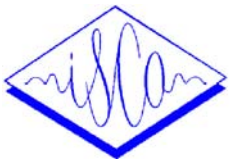

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

SPEAKER ADAPTATION BASED ON PRE-CLUSTERING TRAINING SPEAKERS

Yuqing Gao, Mukund Padmanabhan, Michael Picheny
IBM T. J. Watson Research Center
P. O. Box 218, Yorktown Heights, NY 10598

1 ABSTRACT

A new strategy for speaker adaptation is described that is based on: (1) pre-clustering all the speakers in the training set acoustically into clusters; (2) for each speaker cluster, a system is built using the data from the speakers who belong to the cluster; (3) when a test speaker's data is available, we find a subset of these clusters, closest to the test speaker; (4) we transform each of the selected clusters to bring it closer to the test speaker's acoustic space; (5) we build a speaker-adapted model using transformed cluster models.

This method solves the problem of excessive storage for the training speaker models[1], as it is relatively inexpensive to store a model for each cluster. Also as each cluster contains a number of speakers, parameters of the models for each cluster can be robustly estimated. The algorithm has been evaluated on a large vocabulary system and compared to existing algorithms. The imporvement over existing algorithms such as MLLR[2] is statistically significant.

2 INTRODUCTION

In a previous paper[1], a speaker adaptation scheme was described based on the fact that a speech training corpus contains a number of training speakers, some of whom are closer, acoustically, to the test speaker, than others. Therefore, given a test speaker, if the acoustic models are re-estimated from a subset of the training speakers who are acoustically close to the test speaker, the system should be a better match to the test data of the speaker. A further improvement can be obtained if the acoustic space of each of these selected speakers is transformed to come closer to the test speaker.

Given a test speaker, the adaptation procedure used in [1] is: (1) find a subset of speakers from the training corpus, who are acoustically close to the test speaker; (2) transform the data of each of these speakers to bring it closer to the test speaker, and (3) use only the (transformed) data from these selected speakers, rather than the complete training corpus, to re-estimate the model (Gaussian) parameters. This scheme was shown to produce better speaker adaptation performance than other algorithms, for example MLLR[2], or MAP adaptation[3], when only a small amount of adaptation data was available.

The implementation of [1] used the transformed training data of each selected training speaker to re-estimate the system parameters; this required the entire training corpus to be available on-line for the adaptation process, and is not practical in many situations. This problem can be circumvented if a model is stored for each of the training speakers, and the transformation (to bring the training speaker closer to the test speaker) is applied to the model. The transformed models are then combined to produce the speaker-adapted model. However, due to the large number of training speakers, storing the models of each training speaker would require a prohibitively large amount of storage. Also, we may not have sufficient data from each training speaker to robustly estimate the parameters of the speaker-dependent model for the training speaker.

To solve this problem and retain the advantages of the method in [1], we present a new idea in this paper, which is based on pre-clustering the training speakers acoustically into clusters. The speaker pre-clustering can also be viewed as a partitioning the acoustic space in terms of speakers. For each speaker cluster, an acoustic system (called a "cluster-dependent system") is trained using speech data from the speakers who belong to the cluster. When a test speaker's data is available, we rank these cluster-dependent systems according to the distances between the test speaker and each cluster, and a subset of these clusters, acoustically closest to the test speaker, is chosen. Then the model for each of the selected clusters is transformed[2] to bring the model closer to the test speaker's acoustic

space. Finally these adapted cluster models are combined to form a speaker adapted system. Hence, compared to [1], we now choose clusters that are acoustically close to the test speaker, rather than individual training speakers.

This method solves the problem of excessive storage for the training speaker models, because the number of clusters is far fewer than the number of training speakers, and it is relatively inexpensive to store a model for each cluster. Also as each cluster contains a number of speakers, we have enough data to robustly estimate the parameters of the model for the cluster.

The following problems are relevant in the context of this adaptation strategy. (1) how should we pre-cluster the training speakers; (2) what model should we use to describe a cluster; (3) given a test speaker, how should we select the clusters that are acoustically closest to the test speaker; (4) finally, how should we build a speaker adapted model from the cluster models. We will describe a number of experiments that attempt to solve these problems along with the results that were obtained, in the following sections.

