A COMPARISON OF NOVEL TECHNIQUES FOR INSTANTANEOUS SPEAKER ADAPTATION[1]

Timothy J. Hazen and James R. Glass

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

ABSTRACT

This paper introduces two novel techniques for instantaneous speaker adaptation, *reference speaker weighting* and *consistency modeling*. An approach to hierarchical speaker clustering using gender and speaking rate as the clustering criteria is also presented. All three methods attempt 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. In instantaneous adaptation experiments using the DARPA Resource Management corpus, a reduction in word error rate of 20% has been achieved using a combination of these new techniques.

INTRODUCTION

Speaker adaptation can be viewed as the task of altering the acoustic models of a speech recognition system to match, as closely as possible, the current speaker. Reliable methods exist for performing adaptation when a large amount of adaptation data is available. In particular, a solid mathematical formulation for the maximum *a posteriori* probability (MAP) adaptation of mixture Gaussian model parameters has been derived and algorithms using this approach have been developed and refined [1]. Unfortunately, despite their solid mathematical bases, these standard methods exhibit slow adaptation rates when the amount of adaptation data is limited. To address this problem this paper focuses on the issues of rapid or instantaneous speaker adaptation. Specifically, this paper introduces two novel approaches to instantaneous speaker adaptation, *reference speaker weighting* (RSW) and *consistency modeling*. Additionally this paper presents an approach to supervised hierarchical speaker clustering using gender and speaking rate as the clustering criteria.

In standard speaker independent (SI) training, a phone's acoustic model is trained by pooling together all observations of that particular phone from all training speakers and then estimating the parameters of the acoustic model from the entire pool of observations. Using this approach, the acoustic models are typically heterogeneous and high in variance. The actual acoustic space that the speech of one particular speaker may occupy is typically only a fraction of the space occupied by the entire acoustic model. Furthermore, correlations exist between the acoustic realizations of different phones spoken by the same speaker. These correlations are very strong in some cases and force the acoustic realizations of different phones from an individual speaker to be jointly constrained to specific regions of the entire acoustic space. The difficulties caused by the heterogeneity of the SI models are compounded by the fact that speech recognition systems typically assume that all acoustic observations are independent of each other. To avoid these problems, the goal of rapid speaker adaptation is to quickly focus a speech recognizer's acoustic models onto the unknown yet constrained acoustic space occupied by the current speaker.

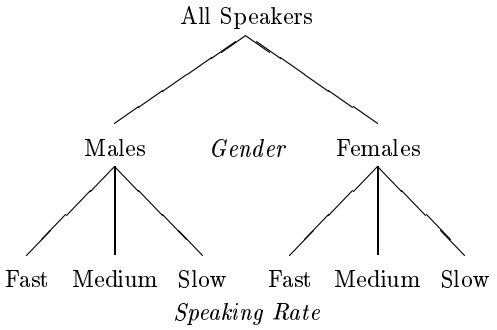

Figure 1: Hierarchical cluster tree utilized by our system.

SPEAKER CLUSTERING

One method of speaker adaptation that has proven successful is hierarchical speaker clustering [2]. Hierarchical speaker clustering allows similar training speakers to be clustered to create models which represent specific speaker types. 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.

Recognition using the speaker clustered models is performed with a two-pass strategy. First, the test utterance is passed through the speaker independent 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. Recognition using only the male and female clustered models is also possible.

Because clustering reduces the amount of data in the tree as the clustering becomes more and more specific, the estimation of the model parameters may suffer from sparse data problems. To increase the robustness of the

[1]This research was supported by DARPA under Contract N66001-94-C-6040, monitored through Naval Command, Control, and Ocean Surveillance Center.

models in the tree, model interpolation is utilized. The final interpolated acoustic models used for each gender dependent phone model, $p_{igd}(\vec{a} \mid p)$, are an interpolation of the maximum likelihood trained gender dependent model, $p_{gd}(\vec{a} \mid p)$ and the speaker independent model, $p_{si}(\vec{a} \mid p)$. The form of this interpolation is:

$$p_{igd}(\vec{a} \mid p) = \lambda p_{gd}(\vec{a} \mid p) + (1 - \lambda) p_{si}(\vec{a} \mid p) \quad (1)$$

The value of λ is determined from the training data using deleted interpolation [3]. Similarly, the final gender and rate specific acoustic models for each phone are an interpolation of the maximum likelihood trained gender and rate specific model, the maximum likelihood trained gender specific model and the speaker independent model.

REFERENCE SPEAKER WEIGHTING

Reference speaking weighting (RSW) techniques, like speaker clustering techniques, utilize a final model constructed from the training speakers most similar to the test speaker. RSW techniques, however, allow the final model to assign varying degrees of weight to each training or *reference speaker* utilized in the model. This differs from hierarchical speaker clustering, in which each training speaker used in a cluster receives an equal weight while all speakers not in the cluster receive weights of zero.

