the SCW process, consider the problem of finding the single optimal set of global weights. The problem is cast in a maximum likelihood framework as follows:

$$\vec{w}' = \arg\max_{\vec{w}} \mathrm{p}_{csw}(X|U, \vec{w}) \tag{12}$$

Here, the weights are represented in the weighting vector $\vec{w}$ as follows:

$$\vec{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_L \end{bmatrix} \tag{13}$$

By assuming each observation is independent of other observations and its surrounding context, this maximization process becomes:

$$\vec{w}' = \arg\max_{\vec{w}} \prod_{n=1}^{N} \mathrm{p}_{scw}(\vec{x}_n|u_n, \vec{w}) \tag{14}$$

This maximization process is easily performed by the EM algorithm.

To perform the maximization process for finding the optimal weights, the phonetic transcription U must be provided. The phonetic transcription from the best path provided by the SI recognizer can be used to approximate the true phonetic transcription.

There are two finals steps in constructing an SCW system. The first step is determining the set of cluster models used by the SCW algorithm. The set of models used in these experiments contain the same nine ML trained model sets appearing at the nine nodes of the hierarchical tree shown in Figure 1. In other words, the set of models contains one SI model set, two GD model sets, and six GRD model sets.

The second step is determining the different phonetic classes, each of which will receive a different weighting vector. Experiments have shown that a very simple set of three phonetic classes works best [Hazen 1998]. The experiments in this paper use three different weight vectors: one for standard phonetic models, one for silence models, and one for the *anti-phone* model (see [Glass 1996] for a description of the anti-phone model and how it is used to normalize segment model scores).

4 Reference Speaker Weighting

In this section we will discuss an adaptation technique which we refer to as *reference speaking weighting* (RSW). This technique is designed to combine the strengths of the parameter sharing techniques utilized by many standard adaptation techniques, such as *maximum likelihood linear regression* (MLLR) adaptation, with the strengths of speaker constraint present in typical speaker clustering techniques. The primary strength of the MLLR adaptation algorithm lies in its ability to jointly adapt the parameters of multiple acoustic models using a shared linear transformation. However, the MLLR assumption that different acoustic models can be jointly adapted using the same linear transformation ignores *a priori* knowledge about the actual underlying relationship between different phonetic events produces by the same speaker. Ideally, the adaptation of a set of models should utilize *a priori* knowledge obtained from training data about how the models of different phonetic units are likely to be jointly constrained. Speaker clustering techniques are one means of defining this constraint.

The basic premise behind reference speaker weighting is that the model parameters of a speaker adapted model can be constructed from a weighted combination of model parameters from a set of individual *reference speakers*. As with hierarchical speaker clustering, the robust training of model parameters is an important issue. Because the amount of data available from each reference speaker may be limited, it might not be possible to robustly train a full acoustic model for every phone for every reference speaker. Thus, our reference speaker weighting technique limits its focus to a small set of model parameters which can be robustly trained for each speaker. Our system only utilizes the *centroid* or *center of mass* of a model (we use these terms instead of the term *mean* to distinguish between the *centroid* of a mixture Gaussian model and the means of the individual mixture components). The *centroid* of a mixture Gaussian model with M components can be expressed as:

$$\vec{c} = \sum_{i=1}^{M} \omega_i \vec{\mu}_i \tag{15}$$

In this expression $\vec{\mu}_i$ is a mixture component's mean vector and ω_i is the component's weight. Using $\vec{c}$, we can re-express each mixture component mean vector as follows:

$$\vec{\mu}_i = \vec{c} + \vec{\nu}_i \tag{16}$$

In this expression $\vec{\nu}_i$ is simply an offset which, when added to $\vec{c}$, yields the mixture component mean, $\vec{\mu}_i$. Using these definitions it can be seen that the location of a model can be altered without changing the model's shape simply by adjusting the vector $\vec{c}$. This type of adjustment will be referred to as model translation.

In deriving the RSW approach, we begin by assuming a set of R different reference speakers exists within the training data. We also assume that for each reference speaker a reasonably accurate estimate of the centroid of the acoustic model for each of P different phonetic units has been obtained. In our experiments, a small number of speakers had little or no training for some of the less common phonetic units. Because this prevents robust estimation of the centroids for these phones using ML estimation, MAP estimation of the centroids was used. Let the centroid for phone p of reference speaker r be represented as $\vec{c}_{p,r}$. Furthermore, the collection of centroid vectors for an individual speaker can be concatenated into a single *speaker vector*. Let the speaker vector for reference speaker r be defined as $\vec{m}_r$. The mathematical representation of the speaker vector $\vec{m}_r$ is thus given as:

$$\vec{m}_r = \begin{bmatrix} \vec{c}_{1,r} \\ \vec{c}_{2,r} \\ \vdots \\ \vec{c}_{P,r} \end{bmatrix} \tag{17}$$

Furthermore, the entire set of reference speaker vectors can be represented by the matrix $\mathbf{M}$ which will be defined as:

$$\mathbf{M} = [\, \vec{m}_1 \,;\, \vec{m}_2 \,;\, \ldots \,;\, \vec{m}_R \,] = \begin{bmatrix} \vec{c}_{1,1} & \vec{c}_{1,2} & \cdots & \vec{c}_{1,R} \\ \vec{c}_{2,1} & \vec{c}_{2,2} & \cdots & \vec{c}_{2,R} \\ \vdots & \vdots & \ddots & \vdots \\ \vec{c}_{P,1} & \vec{c}_{P,2} & \cdots & \vec{c}_{P,R} \end{bmatrix} \tag{18}$$

The portion of $\mathbf{M}$ which contains only the center of mass vectors for the p^{th} model can be represented as $\mathbf{M}_p$ and is expressed as:

$$\mathbf{M}_p = [\, \vec{c}_{p,1} \,;\, \vec{c}_{p,2} \,;\, \cdots \,;\, \vec{c}_{p,R} \,] \tag{19}$$

This allows $\mathbf{M}$ to be expressed as:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \vdots \\ \mathbf{M}_P \end{bmatrix} \tag{20}$$