3 PRE-CLUSTERING TRAINING SPEAKERS

Pre-clustering training speakers is a problem related to finding similarities across different speakers. Firstly, it is necessary to define a set of acoustic characteristics to represent each speaker. In the speaker identification/verification framework, phonetic information is often ignored in representing a speaker. For example, a codebook, where the mixture components are not associated with phones, may be used to represent a speaker. However, in our implementation, we have used the phonetic information in the speaker models. A phone-based HMM system is used to represent a speaker for the speaker pre-clustering purpose. In our experiment, 52 phones are used, consequently, 52 3-state HMMs are used for each speaker. We used a bottom-up scheme to cluster the individual speaker models. A Gaussian log likelihood is used as a distance measure in the bottom-up clustering procedure.

$$logP_i = -c_i \left[\frac{n}{2} log(2\pi) + \frac{1}{2} |\underline{\underline{\Lambda}}_i| \right] \qquad (1)$$

where, c_i is the E-M count, $\underline{\underline{\Lambda}}_i$ is the variance of the Gaussian. n is the dimension.

In measuring the similarity between any 2 speakers, Gaussian log likelihoods are computed between the same arcs of phone models of 2 speakers. The overall likelihood between 2 speakers is a summation over 156 arcs.

If speaker i and j are parameterized by: c_i^k, $\underline{\mu}_i^k$, $\underline{\underline{\Lambda}}_i^k$ and c_j^k, $\underline{\mu}_j^k$, $\underline{\underline{\Lambda}}_j^k$, $k = 1, ..., K$, K is the number of arcs, we compute a merged Gaussian as:

$$\hat{c^k} = c_i^k + c_j^k \qquad (2)$$

$$\hat{\underline{\mu}^k} = \frac{c_i^k \underline{\mu}_i^k + c_j^k \underline{\mu}_j^k}{\hat{c^k}} \qquad (3)$$

$$\hat{\underline{\underline{\Lambda}}^k} = \frac{c_i^k \underline{\underline{\Lambda}}_i^k + c_i^k (\underline{\mu}_i^k - \hat{\underline{\mu}^k})^2 + c_j^k \underline{\underline{\Lambda}}_j^k + c_j^k (\underline{\mu}_j^k - \hat{\underline{\mu}^k})^2}{\hat{c^k}} \qquad (4)$$

Then, using formula (1) to compute the likelihood.

The bottom-up clustering starts with each node represents a speaker. At every step, 2 nodes are merged into one such that the merged node has a larger likelihood than all other possible merges. The merging continues till the bottom-up tree has the desired number of nodes on the top level.

The distance measure used in the bottom-up clustering procedure is the Gaussian log likelihood. We found that this performed better than other distance measures such as Kullback-Leibler, Euclidean, etc.

4 MODELS FOR THE CLUSTERS

After pre-clustering the speakers, we need to build an acoustic characterization for each speaker cluster in order to determine which clusters are close to the test speaker. An acoustic system is trained using speech data from the speakers who belong to the cluster. The cluster-dependent system was chosen to have the same complexity as a speaker-independent system. In our experiment, each cluster contains anywhere from 9 to 31 speakers, and each speaker has 100-200 utterances available in the training corpus, producing 1000-5000 utterances for each cluster. However, 1000-5000 utterances from each cluster still do not suffice to robustly estimate the parameters of the cluster-dependent models. We used Bayesian adaptation techniques[3] to smooth each cluster-dependent model with a speaker-independent model.

Let L denote the total number of Gaussains, d denote the dimension of the acoustic features, and $\underline{\mu}_i^{ind}, \underline{\underline{\Lambda}}_i^{ind}, p_i^{ind} \quad i = 1, \cdots, L$ denote the means, variances and priors of a speaker-independent acoustic model; and let the k^{th} cluster be parametrized by $\underline{\mu}_i^k, \underline{\underline{\Lambda}}_i^k, p_i^k \quad i = 1, \cdots, L$. We use the re-estimation formulae

$$p_i^k = \frac{c_i + \tau_i}{\sum_{j \in J} (c_j + \tau_j)} \qquad \tau_i = p_i^{ind} \tau \qquad (5)$$

where, J is the collection of Gaussians which belong to the same leaf as ith Gaussian does, and τ is a constant.