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:

$$\vec{c} = \sum_{i=1}^{M} \omega_i \vec{\mu}_i \quad (2)$$

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:

$$\vec{\mu}_i = \vec{c} + \vec{\nu}_i \quad (3)$$

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 new 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 for each of P different phonetic classes has been obtained. 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:

$$\vec{m}_r = \begin{bmatrix} \vec{c}_{1,r} \\ \vec{c}_{2,r} \\ \vdots \\ \vec{c}_{P,r} \end{bmatrix} \quad (4)$$

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

$$\mathbf{M} = [\, \vec{m}_1 \,;\, \vec{m}_2 \,;\, \ldots \,;\, \vec{m}_R \,] \quad (5)$$

During rapid speaker adaptation, the goal is to determine the most likely speaker vector, $\vec{m}$, for a test speaker given the speaker's adaptation data. With only a small amount of data, it is likely that no observations exist for many phones. In this case standard MAP estimation does not provide any means of estimating the components of $\vec{m}$ for these phones. One solution to this problem is to utilize RSW on the speaker vectors in $\mathbf{M}$ to constrain the speaker space in which $\vec{m}$ may fall. Specifically, the value of $\vec{m}$ is constrained to be a weighted average of the speaker vectors contained in $\mathbf{M}$. This can be expressed as:

$$\vec{m} = \mathbf{M}\vec{w} \quad (6)$$

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 class p can be expressed as $\vec{c}_p$ and $\mathbf{M}_p$, thus allowing the following expression:

$$\vec{c}_p = \mathbf{M}_p\vec{w} \quad (7)$$

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:

$$\mathcal{X} = \{\, X_1,\, X_2,\, \ldots,\, X_P \,\} \quad (8)$$

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

$$X_p = \{\, \vec{x}_{p,1},\, \vec{x}_{p,2},\, \ldots,\, \vec{x}_{p,N_p} \,\} \quad (9)$$

Here each $\vec{x}_{p,n}$ is a specific observation vector of class p and N_p is the total number of adaptation observations available for class p. Note that it is possible for N_p to be zero for any given class, 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):

$$\arg\max_{\vec{w}} \log p(\mathcal{X} \mid \vec{w}). \quad (10)$$

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

$$\arg\max_{\vec{w}} \sum_{p=1}^{P} \sum_{n=1}^{N_p} \log p(\vec{x}_{p,n} \mid \vec{w}). \quad (11)$$

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 class model. The density function for phone p can thus be expressed as:

$$\mathrm{p}(\vec{x}_{p,n}|\vec{w}) \equiv \mathcal{N}(\vec{c}_p, \mathbf{S}_p) \tag{12}$$

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

It can be shown that the expression in (11) reduces to the following expression:

$$\arg\max_{\vec{w}} \; 2\vec{v}^{\,T}\vec{w} - \vec{w}^{\,T}\mathbf{U}\vec{w}. \tag{13}$$

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

$$\mathbf{U} = \sum_{p=1}^{P}\sum_{n=1}^{N_p} \mathbf{M}_p^T\mathbf{S}_p^{-1}\mathbf{M}_p = \sum_{p=1}^{P} N_p\mathbf{M}_p^T\mathbf{S}_p^{-1}\mathbf{M}_p \tag{14}$$

$$\vec{v}^{\,T} = \sum_{p=1}^{P}\sum_{n=1}^{N_p} \vec{x}_{p,n}^{\,T}\mathbf{S}_p^{-1}\mathbf{M}_p \tag{15}$$

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

$$\forall i \; w_i \geq 0 \;\text{ and }\; \sum_{i=0}^{R} w_i = 1 \tag{16}$$

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.

CONSISTENCY MODELING

Theoretical Framework

Consistency modeling is a novel modeling technique which attempts to utilize the correlation information between acoustic segments which is ignored when these acoustic segments are considered independent. This technique is discussed here as a form of instantaneous speaker adaptation. However, consistency modeling can also be viewed as a new modeling technique for speaker independent recognition.

