ACOUSTIC CLUSTERING AND ADAPTATION FOR ROBUST SPEECH RECOGNITION *

Larry Heck and Ananth Sankar

Speech Technology And Research Laboratory
SRI International
Menlo Park, CA
{heck,sankar}@speech.sri.com

ABSTRACT

We describe an algorithm based on acoustic clustering and acoustic adaptation to significantly improve speech recognition performance. The method is particularly useful when speech from multiple speakers is to be recognized and the boundary between speakers is not known. We assume that each test data segment is relatively homogeneous with respect to the acoustic background and speaker. These segments are then grouped using an agglomerative acoustic clustering algorithm. The idea is to group together all test segments that are acoustically similar. The speech recognition models are then adapted separately to each test data cluster. Finally these adapted models are used to recognize the data from that cluster. This algorithm was used in SRI's system for the 1996 DARPA Hub4 partitioned evaluation. Experimental results are presented on the 1996 H4 development data set. It was found that an improvement of 9.5% was achieved by using this algorithm.

1. INTRODUCTION

Recently there has been much research on acoustic adaptation [1, 2, 3, 4, 5] to improve the performance of speech recognition systems in mismatched training and testing acoustic environments. Adaptation techniques typically adapt a model trained under one condition to the test environment using a small amount of data from the test environment. Common approaches include maximum-likelihood (ML) transformation-based techniques [1, 2, 3], and Bayesian algorithms [5]. These approaches have been shown to give significant improvements in performance for cases such as mismatched speakers, and channels [1, 2, 3]. To use these techniques it is necessary to have a sufficient amount of adaptation data from the test environment. However, in some situations, this may be difficult to get. For example, in broadcast news, the domain for the 1996 DARPA Hub4 benchmark tests, the acoustic environment is continually changing. There are constant switches of channel types, background noise, and speakers. Thus, to adapt the models to each test environment, it is necessary to collect all test segments from a particular environment before adapting the model to that environment. Unsupervised acoustic clustering of the test segments can be used to group acoustically similar segments. Clustering also serves the purpose of more robust adaptation parameter estimation because similar test environments can be clustered, resulting in more data for adaptation. In this paper, we report on an algorithm that uses this idea specifically to cluster the test data speakers in the 1996 H4 partitioned evaluation (PE) development set. Seed recognition models are then adapted to each speaker cluster. Experimental results show that the algorithm gives a significant 9.5% improvement over the seed models.

*THIS WORK WAS SUPPORTED BY DARPA THROUGH THE NAVAL, COMMAND, CONTROL AND OCEAN SURVEILLANCE CENTER CONTRACTS #N66001-94-C-6048.

2. BASELINE SYSTEM

The 1996 DARPA Hub4 evaluation was divided into two individual problems. In the unpartitioned evaluation (UE), the test data consisted of a set of television and radio shows in their entirety. However, commercials and sports reports were not included in the data to be recognized, as the language used was considered to be very different from the rest of the data. In the UE, it is necessary to automatically excise the speech segments from the test data before recognizing them. In the PE, each test show was partitioned into segments of speech. Thus, pure music or noise segments were removed by hand, and only speech segments remained. Each segment contained speech from a single speaker. In addition, the segments were homogeneous with respect to the acoustic background condition or speech style. The segments were classified into seven different acoustic focus conditions, F0, F1, F2, F3, F4, F5, FX, as described in [6], and the labels were provided for use in the evaluation. The experiments reported in this paper use the PE development data.

Our baseline system was a gender-dependent Genonic hidden Markov model (HMM) system [7] adapted to each acoustic focus condition by using the training data provided for that focus condition. This approach results in condition-specific HMMs that can be used to recognize speech from the corresponding focus condition. ML transformation-based adaptation was used to adapt a seed Genonic HMM to each of the acoustic focus conditions. A parametric transformation of the HMMs is postulated, and the parameters of the transformation are estimated by maximizing the likelihood of the training data from the acoustic focus condition [1, 2, 3]. We used a block-diagonal affine matrix transformation of the HMM mean

vectors in this stage [4]. This is a modification of the algorithm presented in [2] that results in more estimation of the adaptation transformations. To approximate more complex transforms, we used multiple block-diagonal affine transformations, where each transformation is tied to a group of Gaussians. In our adaptation algorithm, this tying is achieved using hand-generated phone clusters when the number of transformations is less than the number of phones. If the number of transforms is greater than the number of phones, then the transforms are tied to Gaussian groups generated using the HMM state clustering algorithm used to train our Genonic HMMs [7]. For the condition-specific models, we used 11 block-diagonal affine transforms of the HMM mean vectors. The number of transforms was optimized to give the lowest word error rate on the 1996 Hub4 PE development data. While the results presented in this paper are on this same data set, the number of adaptation parameters is relatively small compared to the amount of training data, and the results have been found to generalize well for both this data set and in other experiments. Furthermore, the performance is not very sensitive to small changes in the number of transforms. Since the F2 condition corresponds to telephone speech, we decided to use seed models trained on the Switchboard and Macrophone databases for F2. For all other focus conditions, we used seed models trained on the Wall Street Journal (WSJ) SI-284 database. Details of this approach are presented in [8].