During adaptation, the goal is to determine the most likely speaker vector, $\vec{m}$, for a test speaker given the available adaptation data. It is also desirable to utilize the *a priori* knowledge provided by the reference speaker vectors about the correlations between the centroids of different models. However, past approaches which

have attempted to build statistical models to govern the adaptation of $\vec{m}$, such as the extended maximum *a posteriori* (EMAP) adaptation approach, have run into difficulty because of the sparse data problem associated with training the large set of correlation parameters required by their models. To avoid this problem we seek a solution where these correlations can be accounted for without having to explicitly train a large *a priori* statistical model containing many parameters. One possible solution is to use the speaker vectors in $\mathbf{M}$ to constrain the speaker space in which $\vec{m}$ may fall. Specifically, the value of $\vec{m}$ can be constrained to be a weighted average of the speaker vectors contained in $\mathbf{M}$. This can be expressed as:

$$\vec{m} = \mathbf{M}\vec{w} \tag{21}$$

Here $\vec{w}$ is a weighting vector which allows a new speaker vector to be created via a weighted summation of the reference speaker vectors in $\mathbf{M}$. The portions of $\vec{m}$ and $\mathbf{M}$ which represent phonetic unit p can be expressed as $\vec{c}_p$ and $\mathbf{M}_p$, thus allowing the following expression:

$$\vec{c}_p = \mathbf{M}_p\vec{w} \tag{22}$$

To find the optimal value of $\vec{w}$ a maximum likelihood approach can be utilized. The goal is to find the value of $\vec{w}$ which maximizes the likelihood of a set of adaptation data. Let $\mathcal{X}$ represent the adaptation data. In particular, let $\mathcal{X}$ be represented as:

$$\mathcal{X} = \{\, X_1,\, X_2,\, \ldots,\, X_P \,\} \tag{23}$$

Here each X_p is a set of example observations from the p^{th} phonetic unit. Furthermore, the sets of observations for each unit will be represented as:

$$X_p = \{\, \vec{x}_{p,1},\, \vec{x}_{p,2},\, \ldots,\, \vec{x}_{p,N_p} \,\} \tag{24}$$

Here each $\vec{x}_{p,n}$ is a specific observation vector of phonetic unit p and N_p is the total number of adaptation observations available for unit p. Note that it is possible for N_p to be zero for any given unit, especially when only a small amount of adaptation data is available. Using the above definitions the goal is to find the optimal value of $\vec{w}$ using the following maximum likelihood expression (as expressed in the log domain):

$$\arg\max_{\vec{w}} \log \mathrm{p}(\mathcal{X}|\vec{w}). \tag{25}$$

In solving for the optimal $\vec{w}$ the common assumption that all observations are independent is made. With this assumption the expression reduces to:

$$\arg\max_{\vec{w}} \sum_{p=1}^{P} \sum_{n=1}^{N_p} \log \mathrm{p}(\vec{x}_{p,n}|\vec{w}). \tag{26}$$

Next, the density function must be defined. A single full covariance Gaussian density function is used to approximate the mixture Gaussian density function used by each phonetic unit model. The density function for phone p can thus be expressed as:

$$\mathrm{p}(\vec{x}_{p,n}|\vec{w}) \equiv \mathcal{N}(\vec{c}_p, \mathbf{S}_p) \tag{27}$$

Here $\mathbf{S}_p$ represents the speaker independent covariance matrix for unit p, which will remain constant.

It can be shown that the expression in (26) reduces to the following expression:

$$\arg\max_{\vec{w}} \; 2\vec{v}^T\vec{w} - \vec{w}^T\mathbf{U}\vec{w}. \tag{28}$$

Here $\mathbf{U}$ and $\vec{v}$ are defined as follows:

$$\mathbf{U} = \sum_{p=1}^{P}\sum_{n=1}^{N_p}\mathbf{M}_p^T\mathbf{S}_p^{-1}\mathbf{M}_p = \sum_{p=1}^{P}N_p\mathbf{M}_p^T\mathbf{S}_p^{-1}\mathbf{M}_p \tag{29}$$

$$\vec{v}^T = \sum_{p=1}^{P}\sum_{n=1}^{N_p}\vec{x}_{p,n}^T\mathbf{S}_p^{-1}\mathbf{M}_p \tag{30}$$

Before, solving for $\vec{w}$ the following two constraints are also applied:

$$\forall i \; w_i \geq 0 \;\text{ and }\; \sum_{i=1}^{R} w_i = 1 \tag{31}$$

A simple hill climbing algorithm can be utilized to find the value of $\vec{w}$ which maximizes the likelihood of the data under the constraints given.

For the experiments that will be presented in this paper, the centroids for all of the models are adapted using one global weighting vector. However, the RSW framework can be easily extended to handle multiple weighting vectors covering different phonetic classes. This is akin to the approach taken by most MLLR systems where a varying number of MLLR transforms can be utilized depending on the amount of available adaptation data. This process is discussed in more detail in [Hazen 1998]. Though not done in our experiments, the size of the weighting vector that must be estimated can also be reduced by applying eigen analysis techniques to the reference speaker matrix $\mathbf{M}$ and utilizing only the most significant eigen vectors. This approach, called *eigenvoices*, was introduced in [Kuhn 1998].

5 Consistency Modeling

5.1 Probabilistic Framework

As discussed in the introduction, one potential method for incorporating speaker constraint into a speech recognition system is to explicitly model the *consistency ratio*. We will refer to this type of modeling as *consistency modeling*. To introduce the theoretical aspects of consistency modeling, consider the probabilistic framework introduced in Section 2. In the probabilistic framework the likelihood of a sequence, X, of N acoustic measurements being produced by the underlying sequence of phonetic units, U, can be expressed as follows:

$$\mathrm{p}(X|U) = \prod_{n=1}^{N} \mathrm{p}(\vec{x}_n|\vec{x}_{n-1},\ldots,\vec{x}_1,U) = \prod_{n=1}^{N} \mathrm{p}(\vec{x}_n|U)\frac{\mathrm{p}(\vec{x}_{n-1},\ldots,\vec{x}_1|\vec{x}_n,U)}{\mathrm{p}(\vec{x}_{n-1},\ldots,\vec{x}_1|U)} \tag{32}$$

In examining this expression, the likelihood of any particular acoustic observation $\vec{x}_n$ can be realized as the product of two separate terms. The first term is the *standard acoustic model*, i.e., the model that is used when the acoustic observations are considered independent. The second term is a ratio which will be referred to as the *consistency ratio*. As discussed earlier, this ratio compares the likelihood of the previously observed phones when considering and not considering the latest observation.