To explain this technique, consider the task of classifying a sequence of N segments. In a segment-based approach a measurement vector is created for each potential segment from the underlying acoustic information. The sequence of measurement vectors for a particular set of N segments can be represented as:

$$A = \{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_N\} \tag{17}$$

For each particular set of N segments, a string of N phones can be hypothesized. This string can be represented as:

$$P = \{p_1, p_2, \ldots, p_N\} \tag{18}$$

Given a particular set of segments, the goal is to find the most likely string of phones. This is represented as:

$$\arg\max_{P} \; \mathrm{p}(P \,|\, A) \tag{19}$$

This expression is equivalently written as:

$$\arg\max_{P} \; \mathrm{p}(A \,|\, P)\,\mathrm{p}(P) \tag{20}$$

The expression $\mathrm{p}(A \,|\, P)$ is typically referred to as the acoustic model of an automatic speech recognition system. This model can be expanded as follows:

$$\begin{aligned} \mathrm{p}(A \,|\, P) &= \mathrm{p}(\vec{a}_N, \vec{a}_{N-1}, \ldots, \vec{a}_1 \,|\, P) && (21)\\ &= \prod_{j=1}^{N} \mathrm{p}(\vec{a}_j \,|\, \vec{a}_{j-1}, \ldots, \vec{a}_1, P) && (22) \end{aligned}$$

At this point, typical speech recognition systems assume that the segments are independent of each other on the acoustic level. This assumption allows the acoustic model to be simplified as follows:

$$\prod_{j=1}^{N} \mathrm{p}(\vec{a}_j \,|\, \vec{a}_{j-1}, \ldots, \vec{a}_1, P) = \prod_{j=1}^{N} \mathrm{p}(\vec{a}_j \,|\, P) \tag{23}$$

We wish to avoid making the segment independence assumption. To begin, consider the right hand side of Equation (22). Bayes' rule can be used to rewrite the probability terms in this expression as follows:

$$\mathrm{p}(\vec{a}_j \,|\, \vec{a}_{j-1}, \ldots, \vec{a}_1, P) = \mathrm{p}(\vec{a}_j \,|\, P)\frac{\mathrm{p}(\vec{a}_{j-1}, \ldots, \vec{a}_1 \,|\, \vec{a}_j, P)}{\mathrm{p}(\vec{a}_{j-1}, \ldots, \vec{a}_1 \,|\, P)} \tag{24}$$

In this form, the original probability term can be viewed as the product of two separate terms. The first term is the *standard acoustic model* for the phone when it is considered independently from all other phones. The second term is a ratio which we will refer to as the *consistency model*. This ratio compares the likelihood of the previously observed phones when considering and not considering the latest observation. If the current phone observation is consistent with the previous observations, this ratio becomes greater than 1 and increases the overall score of the current hypothesized path. Similarly, if the ratio is less than 1 then the current phone hypothesis is considered *inconsistent* with the previous phone hypotheses.

Now that the consistency model is defined the difficultly lies in developing ways to estimate this ratio. Modeling a large joint expression such as $\mathrm{p}(\vec{a}_{j-1}, \ldots, \vec{a}_1 \,|\, P)$ would be extremely difficult with anything but the simplest probabilistic models. For the purpose of practicality, one simplifying assumption will be made. It will be assumed that the following approximation can be made:

$$\frac{\mathrm{p}(\vec{a}_{j-1}, \ldots, \vec{a}_1 \,|\, \vec{a}_j, P)}{\mathrm{p}(\vec{a}_{j-1}, \ldots, \vec{a}_1 \,|\, P)} \approx \prod_{k=1}^{j-1} \frac{\mathrm{p}(\vec{a}_k \,|\, \vec{a}_j, P)}{\mathrm{p}(\vec{a}_k \,|\, P)} \tag{25}$$

This expression can be equivalently expressed as:

$$\prod_{k=1}^{j-1} \frac{\mathrm{p}(\vec{a}_k \,|\, \vec{a}_j, P)}{\mathrm{p}(\vec{a}_k \,|\, P)} = \prod_{k=1}^{j-1} \frac{\mathrm{p}(\vec{a}_j, \vec{a}_k \,|\, P)}{\mathrm{p}(\vec{a}_j \,|\, P)\,\mathrm{p}(\vec{a}_k \,|\, P)} \tag{26}$$

From here it is easy to show that the full score for a hypothesized path can be written in the log domain as:

$$\left(\sum_{j=1}^{n} \log \mathrm{p}(\vec{a}_j \,|\, P)\right) + \left(\sum_{j=1}^{n}\sum_{k=1}^{j-1} \log \frac{\mathrm{p}(\vec{a}_j, \vec{a}_k \,|\, P)}{\mathrm{p}(\vec{a}_j \,|\, P)\,\mathrm{p}(\vec{a}_k \,|\, P)}\right) \tag{27}$$

Note that the consistency model score and the standard acoustic model score are captured in separate terms. Also note that the log ratio for each phone pair in the consistency model is referred to as the pair's *mutual information* in information theory.

