RAPID SPEAKER ADAPTATION USING SPEAKER CLUSTERING

Ernest J. Pusateri and Timothy J. Hazen

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

ABSTRACT

This paper examines an approach to speaker adaptation called speaker cluster weighting (SCW) for rapid adaptation in the Jupiter weather information system. SCW extends the ideas of previous speaker cluster techniques by allowing the speaker cluster models (learned from training data) to be adaptively weighted to match the current speaker. We explore strategies for automatic speaker clustering as well as cluster model training procedures for use with this algorithm. As part of this exploration, we develop a novel algorithm called least squares linear regression (LSLR) clustering for the clustering of speakers for whom only a small amount of data is available.

1. INTRODUCTION

Even with recent advances in speaker independent (SI) speech recognition systems, SI systems still perform considerably worse than their speaker dependent (SD) counterparts. For example, SD systems can achieve word error rates that are 50% lower than those achieved by SI recognizers [1]. Speaker adaptation algorithms attempt to narrow this performance gap.

The task of speaker adaptation is especially difficult in contexts where a very small amount ($<$10 seconds) of adaptation data is available, and where an accurate transcription of the adaptation data is not provided. Under these conditions adaptation is referred to as rapid and unsupervised. These are the conditions under which many telephone-based conversational systems, such as the Jupiter weather information system [2], must operate. Standard algorithms like maximum a posteriori probability adaptation (MAP) [3] and maximum likelihood linear regression (MLLR) [4] do not work well in the rapid, unsupervised case. Many recent efforts have extended these algorithms to improve their performance under these conditions [5, 6, 7].

Our work takes a different approach. The work presented here is based on an algorithm called speaker cluster weighting (SCW) [1, 8]. SCW extends the ideas of previous speaker cluster techniques [9, 10] by allowing the speaker cluster models (learned from training data) to be adaptively weighted to match the current speaker. This approach is similar in spirit to Eigenvoices [11] because it explicitly utilizes the training data to constrain the adaptation. It also only requires a small number of parameters to be learned during adaptation, thereby enabling rapid adaptation.

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

This paper presents an examination of several different strategies that can be utilized by the SCW approach. First, we have tried different techniques for automatically clustering the speakers, including a novel algorithm called least squares linear regression (LSLR) clustering. Second, we have experimented with two different ways of training acoustic models for the clusters: one where only the mixture weights of the original SI model are adapted for each cluster and another where small cluster models are interpolated with the SI model.

2. SPEAKER CLUSTER WEIGHTING

2.1. Overview

SCW consists of three main steps, the first two of which are completed before recognition. In the first step, speakers in the training data are clustered according to acoustic similarity. These clusters should capture intra-speaker inter-class correlations. In the second step, acoustic models are trained for each speaker cluster. These models need to be both focused and robust. The third step occurs during recognition where the acoustic models for each cluster are combined in an optimal way using adaptation data collected from the speaker's previous utterances. The following three sections describe the approaches taken for each of these steps in this work.

2.2. Clustering

A speaker's acoustic characteristics can be shaped by various forces, including physical characteristics of the vocal tract and dialectic influences. Because these forces can be similar for many different speakers, it is possible to form speaker clusters within which speakers are highly similar across all phones. When performing recognition, we can then emphasize those clusters that most resemble the current speaker.

2.2.1. Gender Clustering

Gender-dependent modeling is the most obvious and widely used clustering technique. Gender clustering captures intra-speaker phonetic correlations largely because of the correlation between gender and vocal tract length. In our work, we have subdivided our training data to create three different sets of gender-dependent models, one for males, one for females, and one for children. Indeed, a limitation of gender clustering is that only three manually determined clusters are created. There are presumably additional characteristics beyond gender which can be used to create additional speaker clusters.

In *Proceedings of the 7th International Conference on Spoken Language Processing*, Sep. 16-20, 2002, Denver, Colorado, pp. 61-64.

2.2.2. LSLR Clustering

In order to allow an arbitrary number of clusters to be created and to allow for the use of intra-speaker correlations besides those introduced by gender, another clustering approach was developed. Because the corpus utilized in this work provided no relevant information on the speakers besides gender, it was necessary that this new approach be automatic. The corpus also contained very few training examples per speaker. This meant that the new approach had to be robust under this condition.