With the consistency ratio defined, the difficulty lies in devising a means of modeling this ratio. Modeling a large joint expression such as $\mathrm{p}(\vec{x}_{n-1},\ldots,\vec{x}_1|U)$ would be extremely difficult with anything but the simplest probabilistic models. Even the use of a single full covariance Gaussian model, though easy to construct, would be computationally expensive to use. For the purpose of practicality, one simplifying assumption will be made. It will be assumed that only the correlations between the current observation and each of the individual past observations are necessary to estimate the value of the consistency ratio. With this assumption the consistency ratio can be approximated as follows:

$$\frac{\mathrm{p}(\vec{x}_{n-1},\ldots,\vec{x}_1|\vec{x}_n,U)}{\mathrm{p}(\vec{x}_{n-1},\ldots,\vec{x}_1|U)} \approx \prod_{k=1}^{n-1} \frac{\mathrm{p}(\vec{x}_k|\vec{x}_n,U)}{\mathrm{p}(\vec{x}_k|U)} \tag{33}$$

This assumes that ignoring the correlations between the observations $\vec{x}_{n-1},\ldots,\vec{x}_1$, which exist in both the numerator and the denominator, will not affect the final result. This expression can be equivalently expressed as:

$$\prod_{k=1}^{n-1} \frac{\mathrm{p}(\vec{x}_k|\vec{x}_n,U)}{\mathrm{p}(\vec{x}_k|U)} = \prod_{k=1}^{n-1} \frac{\mathrm{p}(\vec{x}_n,\vec{x}_k|U)}{\mathrm{p}(\vec{x}_n|U)\mathrm{p}(\vec{x}_k|U)} \tag{34}$$

The full score for a hypothesized path can thus be written as:

$$\mathrm{p}(X|U) = \prod_{n=1}^{N} \mathrm{p}(\vec{x}_n|U) \prod_{k=1}^{n-1} \frac{\mathrm{p}(\vec{x}_n, \vec{x}_k|U)}{\mathrm{p}(\vec{x}_n|U)\mathrm{p}(\vec{x}_k|U)} \tag{35}$$

Typically the score of a hypothesized path is expressed in the log domain. In this case, it is straightforward to rewrite the expression as:

$$\log \mathrm{p}(X|U) = \left(\sum_{n=1}^{N} \log \mathrm{p}(\vec{x}_n|U) \right) + \left(\sum_{n=1}^{N} \sum_{k=1}^{n-1} \log \frac{\mathrm{p}(\vec{x}_n, \vec{x}_k|U)}{\mathrm{p}(\vec{x}_n|U)\mathrm{p}(\vec{x}_k|U)} \right) \tag{36}$$

In examining the final score of a hypothesized path using consistency modeling it can be seen that the consistency model contributes a sum of log ratios modeling individual pairs of acoustic observations. In information theory, this log ratio, computed for each pair of observations, is known as the pair's *mutual information*. Ideally, the log ratio for a pair of observations will contribute a positive score if the observations, given the hypothesized phonetic labels, are consistent with each other under the assumption that they were spoken by the same speaker. Likewise, negative scores would indicate that the observations, given the hypothesized phonetic labels, are not consistent with each other under the assumption they were spoken by the same speaker.

5.2 Engineering Issues

In order to utilize the consistency model framework in an actual speech recognition system, several engineering issues must be addressed. These issues are summarized by the following 5 questions:

1. How are the consistency model's joint probability density functions created?
2. What acoustic measurements should the consistency model utilize?
3. What phone pairs should be scored by the consistency model?
4. How should the consistency model be scaled relative to the standard acoustic model?
5. How can a recognizer's search mechanism incorporate consistency modeling?

5.2.1 Constructing Joint Density Functions

When utilized in a context independent mode, the consistency ratio is modeled utilizing the following expression:

$$\frac{\mathrm{p}(\vec{x}_j, \vec{x}_k | u_j, u_k)}{\mathrm{p}(\vec{x}_j | u_j)\mathrm{p}(\vec{x}_k | u_k)} \tag{37}$$

This expression requires the creation of a joint density function $\mathrm{p}(\vec{x}_j, \vec{x}_k | u_j, u_k)$. The independent density functions $\mathrm{p}(\vec{x}_j | u_j)$ and $\mathrm{p}(\vec{x}_k | u_k)$ are simply the marginal densities for $\vec{x}_j$ and $\vec{x}_k$ and can be extracted directly from $\mathrm{p}(\vec{x}_j, \vec{x}_k | u_j, u_k)$.

In order to train $\mathrm{p}(\vec{x}_j, \vec{x}_k | u_j, u_k)$ using standard methods, a set of joint observation vectors representing the observations of $\vec{x}_j$ and $\vec{x}_k$, as spoken by the same speaker, must be constructed. One potential method for creating joint vectors for a particular phone pair is by concatenating individual observation vectors from each of the two phones collected from one speaker. For example, suppose a training speaker has spoken 2 examples of the phone [s] and 3 examples of the phone [t]. The observation vectors for the [s] examples can be represented as $\vec{x}_{s,1}$ and $\vec{x}_{s,2}$. Likewise observation vectors for the [t] examples can be represented as $\vec{x}_{t,1}$, $\vec{x}_{t,2}$, and $\vec{x}_{t,3}$. From the examples of these two phones a set of joint observation vectors, $X_{s,t}$, for this one speaker can be created. If all combinations of the two phones are considered then six total joint observation vectors would be created. The joint vectors in the $X_{s,t}$ set would be represented as:

$$X_{s,t} = \left\{ \begin{bmatrix} \vec{x}_{s,1} \\ \vec{x}_{t,1} \end{bmatrix}, \begin{bmatrix} \vec{x}_{s,1} \\ \vec{x}_{t,2} \end{bmatrix}, \begin{bmatrix} \vec{x}_{s,1} \\ \vec{x}_{t,3} \end{bmatrix}, \begin{bmatrix} \vec{x}_{s,2} \\ \vec{x}_{t,1} \end{bmatrix}, \begin{bmatrix} \vec{x}_{s,2} \\ \vec{x}_{t,2} \end{bmatrix}, \begin{bmatrix} \vec{x}_{s,2} \\ \vec{x}_{t,3} \end{bmatrix} \right\} \tag{38}$$

This process of constructing joint vectors must then be repeated for the remaining training speakers in the training set. Figure 2 illustrates how the joint vectors from three different speakers can be created. In this figure each phone observation is represented by a single measurement, giving the joint phone vectors two dimensions. For this example, speaker 1 has two examples of [s] and three examples of [t]. Similarly, speaker 2 has four examples of [s] and two examples of [t], while speaker 3 has three examples each of [s] and [t].