$$\underline{\mu}_i^k = \frac{\underline{\eta}_i + \tau_i \underline{\mu}_i^{ind}}{c_i + \tau_i} \tag{6}$$

$$\underline{\underline{\Lambda}}_i^k = \frac{\underline{\underline{\gamma}}_i + \tau_i \left[\underline{\underline{\Lambda}}_i^{ind} + \underline{\mu}_i^{ind} \underline{\mu}_i^{ind,T}\right]}{c_i + \tau_i} - \underline{\mu}_i^k \underline{\mu}_i^{k,T} \tag{7}$$

where

$$c_i = \sum_t c_i(t), \qquad \underline{\eta}_i = \sum_t c_i(t) \underline{y}_t,$$

$$\underline{\underline{\gamma}}_i = \sum_t c_i(t) \underline{y}_t \underline{y}_t^T.$$

Here $c_i(t)$ is the *a-posteriori* probability of the Gaussian i at time t, conditioned on all acoustic observations $\underline{y}_1^T$, and the terms $c_i(t), \underline{\eta}_i, \underline{\underline{\gamma}}_i$ are usually referred to as the E-M counts.

5 SELECTING A SUBSET OF CLUSTERS

Here we discuss, how to find a subset of the clusters which are acoustically closest to the test speaker. We ran some initial speaker recognition experiments to decide on the distance measure to be used in the cluster selection procedure.

In the speaker recognition experiment, we selected a subset of the training speakers and evaluated the phone-based speaker model that was closest to each training speaker based on various distance measures. Ideally of course, the closest model should be the model of the selected training speaker. Our results indicated that a Euclidean distance measure was better than the other measures that we tried. When the Mahalanobis distance was used in the speaker recognition experiment, the correct model was selected only 62.3% of the time. While when Euclidean distance was used, the correct model was selected 100% of the time.

Consequently, it appears that the Euclidean distance is a better measure to use in the cluster selection process. The results of the speech recognition experiments we report in the following sections also confirm this inference.

The speech data from a test speaker is first decoded using a speaker-independent system to generate a transcription. Subsequently, the data is Viterbi-aligned against the transcription and each acoustic observation is tagged with an allophone id. The distance of the adaptation data, conditioned on the Viterbi alignment, to each cluster-dependent model is then calculated using each cluster model, and the clusters are ranked in the order of this distance. The top N clusters are then selected as being acoustically close to the test speaker. In computing this distance, we have the option of using the Euclidean distance or the Mahalanobis distance depending on the final recognition accuracy.

6 TRANSFORMATION AND RE-ESTIMATION OF GAUSSIANS

We next use the MLLR[2] technique to transform a cluster-dependent model and bring it closer to the test speaker. Given some observations from a test speaker, a subset of clusters can be selected using the above procedure. Given a selected cluster model, mean $\underline{\mu}_i^k$ and variance $\underline{\underline{\Lambda}}_i^k$, one can compute a posterior probability, $c_i^k(t)$ of the i^{th} Gaussian at time t, conditioned on all the acoustic observations in the adaptation data, and compute the transformations so as to maximize the likelihood of the adaptation data, given the selected cluster model. This is equivalent to minimizing the following objective function[2]:

$$\sum_{i,t} c_i^k(t) \left[(\underline{x}_t - \underline{\underline{A}}^k \hat{\underline{\mu}}_i^k)^T \underline{\underline{\Lambda}}_i^{k^{-1}} (\underline{x}_t - \underline{\underline{A}}^k \hat{\underline{\mu}}_i^k) + \log\left(|\underline{\underline{\Lambda}}_i^k|\right)\right] \tag{8}$$

with respect to $\underline{\underline{A}}^k$. Here, $\underline{\underline{A}}^k$ is a (d)x$(d+1)$ matrix, and $\hat{\underline{\mu}}_i^k$ is a $(d+1)$x1 vector obtained from $\underline{\mu}_i^k$ as $(\hat{\underline{\mu}}_i^k)^T = \left[(\underline{\mu}_i^k)^T \;\; 1\right]^T$.

Once the transformations have been computed, the transformed means of cluster model become $\underline{\underline{A}}_i^k \underline{\mu}_i^k$. The Gaussian means of the adapted model can be formed by accumulating the transformed model means of all selected clusters using the re-estimation formulae given below, while the variances of the model is left unchanged at the speaker-independent values.

$$\underline{\mu}_i^{adapted} = \frac{\sum_k c_i^k \left[\underline{\underline{A}}_i^k \underline{\mu}_i^k\right]}{\sum_k c_i^k} \tag{9}$$

where $c_i^k = \sum_t c_i^k(t)$.