Constructing the Consistency Models

The most important issue in using the consistency modeling technique is the construction of the joint models. In a context independent mode, the consistency model utilizes the following expression:

$$\frac{\mathrm{p}(\vec{a}_j, \vec{a}_k \,|\, p_j, p_k)}{\mathrm{p}(\vec{a}_j \,|\, p_j)\,\mathrm{p}(\vec{a}_k \,|\, p_k)} \tag{28}$$

This expression requires the creation of a joint density function $p(\vec{a}_j, \vec{a}_k \mid p_j, p_k)$. The independent density functions $p(\vec{a}_j \mid p_j)$ and $p(\vec{a}_k \mid p_k)$ are simply the marginal densities for $\vec{a}_j$ and $\vec{a}_k$ as extracted from $p(\vec{a}_j, \vec{a}_k \mid p_j, p_k)$.

For our experiments the joint models are constructed in the following fashion for any given phone pair:

1. Train a standard single diagonal Gaussian model of the acoustic measurements for each phone in the pair for each speaker in the training set.
2. For each training speaker concatenate the diagonal Gaussians from each of the two phones into one joint diagonal Gaussian.
3. Giving all training speakers equal weight, combine the joint diagonal Gaussians from each training speaker into one large mixture Gaussian model.

Because the consistency model can be completely separated from the original acoustic model, the consistency model need not use an identical set of acoustic measurements as the acoustic model. To increase robustness, the consistency model in our system only uses the first 10 principal components of the 36 measurements used by the standard acoustic model.

The consistency model does not need to utilize all phone pairs during its scoring. Because the consistency model's score is a log ratio, a phone pair that is not used simply contributes a score of zero to the final score. Because the consistency model may not be as robustly trained as the standard acoustic model, it is wise to use only the phone pairs which exhibit the most within-speaker correlation. In our experiments we use only 63 phone pair models in the consistency model, 40 of which are self-pairs.

Incorporating the Consistency Model

Because the consistency model is not trained as robustly as the standard acoustic model it is wise to scale its score relative to the standard acoustic model score. In our experiments, a scale factor of .2 is typically used. Experiments have shown that varying this scale factor by as much as 50% has only marginal effects on the final accuracy of the system, although larger changes in its value do begin to degrade the final performance. This scale factor should be set automatically by maximizing performance on either a separate development test set or data jackknifed from the training set.

Our system utilizes a two step search process when incorporating the consistency model. First an N-best list is generated using only the standard acoustic models. The top N hypotheses are then rescored using the consistency model. The value of N was set to 10 in our experiments.

EXPERIMENTAL RESULTS

Our various instantaneous adaptation methods were tested on the DARPA Resource Management (RM) corpus on the task of word recognition [4]. The experiments utilized the 109 speakers in the training and development sets for training purposes. The entire 1200 utterance test set was used for testing. Each test utterance was independently used for simultaneous adaptation and word recognition. Recognition was performed by the SUMMIT recognition system using the standard RM word pair grammar and context independent acoustic models [5].

Our system was initially tested with standard speaker independent models (SI), gender dependent (GD) models, and gender and speaking rate dependent (GRD) models. Each of these models could be further adapted using any of the following methods: (1) unsupervised maximum *a posteriori* probability (MAP) model translation, (2) unsupervised RSW model translation, (3) consistency modeling (CM), or (4) a combination of any of the first three methods. A summary of the results under various conditions is presented in Table 1.

Adaptation Method	Word Error Rate	Total Errors	Error Reduction
SI	8.6%	884	- - - -
SI + MAP	8.5%	877	0.8%
SI + RSW	8.0%	826	6.6%
SI + CM	7.9%	812	8.1%
SI + RSW + CM	7.7%	793	10.3%
GD	7.7%	791	10.5%
GD + RSW	7.6%	785	11.2%
GD + CM	7.1%	730	17.4%
GRD	7.2%	739	16.4%
GRD + CM	6.9%	706	20.1%

Table 1: Table of instantaneous adaptation results.

DISCUSSION

Our experiments have shown the importance of incorporating within-speaker correlation information into a system performing instantaneous speaker adaptation. Our results indicate that information about the gender and speaking rate of a speaker accounts for a large amount of the error reduction observed by our system. It can also be 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 sustained 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.

REFERENCES

[1] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, 291-298, April 1994.

[2] T. Kosaka, S. Matsunaga, and S. Sagayama, "Tree-structured speaker clustering for speaker-independent continuous speech recognition," in *Proc. ICSLP* (Yokohama), 1375–1378, 1994.

[3] X. D. Huang, M.-Y. Hwang, L. Jiang, and M. Mahajan, "Deleted interpolation and density sharing for continuous hidden Markov models," in *Proc. ICASSP* (Atlanta), 885–888, 1996.

[4] P. Price, et al, "The DARPA 1000-word Resource Management database for continuous speech recognition," in *Proc. ICASSP* (New York), 651–654, 1988.

[5] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. ICSLP* (Philadelphia), 2277-2280, 1996.