It is worthwhile to note that the process of constructing the joint observation vectors need not be performed on a speaker by speaker basis. If one wishes to capture additional correlation information about factors which could vary from day to day (the speaker's health or environment) or even from utterance to utterance (the speaker's speaking style or speaking rate) then the joint vectors could be created on a session

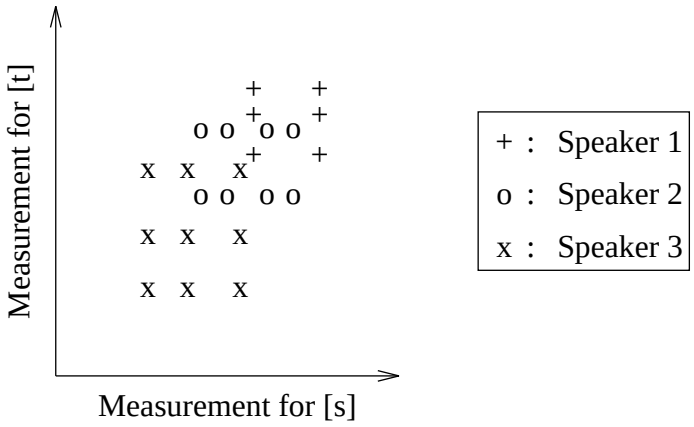

Figure 2: Illustration of joint vectors created for the pair of phones [s] and [t] as collected from three different training speakers.

by session or even an utterance by utterance basis. In our experiments, all utterances from a single training speaker were recorded in a single session using the same speaking style, thus justifying the speaker by speaker approach.

It is also important to note that the process described above for creating joint observation vectors for a single speaker results in a collection of joint observation vectors whose individual observation spaces (e.g., the separate observation spaces for [s] and [t] in the example above) are uncorrelated with each other. This can be observed visually in Figure 2 by noting from the geometric symmetry of the collection of joint observations created each for the three example speakers that the observation spaces of [s] and [t] are uncorrelated with each other for each individual speaker. This is consistent with the assumption discussed in Section 2 that different observations can be treated as independent when utilizing speaker dependent modeling. However, when the joint observations from all training speakers are combined, then the within-speaker correlations between [s] and [t] in the figure become evident.

There are various ways in which these joint vectors can be used to train a set of consistency models. The training method that proved most effective in our work is a technique we refer to as speaker mixture training. In this approach, a joint model is first created for each individual training speaker. Next, the final model is created by combining all of the individual joint models from each speaker into one large mixture model. In our experiments, the models from each individual speaker receive an equal weighting in the final mixture model. An equal weighting was used because each speaker had roughly the same amount of training data. However, in the general case different speaker models could receive different weighting factors in the final mixture

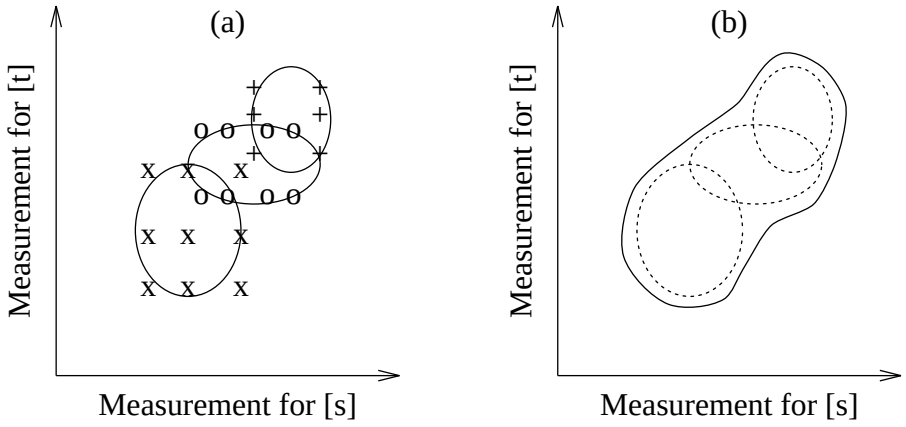

Figure 3: Illustration of joint models created for the pair of phones [s] and [t] as collected from three different training speakers. In (a) diagonal Gaussian models are created for each speaker. In (b) the individual diagonal Gaussians for each speaker are combined to make one large mixture of Gaussians.

model. In these experiments each individual speaker is modeled using only a single diagonal Gaussian density function. The use of a diagonal Gaussian is justified by the observations detailed in the previous paragraph. In experiments this method worked better than the more obvious method of training a model directly from the collection of joint observation vectors pooled over all speakers.

Formally, the training procedure used when creating the joint model for any particular phone pair is as follows:

1. Train a single diagonal Gaussian model from the collection of joint vectors for each individual training speaker.

2. Giving all training speakers equal weight, combine the diagonal Gaussians for the joint vectors from each training speaker into one large mixture Gaussian model.

This approach is illustrated by the example in Figure 3.

5.2.2 Measurement Selection

Because the consistency model score can be computed independently of the standard acoustic model score, the measurement sets used by the two different models need not

be the same. Because the consistency model is more difficult to train, a small set of measurements which exhibit a large amount of the correlation between phones may be more appropriate than the full set of measurements used by the standard acoustic model.

The recognizer used is these experiments utilizes 36 acoustic measurements in the standard acoustic model. These measurements are rotated using principal components analysis. In this work, the dimensionality of the acoustic measurement vectors used by the consistency model is reduced by using the top n principal components. Thus, the joint vector used by the consistency model would be of length $2n$. In our experiments, a value of $n = 10$ was found to work best. In [Hazen 1998] the primary principle components are shown to exhibit more correlation between phonetic units than the lesser principle components, thus justifying this approach.

5.2.3 Phone Pair Selection

The consistency model need not score all of the phone pairs that it encounters. Because creating robust consistency models is a difficult estimation problem, it is wise to score only the phonetic pairs which exhibit a high amount of within-speaker correlation in the training data. If two phones do not exhibit a high amount of correlation, the estimation noise inherent in the phone pair's model could be more significant than the actual information to be gained from the correlation between the two phones. In these cases it is wise to assume that these phone pairs are uncorrelated and not score them. Phone pairs that are not used contribute a score of zero to the final log score, the same score that truly uncorrelated pairs should contribute.