7 EXPERIMENTS AND DISCUSSIONS

In our experiments, the size of our model was 36,000 Gaussians. The training corpus (Wall Street Journal database) has 284 speakers, half male and half female.

Each speaker has about 100 sentences of speech data in the training corpus.

The baseline system and the new algorithm are evaluated using 10 test speakers. We have also compared our algorithm with standard MLLR[2] adaptation as well. Each of the 10 test speakers has 50 sentences of adaptation data and 61 sentences of testing data.

The baseline speaker-independent system has an average word error rate of 16.35% over the 10 test speakers. MLLR adaptation reduces the error rate to 12.67% using 50 adaptation utterances with an average of 25 transformations for each speaker.

For the speaker-pre-clustering purpose, each training speaker is modelled by 156 gaussians (one gaussian per state, with 3 states for each of 52 phones). When these training speaker models are bottom-up clustered, the scheme generates 15 clusters, each containing anywhere from 9 to 31 speakers. Each cluster has about 1000-5000 utterances and a 36,000-Gaussian system is built for each cluster by using the data from the speakers belonging to the cluster.

In the tables below, we compared the performance with the use of the Mahalanobis distance and Euclidean distance in selecting close clusters for each test speaker. We also compared 2 different ways of selecting the number of clusters: in the first way, we fixed the number of selected clusters. In the second way, we selected a subset of clusters subject to a maximum number of speakers in the selected clusters. For a given threshold of maximum number of selected speakers, the selected number of clusters for test speakers may be different from one test speaker to another depending on how many speakers each cluster contains.

# of clusters	Mahalanobis	Euclidean
3	12.56%	12.11%
5		12.07%
6	12.44%	
8	12.29%	
10	12.18%	
12	12.20%	
15	12.25%	

# of spkrs: Max	Ave	Mahalanobis	Euclidean
80	69	12.51%	12.18%
90	76		12.05%
100	89	12.40%	12.01%
110	100		12.03%
120	110		12.18%
150	141	12.26%	
200	184	12.12%	

From the viewpoint of recognition error rate of the final speaker adapted system, the Euclidean distance measurement is superior to the Mahalanobis distance. We can also compare different distance measures in terms of number of selected speakers. In [1], the optimum performance was obtained when the 50 closest training speakers were selected to build a speaker-adapted system. However, in our experiments with the Mahalanobis distance, we see that the performance is better when more speakers (the best performance is obtained when the number of clusters selected is 10, which corresponds to selecting a total of 184 speakers on average) are used. Meanwhile, from the experiments that use the Euclidean distance, we see that fewer clusters need to be selected (average number of speakers is 89) to obtain optimal performance. We believe these experiments reflect the quality of different cluster selecting techniques, and clearly, the Euclidean distance is better than the Mahanalobis distance in cluster selection.

Comparing these results with the MLLR technique[2], MLLR provides a 22.5% relative improvement over the speaker-independent system, while our scheme provides a relative improvement of 26.5% over the speaker-independent performance. The NIST "standard" benchmark testing program[4] is used to compary the performance of our method with MLLR in terms of statistical significance. It shows the improvement over MLLR is significant.

We are currently investigating other ways of pre-clustering speakers. For example, one can totally ignore the phonetic information during speaker clustering and use VQ based acoustic characteristics to represent a speaker. On the other hand, one can emphasize the context information and do a phone-dependent speaker clustering. For example, certain phones of speaker A belong to cluster k, and some other phones of the same speaker belong to cluster j, and so on.

8 REFERENCES

1. M. Padmanabhan, et al, "Speaker Clustering Transformation for Speaker Adaptation in Large-Vacobulary Speech Recognition Systems", ICASSP'96, also to appear in IEEE Trans. Speech and Audio Processing

2. C. J. Leggetter, et al, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, Vol 9, 1995.

3. J. L. Gauvain and C. H. Lee, "Maximum-a-Posteriori estimation for multivariate Gaussian observations and Markov chains", IEEE Trans. Speech and Audio Processing, vol. 2, no. 2, Apr 1994.

4. L. Gillick, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", ICASSP'89.