Two straightforward approaches were attempted. In the first, the distance between two speakers was defined as the Euclidean distance between the global mean of the principal component normalized feature vectors of the training data for each speaker [9]. In the second approach, the distance between two speakers was calculated as the average distance between the average principal component normalized feature vectors of all of the phones shared by the two speakers. In both cases, the resulting clusters were poor. In the first approach, this was due to the lack of phonetic diversity within each speaker's data. In the second approach, it was due to the fact that speakers often shared very few phones.

The failure of these two approaches led to the development of an algorithm we call least-squares linear regression (LSLR) clustering. LSLR clustering is automatic and reasonably insensitive to the phonetic content of the data available for a speaker. Using this approach, a two step process is utilized to compute a characteristic vector for each speaker. First, the observed feature vectors from different phones are transformed into a generic phonetic space that is shared by all phones. Second, the transformed features from all observations of a subset of the phones from one speaker are averaged to create that speaker's characteristic vector. The transformation is computed in order to minimize the intra-speaker variation of observations from different phones. Our work is partially influenced by Doh and Stern who demonstrated the utility of an inter-class affine transformation[12].

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

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

The transform for phone i is shared over all speakers s and is computed to minimize the average intra-speaker distance between the transformed mean vector for phone i and the destination mean vector. An LSLR transformation is computed independently for each phone. The minimization was performed with a standard least-squares algorithm.

Once the transforms for each phone are computed, the characteristic feature vector for each speaker is computed by averaging the transformed feature vectors from all training observations of a selected subset of phones for that speaker. The speakers are clustered based on this characteristic feature vector. This is done using K-means clustering with clusters initialized through a bottom-up procedure.

2.3. Cluster Model Combination

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

2.3.1. Cluster Weighting

In cluster weighting, adapted acoustic models are weighted combinations of the cluster models. The acoustic model resulting from cluster weighting for L clusters and a given phone p can be represented as:

$$\mathrm{p}_{scw}(\vec{x}|p) = \sum_{l=1}^{L} w_l \mathrm{p}_l(\vec{x}|p) \quad (2)$$

If U is the transcription of the adaptation data and X the set of all adaptation vectors, the optimal weights, $\vec{w}$ will satisfy the equation:

$$\vec{w} = \arg\max_{\vec{w}} \mathrm{p}_{scw}(X|U, \vec{w}) \quad (3)$$

The EM algorithm is used to perform this maximization. A different set of weights could be chosen for different phonetic classes, but that was not done in this work.

2.3.2. Best Cluster

In the best cluster approach, the set of cluster acoustic models is chosen that maximizes the adaptation data probability. While this approach does not allow as much flexibility in the adaptation process as cluster weighting, it may be more robust in the presence of small amounts of adaptation data.

2.4. Cluster Model Training

The method in which the acoustic models are trained for each speaker cluster has a significant impact on both the recognition accuracy of the adapted models as well as the computation required to perform recognition. It is also intimately related to the cluster combination algorithm being used.

2.4.1. Weight Adaptation

When using cluster weighting, the size of the resulting adapted acoustic models is of particular concern. If each of the cluster models is as large as the SI model, the resulting adapted acoustic models will require approximately L times as much computation as the SI models (where L is the number of clusters.) However, if the cluster models share the same Gaussian components, this problem is avoided. Training cluster models using weight adaptation consists of adapting only the weights of the Gaussian components of the SI model for each cluster. This is done for each class using the EM algorithm. Using this approach, the component weights of the adapted models are simply the interpolation of the component weights in each of the cluster models.

2.4.2. Model Interpolation

The best cluster approach allows each cluster model to be the same size as the SI model without creating a computational problem. However, using the best cluster requires that each of the models be very robust. If not, an error in the model selection process or the

selection of a model corresponding to a cluster with very little data will result in suboptimal performance, potentially even worse than that of the SI model.

This robustness is achieved through model interpolation. With model interpolation, a set of small acoustic models is trained for each cluster. These models are then interpolated with the SI model. For those classes for which a large amount of training data is present in the cluster, the cluster model is weighted more heavily. For those with a smaller amount of training data present in the cluster, the cluster model gets less weight. The exact formula for these weights is presented in [13] and shown here:

$$\mathrm{p}_{int}(\vec{x}|p) = \frac{N_p}{N_p + K}\mathrm{p}_{clust}(\vec{x}|p) + \frac{K}{N_p + K}\mathrm{p}_{si}(\vec{x}|p) \quad (4)$$

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

3. EXPERIMENTS AND RESULTS

3.1. Corpus and Recognizer

In these experiments, rapid unsupervised adaptation is applied in the Jupiter domain. Jupiter is a telephone-based spoken dialogue system for weather information where users can ask multiple questions during one call. The training corpus used in this work consisted of 17,116 calls. Each call had an average of 4.7 utterances per call giving the training corpus a total of 80,487 utterances. Utterances averaged about 3 seconds in length. The test set consisted of 771 calls (499 male, 168 female, and 104 children), and each call consisted of at least 6 utterances. In both the training corpus and test set, some calls may have shared the same speaker. However, the Jupiter system does not keep track of callers' identities, so each call was treated as if it came from a different speaker.