To decide which pairs the consistency model will score, two criteria are utilized. First, only pairs with high within-speaker correlation values will be scored. A method for estimating the within-speaker correlation of two phones is presented in detail in [Hazen 1998]. Second, only pairs with enough training data to sufficiently train a joint model will be used. For these experiments, phone pairs were eliminated from consideration by the consistency model if the training corpus contained less than 3000 joint vector exemplars of the pair in the training data.

The phone-pairs that have a suitable amount of training data are ranked by their within-speaker correlation values. In examining the ranked list, several patterns are obvious. The top of the list is dominated self pairs, vowel-vowel pairs and nasal-nasal pairs. Of the top 60 phone pairs, 36 are self pairs, 31 are vowel-vowel pairs, ten are fricative-fricative pairs, eight are nasal-nasal pairs, and only one is a stop-stop pair. Table 5.2.3 shows the top ten phone pairs as ranked by their estimated within-speaker

Rank	Phone Pair
1	[ŋ],[ŋ]
2	[š],[š]
3	[ɾ̃],[ɾ̃]
4	[ɑ^{y}],[ɑ^{y}]
5	[n],[n]
6	[o],[o]
7	[m],[m]
8	[ɾ],[ɾ]
9	[n],[ɾ̃]
10	[e],[e]

Table 1: Top ten phone pairs as ranked by the amount of their within-speaker correlation in the training data.

correlation. The list contains nine self pairs, with the final pair being the nasal [n] and it's flapped counterpart [ɾ̃]. Five of the pairs are nasal-nasal pairs indicating that nasals exhibit a large amount of within-speaker correlation. This is expected because the acoustic realization of nasals is dominated by the speaker's nasal cavity. The nasal cavity's physical characteristics typically undergo little to no variation during the course of a conversation, thus allowing different observations of the same nasal to be highly correlated. In our experiments, using only the top 60 phone-pairs in the consistency model was empirically found to work best.

5.2.4 Consistency Model Scaling

Experiments using the consistency model demonstrated the need for the consistency model score to be scaled relative to the score of the standard acoustic model. The scaling factor will be represented as κ. In our experiments a κ of around 0.2 was empirically found to work best. With the scaling factor the full acoustic model score is expressed as:

$$\log \mathrm{p}(X|U) = \left(\sum_{n=1}^{N} \log \mathrm{p}(\vec{x}_n|U) \right) + \kappa \left(\sum_{n=1}^{N} \sum_{k=1}^{n-1} \log \frac{\mathrm{p}(\vec{x}_n, \vec{x}_k|U)}{\mathrm{p}(\vec{x}_n|U)\mathrm{p}(\vec{x}_k|U)} \right) \tag{39}$$

5.2.5 Search Issues

As discussed earlier, when the utterance is processed in a time synchronous fashion, the acoustic model score for a particular segment is represented as:

$$\mathrm{p}(\vec{x}_n|\vec{x}_{n-1},\ldots,\vec{x}_1,U) = \mathrm{p}(\vec{x}_n|U)\frac{\mathrm{p}(\vec{x}_{n-1},\ldots,\vec{x}_1|\vec{x}_n,U)}{\mathrm{p}(\vec{x}_{n-1},\ldots,\vec{x}_1|U)} \tag{40}$$

From this equation it is clear that the score for a particular segment is dependent on all segment observations preceding it (as well as the segment labels U and the particular segmentation being considered). Because of this dependence on the full past context of the acoustic observations, the consistency model can not be incorporated into a standard Viterbi search. Furthermore, because the number of phones pairs that could be scored by the consistency model could be $O(n^2)$, it may be very inefficient to incorporate the consistency model into a best-first search such as the A^* search.

An alternative to incorporating the consistency model directly into an A^* search is to use an A^* search to generate an N-best list and then rescore the N-best hypotheses using the consistency model. This approach greatly reduces the amount of computation that would potentially be performed by an A^* search directly incorporating the consistency model. If the N-best list has a high probability of containing the correct answer then this approach is not likely to suffer any severe degradation in performance as compared to implementing an A^* search which utilizes the consistency model. In the case of the Resource Management task on which we conducted our experiments , the correct answer is one of the top two hypotheses 75% of the time and is one of the top ten hypotheses 90% percent of the time when the standard SI recognizer is used. For the experiments presented later in the paper, the consistency model is used to rescore the 10-best hypotheses proposed by the recognizer.

6 Results

The techniques discussed in this paper (hierarchical speaker clustering, speaker cluster weighting, reference speaker weighting, and consistency modeling) were evaluated using a word recognition task. The techniques were incorporated into the system for the purpose of performing rapid unsupervised speaker adaptation. In our experiments, the system attempts to adapt to the characteristics of the current speaker using the same utterance it is trying recognize. The corpus used for these experiments was the DARPA Resource Management corpus [Price 1988]. The experiments utilized the 109 speakers in the training and development sets for training purposes. The entire 40 speaker, 1200 utterance test set was used for testing. The SUMMIT system was used for recognition [Glass 1996]. The recognizer utilized segment-based, context-independent models for 68 different phonetic units. The standard word-pair grammar distributed with the corpus was used for the language model.

All of the techniques presented in this paper require a transcription of the adaptation data when performing adaptation. Unfortunately, the underlying transcription of an utterance is not known during unsupervised adaptation. The simplest solution to this problem is to run the standard SI recognizer on the adaptation data and then use the best path proposed by the recognizer as a substitute for the true transcription when performing adaptation. This approach can cause problems if the adaptation routine is sensitive to errors in the transcription. This is especially problematic for techniques which try to adapt a large number of specific parameters (such as the standard MAP adaptation algorithm) instead of a small number of general parameters (such as the RSW technique or MAP algorithms that incorporate shared parameter techniques [Kannan 1997, Shashahani 1997, Shinoda 1997, Zavaliagkos 1995c]). When adapting a small number of general parameters it is possible for the correct segments in the best path to overwhelm the errors during the adaptation routine's estimation phase. This is the case with the RSW and speaker cluster weighting techniques.

Figure 4 diagrams the system architecture used for the adaptation experiments presented in this chapter. The system uses a two-pass recognition approach. First, the SI recognizer is run to generate a best path. This best path is then utilized by the speaker cluster selection module. If hierarchical speaker clustering is being used then this module determines the gender and speaking rate of the utterance and outputs the appropriate gender and speaking rate dependent set of models. If speaker cluster weighting is being used then this module determines the optimal weighting of the different cluster models and outputs the final speaker cluster weighted set of models. The best path from the SI recognizer is also used by the RSW adaptation module.