3. TEST DATA CLUSTERING AND ADAPTATION

The condition-specific models described in Section 2 are estimated using adaptation algorithms and the training data for each focus condition. However, there may still be a mismatch between these condition-specific models and test data from the same acoustic condition. Such mismatches are largely due to different speakers between training and testing. In addition, there may be small differences in the training and test acoustical conditions, leading to a mismatch. Since the main source of variability between the training and test conditions is the different speakers, we used an unsupervised bottom-up agglomerative clustering algorithm to cluster acoustic segments that were similar to each other. Since acoustic segments of the same speaker are similar, the resulting clusters are typically homogeneous with respect to speakers.

Once the segments are clustered, the condition-specific models are separately adapted to each cluster by using the block-diagonal mean transformation [4], followed by a variance scaling transformation described in [3, 4]. The variance scaling transform has also been studied more recently in [9]. In this stage we used three separate transformations, including a separate transformation for the silence Gaussians. The reference transcriptions for adaptation were derived by running a one-pass Viterbi recognition search through word lattices [10] using the condition-specific models described in Section 2. Once the models are adapted, it is possible to re-recognize the acoustic segments for each cluster and then re-adapt the models by using these new hypotheses. However, we did not observe a significant improvement with multiple iterations of this kind, and hence we used only one iteration.

For clustering, the distance between two acoustic segments $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,T_i}\}$ and $\boldsymbol{X}_j = \{\boldsymbol{x}_{j,1}, \ldots, \boldsymbol{x}_{j,T_j}\}$ was computed using a symmetric relative entropy distance,

$$D(i,j) = \frac{1}{T_i}\sum_{t=1}^{T_i} \log \frac{p(\boldsymbol{x}_{i,t} \mid \lambda_i)}{p(\boldsymbol{x}_{i,t} \mid \lambda_j)} + \frac{1}{T_j}\sum_{t=1}^{T_j} \log \frac{p(\boldsymbol{x}_{j,t} \mid \lambda_j)}{p(\boldsymbol{x}_{j,t} \mid \lambda_i)}, \quad (1)$$

where Λ_i and Λ_j are the underlying statistical models of X_i and X_j. In our work, we used a Gaussian mixture model (GMM) to model each test segment. The distance between two clusters was then computed as the maximum distance between segments in the two clusters. A threshold on the minimum distance between any pair of clusters defines a cut in the agglomerative cluster tree and hence a set of test segment clusters. The maximum, or furthest neighbor distance, generally resulted in more speaker-homogeneous clusters than a nearest neighbor distance, or average distance between clusters. This can be seen for the case of the maximum and minimum distances in Figure 1 which plots the average speaker-class entropy over all the clusters against the distance threshold. The speaker-class entropy measures the speaker homogeneity and is computed as

$$H = \frac{1}{N_C}\sum_{i=1}^{N_C}\sum_{j=1}^{N_{S_i}} -P(S_j|C_i)\log P(S_j|C_i), \quad (2)$$

where N_C is the number of clusters, N_{S_i} is the number of speakers in cluster i, $P(S_j|C_i)$ is the conditional probability of speaker j in cluster i, and the term being summed is the speaker-class entropy for cluster i, which is a measure of the homogeneity of that cluster. As the distance threshold increases from 0 to ∞, the number of clusters decreases from the number of test segments to 1, and the entropy goes from 0 to a maximum. For more speaker homogeneous clusters, we expect the entropy to increase slowly as the distance threshold is increased. Figure 1 shows that for the maximum distance, the entropy increase is much more gradual than for the minimum distance, indicating more speaker homogeneous clusters for the maximum distance. For our system, the distance threshold was empirically determined. By examining the clusters on the 1996 H4 development set, we found that the clusters were indeed quite homogeneous with respect to speakers. This clustering procedure was previously described by us in [11], but applied to cluster the training data speakers. In the work reported here, we used it to cluster the test data segments.