The recognizer used was the SUMMIT segment-based speech recognition system [14]. Landmark-based acoustic models were used. These models are mixtures of Gaussians which model transitions between phones as well as landmarks internal to phonetic segments.

3.2. Speaker Clustering

For the recognition experiments in the next section, 5 clusters were created using LSLR clustering. As described in section 3.2, LSLR clustering requires that a set of classes be chosen on which to base the clustering. The first set consisted of context-independent models for all the vowels, the second set for a selected subset set of 7 common unreduced vowels ([i], [ɪ], [e], [æ], [ɑ], [ɔ], [u]), the third set for all the nasals and the fourth set for all the fricatives. To evaluate the quality of these clusterings without running recognition experiments, we examined gender distributions of the clusters determined using each of the phone sets.

As was expected, the vowel sets perform best. Because the smaller vowel set performed as well as the set consisting of all the vowels but requires less computation during clustering, it was chosen for use in the recognition experiments. Table 1 shows the gender distribution of each of the clusters created using this vowel set. We see that clusters 1, 2, and 3 are almost entirely male, while cluster 5 contains most of the children in the training set and almost no males. This suggests that the algorithm is effectively separating voices with a great deal of acoustic disparity.

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

Table 1. Gender distribution for 5 clusters created using LSLR clustering.

3.3. Recognition Experiments

Recognition experiments were completed using cluster weighting with weight adaptation and using best cluster with model interpolation. In both cases experiments were run using both gender and LSLR clustering. For gender clustering, three clusters were used, one containing all the males in the training data (13,030 calls, 59,076 utterances), one containing all the females (3,297 calls, 14,386 utterances) and one containing all the children (1206 calls, 7034 utterances.) When LSLR clustering was used, the clusters described in the previous section were utilized.

For weight adaptation, an SI model with a maximum of 50 Guassians per class was trained. Cluster models were then trained using the method described in Section 2.4.1. When model interpolation was applied, cluster models with a maximum of 10 Gaussians per class were created and then interpolated with the SI model with a maximum of 40 Guassians per class according to the method described in Section 2.4.2.

Recognition was performed on each utterance in the test set, using up to 5 of the same speaker's utterances to adapt. The utterance being recognized was never included in the adaptation data. The best path obtained using the SI model was used as the transcription for the adaptation data, making the task unsupervised.

In Figure 1 results are presented for adaptation after one utterance using cluster weighting with weight adaptation and best cluster with model interpolation and both types of clustering. We see that neither cluster weighting nor best cluster is consistently superior on females and children. However, cluster weighting results in a slight degradation over the SI model for males, while best cluster results in a slight improvement. Comparing the clustering approaches, we see that the 5 LSLR clusters perform comparably to the 3 gender clusters.

Figure 2 allows us to compare the speed of adaptation for the two clustering approaches. We see that the difference in WER after 1 adaptation utterance and after 5 adaptation utterances is more pronounced when using the 5 LSLR clusters than when using 3 gender clusters. This is probably because, when using 5 LSLR clusters, the number of degrees of freedom for the weighting algorithm is increased. This results in less robust parameter estimation after 1 adaptation utterance, but also can allow for more effective adaptation with more adaptation data. The superior performance of the 5 LSLR clusters on the children speakers suggests that the added flexibility provided by the increased number of clusters allows for better adaptation to those speakers least represented in the training data.

Using Figure 3 we can compare the speed of adaptation for cluster weighting and best cluster. The difference in WER between using 1 adaptation utterance and using 5 adaptation utterances is

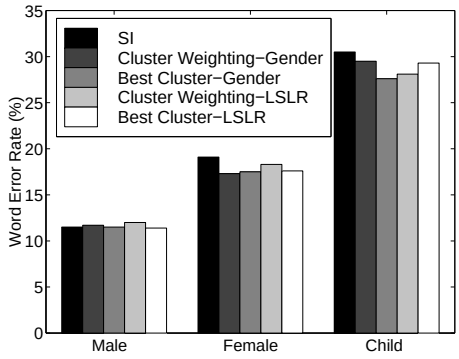

Fig. 1. WER using cluster weighting and best cluster with gender clustering and LSLR clustering after 1 adaptation utterance.