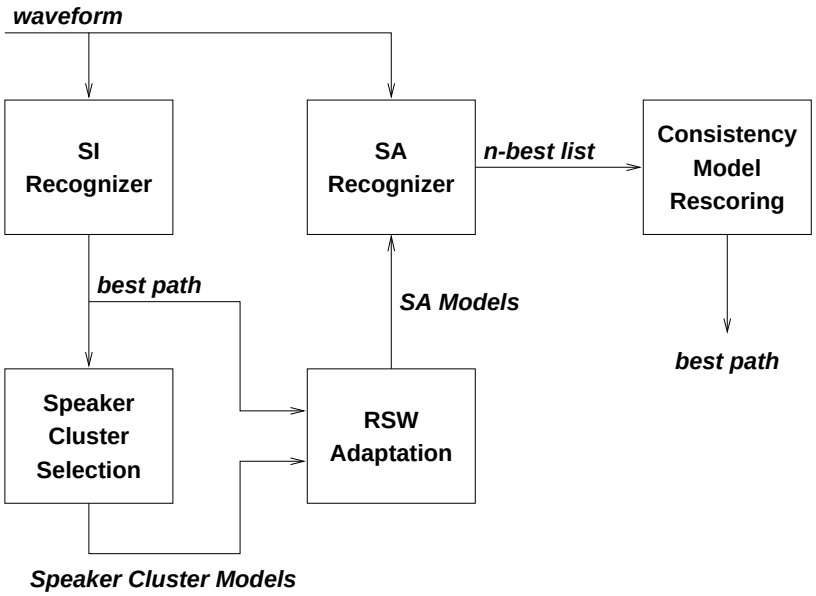

Figure 4: Architecture of recognizer using the rapid adaptation techniques described in the paper.

This module takes the set of models provided from the speaker clustering module and adapts them using RSW adaptation based on the best path provided by the SI recognizer. The RSW module outputs a speaker adapted (SA) set of models which can then be utilized for the second recognition pass. The SA recognizer is then used to generate an N-best list which can be rescored by the consistency model module.

Table 6 shows the recognition results using various combinations of the different adaptation algorithms. The table is broken down into three subsections corresponding to the three different speaker clustering gradations used: speaker independent (SI), gender dependent (GD), and gender and speaking rate dependent (GRD). The speaker clustering can also be augmented with the speaker cluster weighting (SCW) adaptation technique. For each type of speaker clustering, RSW adaptation and/or consistency modeling (CM) can be applied in addition to the speaker clustering. The type of hierarchical speaker clustering being used is listed in the first column. The second column contains the types of adaptation being utilized in addition to the speaker clustering. The next three columns show the total number of errors, the word error rate, and the reduction in word error rate relative to the performance of the speaker cluster models being used by the recognizer.

The most significant improvements in the system are gained by utilizing speaker

Exp. #	Initial Models	Adaptation Method	Word Error Rate	Total Errors	Error Reduction
1	SI	—	8.6%	882	—
2	SI	MAP	8.5%	875	0.8%
3	SI	RSW	8.0%	825	6.5%
4	SI	CM	7.9%	810	8.2%
5	SI	RSW + CM	7.9%	808	8.4%
6	GD	—	7.7%	789	—
7	GD	RSW	7.6%	783	0.8%
8	GD	CM	7.2%	738	6.5%
9	GRD	—	7.2%	737	—
10	GRD	CM	6.9%	715	3.0%
11	GRD	SCW	6.9%	715	3.0%
12	GRD	SCW + CM	6.8%	701	4.9%

Table 2: Table of recognition results using various forms of rapid, unsupervised adaptation, where the adaptation is performed on the same utterance the system is trying to recognize.

cluster models instead of standard SI models. This can be seen in the table as the error rates are reduced as the specificity of the clusters models increases from the SI models (exp. 1), to the GD models (exp. 6), to the GRD models (exp. 9). The error rate reduction from the SI models to the GD models was 10.5% while the the error rate reduction from the SI models to GRD models was 16.4%. These results indicate that large improvements in recognition accuracy can be gained simply by adapting to generic speaker properties such as gender and speaking rate. Note that the SCW adaptation technique further improves the results obtained with hierarchical speaker cluster modeling (exp. 11). This indicates that it is better to let the system make a *soft* decision about the characteristics of speaker (as is done in SCW) than to force the system to make a *hard* decision about what speaker cluster model to use (as in done in standard hierarchical speaker clustering).

When examining the different adaptation techniques applied to the SI model set, the first adaptation result is from the application of standard MAP model translation to the SI recognizer (exp 2.). Past results have indicated that standard MAP adaptation techniques, though based on solid mathematical principles, are slow to adapt to a new speaker and are better suited for long term adaptation [Zavaliagkos 1995a]. As expected, when MAP adaptation is incorporated into our adaptation system, it did not significantly improve the recognizer performance. It should be noted that we

did not attempt to incorporate MLLR adaptation into our system because past research efforts have shown that MLLR also performs poorly when only small amounts of adaptation data (three utterances or less) is available [Leggetter 1995]. We also did not attempt to duplicate any form of EMAP adaptation because past efforts have required fairly complex modeling techniques while yielding results only marginally better then MAP adaptation [Huo 1997, Zavaliagkos 1995b]. Since the completion of the experiments in this paper, several promising techniques, in which parameter sharing techniques have been incorporated in a MAP adaptation algorithm, have also been published [Kannan 1997, Shashahani 1997, Shinoda 1997]. Comparison against these techniques were not possible when the experiments in this paper were originally conducted.

Next, the table shows that RSW model translation does improve the performance of the SI system significantly despite the fact that its adaptation is guided by the error prone best path from the SI recognizer (exp. 3). This indicates that RSW model translation adaptation is far more robust to errors in the recognizer's best path and adapts more rapidly than MAP model translation. However, when RSW adaptation is performed on the GD cluster models, no significant improvement is observed (exp. 7). There are two possible explanations for this. First, the GD models have a smaller variance than the SI models and, as such, their likelihood estimates are affected more when their centers of mass are altered than models with larger variance. Thus, as the cluster models become more specific, model translation adaptation techniques become more sensitive to the noise in the center of mass estimation. Second, much of the gain of RSW adaptation might be due to the techniques adaptation to the gender, and not the specific acoustic characteristics, of the current speaker. Because of this result we did not attempt to use RSW to adapt the GRD models.