Since a mixture model must be trained for each segment to compute the relative entropy measure, and many of the segments were short in duration (some less than 1 second), we varied the number of Gaussians in the model of each segment based on a heuristic function of the segment

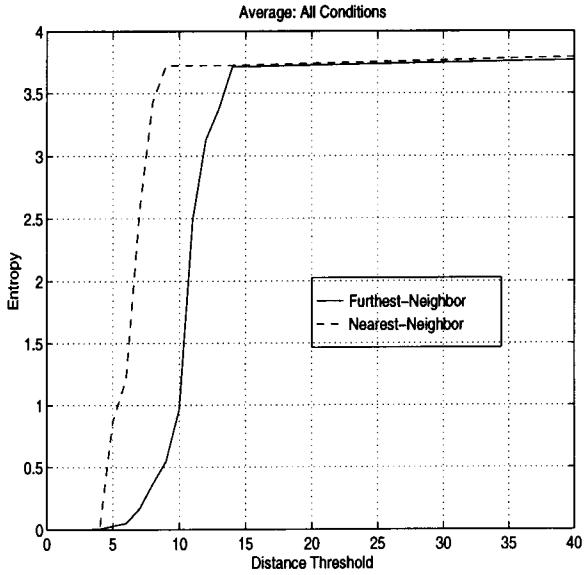

Figure 1. Comparison of agglomerative clustering methods

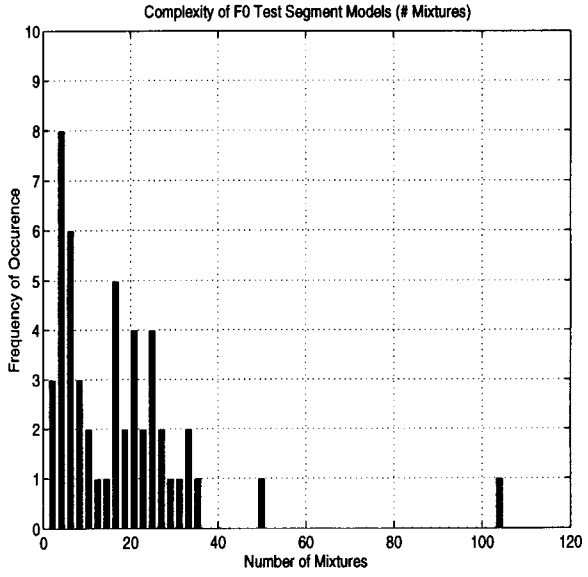

Figure 2. GMM sizes used to model the test segments

duration length. This prevented over-fitting of the model to the short data lengths. Figure 2 shows a histogram of the number of Gaussians used in the GMMs for each development segment of the F0 condition. As can be seen, the number of Gaussian mixtures varies from one to more than one hundred.

4. EXPERIMENTAL RESULTS

The Hub4 PE test data consisted of test segments that were marked as belonging to the different focus conditions. Since some of the test segments were very long, we further segmented these into nominally 10-second segments so as to reduce the memory and computation requirements on the decoding process. This was done by using an automatic segmentation algorithm described in [8].

The front-end feature extraction was based on mel-frequency cepstrum processing. The original speech data was sampled at 16,000 samples per second. For the F2 (telephone) segments, the speech was band limited, and down-sampled to 8,000 samples per second. To extract features, the speech was then hamming-windowed with a 25.6-ms window, and the window was advanced every 10 ms. Each frame was represented by 12 mel-frequency cepstrum coefficients, the log energy, and their first- and second-order time derivatives (delta and delta-delta features), for a resulting 39-dimensional feature vector.

The performance of the baseline models described in Section 2 and the cluster-adapted models described in Section 3 on the Hub4 PE development test data is shown in Table 1. The baseline models were generated using 11 block-diagonal affine transformations of the seed model mean vectors. The cluster-adapted models used three transformations. The number of transformations was experimentally chosen to give the lowest error rate. Whereas for the baseline we used only the block-diagonal affine transformation of the HMM mean vectors, for cluster-adaptation we tried both a block-diagonal affine transformation of the HMM means, and the mean transformation followed by a variance scaling transformation [3, 4].