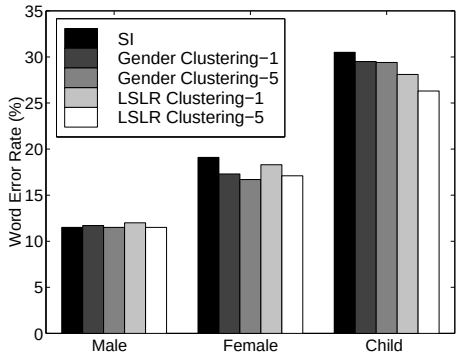

Fig. 2. WER using cluster weighting with gender clustering and LSLR clustering for 1 and 5 adaptation utterances.

slightly larger for cluster weighting. This makes sense for the same reason that adaptation was slower with 5 LSLR clusters than with 3 gender clusters: cluster weighting allows for a great deal more flexibility in creating the final adapted model than does best cluster.

4. CONCLUSION

This work has demonstrated the effectiveness of SCW for rapid speaker adaptation. Using this approach with the best cluster weighting procedure, model interpolation and gender clustering lowers WER from 19.1% to 17.5% for women (a 9% relative improvement) and lowers WER from 30.5% to 27.6% for children (a 10% relative improvement) after only 1 adaptation utterance. We have also demonstrated the effectiveness of LSLR clustering for the purposes of SCW. Using SCW with LSLR clusters achieves comparable performance to that achieved when using gender clusters. LSLR allows for the automatic creation of an arbitrary number of clusters. This ability to create a larger number of clusters enables SCW to adapt better to those speakers most different from the training data.

5. REFERENCES

[1] T. J. Hazen, "A comparison of novel techniques for rapid speaker adaptation," *Speech Communication*, vol. 31, pp. 15–33, May 2000.

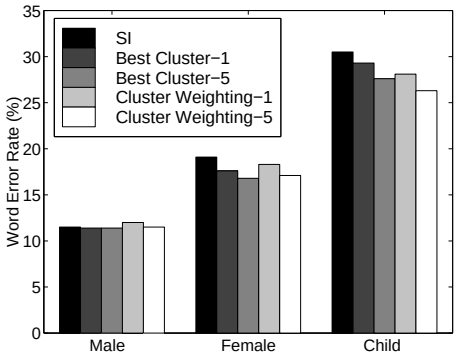

Fig. 3. WER using best cluster and cluster weighting with LSLR clustering for 1 and 5 adaptation utterances.

[2] T. J. Hazen J. R. Glass and I. L. Hetherington, "Real-time telephone-based speech recognition in the Jupiter domain," in *Proc. of ICASSP*, Phoenix, 1999, pp. 61–64.

[3] J. Gauvain, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[4] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adpatation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, April 1995.

[5] K. Shinoda and C. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barabara, 1997, pp. 381–387.

[6] O. Siohan, T. Myrvoll and C. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Speech Communication*, vol. 16, no. 1, pp. 5–24, 2002.

[7] A. Kannan and M. Ostendorf, "Modeling dependency in adaptation of acoustic models using multiscale tree proceesses," in *Proc. of EUROSPEECH*, Rhodes, Greece, 1997, vol. 1, pp. 1863–1866.

[8] Y. Gao, M. Padmanabhan, and M. Pichney, "Speaker adaptation based on pre-clustering training speakers," in *Proc. of EUROSPEECH*, Rhodes, Greece, 1997, vol. 4, pp. 2091–2094.

[9] S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," in *Proc. of ICASSP*, Glasgow, Scottland, 1989.

[10] T. Kosaka and S. Sagayama, "Tree-structured speaker clustering for fast speaker adaptation," in *Proc. of ICASSP*, Adelaide, Australia, 1994, vol. 1, pp. 245–248.

[11] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in Eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[12] S. Doh and R. Stern, "Inter-class MLLR for speaker adaptation," in *Proc. of ICASSP*, Istanbul, Turkey, 2000, pp. 1775–1778.

[13] T. J. Hazen, *The Use of Speaker Correlation Information for Automatic Speech Recognition*, Ph.D. thesis, MIT, Jan. 1998.

[14] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. of ICSLP*, Philadelphia, Oct. 1996, pp. 2277–2280.