When consistency modeling is used, the system's performance is almost universally improved regardless of the models that they are used in conjunction with. It should be noted that the relative improvements from consistency modeling decrease as the cluster models become more specific. The improvements are significant when CM is used in conjunction with the SI and GD models (exp. 4 and exp. 8). However, the improvement is only marginally significant when CM is applied with the GRD models (exp. 10). The reduced effectiveness of the consistency modeling approach as the speaker cluster models get more specific are expected because the contribution of the consistency model should decrease as the resemblance of the standard acoustic models to the true underlying speaker dependent models increases.

When examining the results obtained using consistency model, one might wonder how the consistency modeling approach compares with approaches which attempt to model the correlations of successive observations [Paliwal 1993, Szarvas 1998]. It is

easily reasoned that successive frames of sampled speech are highly correlated because the physical limitations and inertia of a speaker's articulatory mechanisms typically constrain the acoustic characteristics of successive frames of speech to be highly similar. Though the rational for employing this approach is different than the rational for consistency modeling, these techniques do share to common idea of jointly modeling two observations in order to condition the likelihood of one observation on a previous observation. Thus, one might wonder how many of the consistency model phone pairs utilized in a typical utterance are successive observations and what percentage of any improved recognition results is the result of scoring these successive pairs. In our experiments the scoring of successive pairs of observation with the consistency model was actually very uncommon. This is because a majority of the consistency pairs are self-pairs and our system's phonological component rarely allows the same phonetic event to occur twice in a row (because it typically treats sequences of the same fricative or nasal as a single geminate unit, and other duplicate phonetic sequences are either exceptionally rare or cannot happen by rule). As a result, preventing successive pairs from being scored in the consistency model approach used here has no significant effect on the performance of the system. Hence, it is conceivable that incorporating common successive observation pairs into the modeling to accompany the consistency pairs determined to have high within-speaker correlation could further improve upon the results obtained here.

7 Discussion

The experiments presented in this paper have shown the importance of incorporating within-speaker correlation information into a system performing rapid speaker adaptation. By accounting for these correlations using the speaker clustering adaptation methods, models which more closely resemble the current speaker can be quickly constructed using only one adaptation utterance. Furthermore, it was found that mistakes in hypotheses, which were likely caused because the system did not enforce any speaker constraint within its framework, could be corrected by enforcing the speaker constraint with the consistency model. Overall, combinations of the various adaptation techniques described in this paper reduced the error rate of our system by 4.9% to 8.4% depending on the initial speaker cluster models being used. When combining speaker clustering techniques with the rapid adaptation techniques presented in this paper, an overall relative error rate reduction of 20% from the baseline SI system was achieved. Most of the 20% error rate reduction can be attributed to utilizing gender and speaking rate dependent models. However, it was observed that the use of the consistency model improved all versions of our system including the gender and speaking rate dependent version. This indicates that additional information beyond gender and speaking rate is being provided by the consistency model.

It is our belief that the formulation of the *consistency model* technique is an important step forward in the development of our speaker independent recognition system. With this model we are attacking the segment independence assumption, which has long been considered a weak link in the mathematical formulation of typical speech recognition systems. Though the modeling techniques employed in the creation of the consistency models used in this paper are simplistic, the system obtained significant reductions in error rate when these models were used. We believe that further study of the consistency model approach will yield a better understanding of the within-speaker correlation information which the model is attempting the capture, hopefully resulting in further improvements in our system's performance.

It must be stated that we acknowledge that the true value of the adaptation techniques presented in this paper will not be known until the techniques can be tested on a state-of-the-art recognizer. The first step in achieving this is to incorporate the techniques presented in this paper into a context-dependent system which is closer to the state-of-the-art in recognition performance than the context independent recognizer utilized in this paper. We hope to attempt this in the future. At this time we do not have any preconceptions about how well these techniques will scale to a context-dependent large vocabulary system. However it is our hope that, like MAP and MLLR, these techniques can be engineered to produce significant improvements

in performance in a state-of-the-art system.

We also hope to incorporate some of the ideas presented in this paper into our real world spoken language understanding systems such as the JUPITER system [Zue 1997]. These systems must handle short conversations (typically 5 turns or less) which contain spontaneous, telephone speech from a wide variety of speakers, telephone types (speaker phones, cell phones, etc.) and channel qualities. Under these circumstances the ability for a recognizer to produce hypotheses which are consistent across the length of the utterance will be strained and methods for rapid adaptation could prove extremely helpful.

Aknowledgements

This research was supported by DARPA under Contract N66001-94-C-6040, monitored through the Naval Command, Control and Ocean Surveillance Center. The author also wishes to thank the ATR Intrepreting Telephony Laboratory for hosting the author for a three month span in 1995 during which the author was given the freedom to investigate some of the early ideas which led to the research presented in this paper. Finally, the author wishes to thank Dr. James Glass who served as the the PhD. supervisor for the work presented in this paper and was the co-author of the Eurospeech97 paper that this paper expands upon.

References

[Bahl 1995] L. Bahl, *et al.* (1995), "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task," In *Proceedings of the 1995 International Conference on Acoustics, Speech and Signal Processing*, Detroit, 9–12 May 1995, IEEE, Piscataway, pp. 41–44.

[Bahl 1991] L. Bahl, *et al.* (1991), "A fast algorithm for deleted interpolation," In *Proceedings of the 2nd European Conference on Speech Communication and Technology*, Genova, 24–26 September 1991, ESCA, Grenoble, pp. 1209–1212,.

[Bahl 1983] L. Bahl, F. Jelinek, and R. Mercer (1984), "A maximum likelihood approach to continuous speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume PAMI-5, Issue 2, March 1984, pp. 179–190.

[Baker 1975] J. Baker (1975), *Stochastic Modeling as a Means of Automatic Speech Recognition.* PhD thesis, Carnegie Mellon University.

[Furui 1989] S. Furui (1989), "Unsupervised speaker adaptation method based on hierarchical spectral clustering," In *Proceedings of the 1989 International Conference on Acoustics, Speech and Signal Processing*, Glasgow, 23–26 May 1989, IEEE, Piscataway, pp. 286–289.