Condition	Models		
	Condition-specific	Test-cluster-adapted	
		Mean	Mean and variance
F0	22.6	21.3	20.8
F1	41.2	38.8	38.8
F2	47.2	44.0	42.3
F3	45.6	42.5	42.4
F4	36.9	34.4	33.8
F5	36.3	28.3	28.1
FX	63.8	57.8	57.2
All	41.2	37.8	37.3

Table 1. Performance of test-condition-adaptation

From the table, we see that the acoustic clustering and adaptation algorithm gives a significant improvement over the baseline condition-specific models for every focus condition. Over all the data, we get a 8.3% relative improvement when using only mean adaptation for the test clusters and a 9.5% relative improvement when using mean and variance adaptation. We had also tried variance adaptation to train the baseline condition-specific models of Section 2. However, we did not observe any significant improvement over just mean adaptation. It is interesting to note that for the cluster-adapted models, variance adaptation gives a small but consistent improvement for all the focus conditions. This improvement is statistically significant at the 0.3% significance level using a paired-sample sign test on the word error rate of the test sentences.

To evaluate the advantage of performing the unsupervised speaker clustering, we also adapted the condition-specific models to each test condition without doing any

Condition	Adapt to test conditions			
	Single Cluster		Multiple clusters	
	Mean	Mean and variance	Mean	Mean and variance
F0	22.4	22.7	21.3	20.8
F1	40.0	40.3	38.8	38.8
F2	46.9	46.2	44.0	42.3
F3	44.7	44.8	42.5	42.4
F4	35.0	34.5	34.4	33.8
F5	31.4	31.1	28.3	28.1
FX	62.2	62.2	57.8	57.2
All	39.9	39.8	37.8	37.3

Table 2. Effect of clustering

clustering. Thus, in this case, all the data from any test condition was used to adapt the models as opposed to only the data in each of the speaker clusters in the test data. Since using all the test data allowed us to estimate a larger number of transformations, we used 11 transformations, including a separate transformation for the silence Gaussians, as compared to 3 transformations in the case of adapting to the speaker clusters.

Table 2 shows the advantage of using the unsupervised clustering method over simply adapting to the test conditions. The second two columns show the word-error rate when the condition-specific models were adapted to the test conditions, using all the data in each acoustic focus condition (single cluster). The last two columns are replicated from Table 1 and show the performance after adaptation to the test data clusters (multiple clusters).

We can see that adapting to the individual test conditions gave a relative improvement of 3.4% as compared to the condition-specific models, and adapting to the test data clusters gave a further 6.3% improvement, resulting in a total relative improvement of 9.5% compared to the condition-specific models. Both these improvements are statistically significant at lower than 0.3% level of significance using the paired-sample sign test referred to earlier. It is clear that adapting to the test data clusters gave a consistent improvement compared to adapting to only the test conditions for all acoustic conditions.

5. SUMMARY AND CONCLUSIONS

We presented a novel algorithm for adaptation during testing, which used an unsupervised agglomerative clustering algorithm to cluster the test segments, followed by ML transformation-based adaptation of the condition-specific models to these clusters. A symmetric relative entropy distance between test segments was used for clustering. We described a robust method to estimate the models for each test segment necessary for the computation of the distance measure. It was shown that adapting to all the test data in each focus condition gave a 3.4% decrease in error rate as compared to the condition-specific models. However, adapting to the individual clusters proved to be even more important and gave a 9.5% improvement over the condition-specific models.

REFERENCES

[1] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, 1995.

[2] C. J. Legetter and P. C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression," in *Proceedings of the Spoken Language Systems Technology Workshop*, pp. 110–115, 1995.

[3] A. Sankar and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 190–202, May 1996.

[4] L. Neumeyer, A. Sankar, and V. Digalakis, "A Comparative Study of Speaker Adaptation Techniques," in *Proceedings of EUROSPEECH*, pp. 1127–1130, 1995.

[5] J. Gauvain and C.-H. Lee, "Maximum *a posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, April 1994.

[6] R. Stern, "Specification of the 1996 Hub4 Broadcast News Evaluation," in *Proceedings of the DARPA Speech Recognition Workshop* (Chantilly, VA), 1997.

[7] V. Digalakis, P. Monaco, and H. Murveit, "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 281–289, 1996.

[8] A. Sankar, L. Heck, and A. Stolcke, "Acoustic Modeling for the SRI Hub4 Partitioned Evaluation Continuous Speech Recognition System," in *Proceedings of the 1997 DARPA Speech Recognition Workshop* (Chantilly, VA), 1997.

[9] M. J. F. Gales and P. C. Woodland, "Mean and Variance Adaptation within the MLLR Framework," *Computer Speech and Language*, vol. 10, pp. 249–264, 1996.

[10] H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, "Large-Vocabulary Dictation Using SRI's DECIPHER(TM) Speech Recognition System: Progressive-Search Techniques," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. II–319–II–322, 1993.

[11] A. Sankar, F. Beaufays, and V. Digalakis, "Training Data Clustering for Improved Speech Recognition," in *Proceedings of EUROSPEECH*, 1995.