[Gauvain 1995] J. Gauvain, L. Lamel, and M. Adda-Decker (1995), "Developments in continuous speech dictation using the ARPA WSJ task," In *Proceedings of the 1995 International Conference on Acoustics, Speech and Signal Processing*, Detroit, 9–12 May 1995, IEEE, Piscataway, pp. 65–68.

[Gauvain 1994] J. Gauvain and C. Lee (1994), "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chains," *IEEE Transactions on Speech and Audio Processing*, Volume 2, Issue 2, April 1994, pp. 291–298.

[Glass 1996] J. Glass, J. Chang, and M. McCandless (1996), "A probabilistic framework for feature-based speech recognition," In *Proceedings of the 1996 International Conference on Spoken Language*

Processing, Philadelphia, 3–6 October 1996, IEEE, Piscataway, pp. 2277–2280.

[Hazen 1998] T. Hazen (1998), *The Use of Speaker Correlation Information for Automatic Speech Recognition.* PhD thesis, Massachusetts Institute of Technology.

[Huang 1996] X. Huang, *et al.* (1996), "Deleted interpolation and density sharing for continuous hidden Markov models," In *Proceedings of the 1996 International Conference on Acoustics, Speech and Signal Processing*, Atlanta, 7–10 May 1996, IEEE, Piscataway, pp. 885–888.

[Huo 1997] Q. Huo and C. Lee (1997), "Combined on-line model adaptation and Bayesian predictive classification for robust speecg recognition," In *Proceedings of the 5th European Conference on Speech Communication and Technology*, Rhodes, 22–25 September 1997, ESCA, Grenoble, pp. 1847–1850.

[Kannan 1997] A. Kannan and M. Ostendorf (1997), "Modeling dependency in adaptation of acoustic models using multiscale tree processes," In *Proceedings of the 5th European Conference on Speech Communication and Technology*, Rhodes, 22–25 September 1997, ESCA, Grenoble, pp. 1863–1866.

[Kosaka 1994a] T. Kosaka, S. Matsunaga, and S. Sagayama (1994), "Tree-structured speaker clustering for speaker-independent continuous speech recognition," In *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, 18–22 September 1994, ASJ, Tokyo, pp. 1375–1378.

[Kosaka 1994b] T. Kosaka and S. Sagayama (1994), "Tree-structured speaker clustering for fast speaker adaptation," In *Proceedings of the 1994 International Conference on Acoustics, Speech and Signal Processing*, Adelaide, 19–22 April 1994, IEEE, Piscataway, Volume I, pp. 245–248.

[Kubala 1997] F. Kubala, *et al.* (1997), "Advances in transcription of broadcast news," In *Proceedings of the 5th European Conference on Speech Communication and Technology*, Rhodes, 22–25 September 1997, ESCA, Grenoble, pp. 927–930.

[Kuhn 1998] R. Kuhn, *et al.* (1998), "Eigenvoices for speaker adaptation," In *Proceedings of the 1998 International Conference on Spoken Language Processing*, Sydney, 30 November–4 December 1998, ASSTA, Canberra, pp. 1771–1774.

[Lasry 1984] M. Lasry and R. Stern (1984), "*A posteriori* estimation of correlated jointly Gaussian mean vectors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume PAMI-6, Issue 4, July 1984, pp. 530–535.

[Lee 1988] K. Lee (1988), *Large Vocabulary Speaker-Independent Continuous Speech Recognition: The Development of the SPHINX System.* PhD thesis, Carnegie Mellon University.

[Leggetter 1995] C. Leggetter and P. Woodland (1995), "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, Volume 9, Issue 2, April 1995, pp. 171–185.

[Mathan 1990] L. Mathan and L. Miclet (1990), "Speaker hierarchical clustering for improving speaker-independent HMM word recognition," In *Proceedings of the 1990 International Conference on Acoustics, Speech and Signal Processing*, Albuquerque, 3–6 April 1990, IEEE, Piscataway, pp. 149–152.

[Paliwal 1993] K. Paliwal (1993), "Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer," In *Proceedings of the 1993 International Conference on Acoustics, Speech and Signal Processing*, Minneapolis, 27–30 April 1993, IEEE, Piscataway, Volume II, pp. 215–218.

[Price 1988] P. Price, *et al.* (1988), "The DARPA 1000-word Resource Management database for continuous speech recognition," In *Proceedings of the 1988 International Conference on Acoustics, Speech and Signal Processing*, New York, 11–14 April 1988, IEEE, Piscataway, pp. 651–654,.

[Shashahani 1997] B. Shahshahani (1997), "A Markov random field approach to Bayesian speaker adaptation," *IEEE Transactions on Speech and Audio Processing*, Volume 5, Issue 2, March 1997, pp. 183–191.

[Shinoda 1997] K. Shinoda and C. Lee (1992), "Structural MAP speaker adaptation using hierarchical priors," In *Proceedings of the 1997*

IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara, 14–17 December 1997, IEEE, Piscataway, pp. 381–387.

[Szarvas 1998] M. Szarvas and S. Matsunaga (1998), "Acoustic observation context modeling in segment based speech recognition," In *Proceedings of the 1998 International Conference on Spoken Language Processing*, Sydney, 30 November–4 December 1998, ASSTA, Canberra, pp. 2967–2970.

[Zavaliagkos 1995a] G. Zavaliagkos, R. Schwartz, and J. Makhoul (1995), "Adaptation algorithms for BBN's phonetically tied mixture system," In *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, Austin, 22–25 January 1995, Morgan Kaufmann, San Francisco, pp. 82–87.

[Zavaliagkos 1995b] G. Zavaliagkos, R. Schwartz, and J. Makhoul (1995), "Batch, incremental and instantaneous adaptation techniques for speech recognition," In *Proceedings of the 1995 International Conference on Acoustics, Speech and Signal Processing*, Detroit, 9–12 May 1995, IEEE, Piscataway, 1995, pp. 676–679.

[Zavaliagkos 1995c] G. Zavaliagkos, R. Schwartz, and J. Makhoul (1995), "Speaker adaptation using a predictive model," In *Proceedings of the 4rd European Conference on Speech Communication and Technology*, Madrid, 18–21 September 1995, ESCA, Grenoble, pp. 1131–1135.

[Zue 1997] V. Zue, *et al.* (1997), "From interface to content: translingual access and delivery of on-line information," In *Proceedings of the 5th European Conference on Speech Communication and Technology*, Rhodes, 22–25 September 1997, ESCA, Grenoble, pp. 2227–2230.