CLUSTER-DEPENDENT ACOUSTIC MODELING

Bing Xiang, Long Nguyen, Spyros Matsoukas, and Richard Schwartz

BBN Technologies
10 Moulton St., Cambridge, MA 02138, USA
{bxiang,ln,smatsouk,schwartz}@bbn.com

ABSTRACT

In this paper, we present *cluster-dependent acoustic modeling* for large-vocabulary speech recognition. With large amount of acoustic training data, we build multiple *cluster-dependent models* (CDM), each focusing on a group of speakers in order to represent speaker-dependent characteristics. It is motivated by the fact that a sufficiently trained speaker-dependent (SD) model is better than the speaker-independent (SI) model. During decoding, we decode the data of each test speaker using CDMs selected under certain criteria to achieve high recognition accuracy. Various speaker clustering and model selection techniques are proposed and compared in the task of Broadcast News (BN) transcription. The CDM provided more than 1% absolute gain in unadapted decoding and 0.5% gain in adapted decoding when compared to our baseline system on the EARS BN 2003 development test set.

1. INTRODUCTION

Recently, large amount of acoustic training data have become available for large-vocabulary speech recognition. How to effectively utilize these data is important. A simple way is to train a large model with increased number of parameters [1]. But large model may eventually be saturated at certain point and will also slow down the recognition, which is a critical problem for real-time applications.

Instead of training a large model, we have been experimenting with multiple cluster-dependent models (CDM), with each of them focusing on a group of speakers. Our goal is to approach the performance of speaker-dependent (SD) model because we know that, when having sufficient training data, the SD model is better than the speaker-independent (SI) model. Since building an SD model for each test speaker for the BN transcription task is not realistic, CDM can be regarded as a feasible compromise between the SD and SI models.

During decoding, we select one or several CDMs for each test speaker under certain criteria, e. g. using speaker identification (speaker ID) technique. As presented later, different techniques for speaker clustering and model selection are compared. Overall, the CDM resulted in more than 1% absolute gain in unadapted decoding and 0.5% gain in adapted decoding when compared to our baseline on the EARS BN 2003 development test set.

The paper is organized as follows. In Section 2, we describe the training procedure of CDM, including speaker clustering and model adaptation. The model selection strategy and model merging are presented in Section 3. We then report experimental results in Section 4 and conclude in Section 5.

2. CDM TRAINING

The high-level procedure of the CDM training is depicted in Figure 1. We first clustered the training data into several speaker clusters. Then for each cluster, a CDM was trained on the data from that cluster.

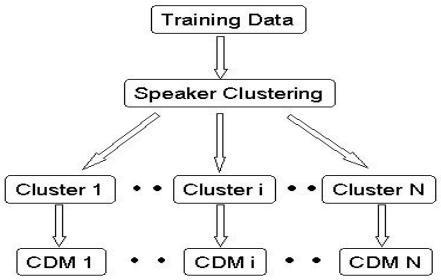

Figure 1: High-level structure of CDM training

2.1. Speaker Clustering

In this work, two different speaker clustering techniques are compared. The first is the online speaker clustering approach developed recently at BBN [2]. For each incoming speaker turn, a decision was made based on the single Gaussian covariance likelihood ratio between the new speaker turn and the current clusters. The new data was either merged with one of those clusters or used to create a new cluster. Since there is no need to look back, this technique is computationally efficient.

We also explored another clustering approach using one Gaussian per state (1gps) State Clustered Tied Mixture (SCTM) model [3]. As shown in Figure 2, first we randomly selected N initial training speakers and adapted the general *1gps* model through Maximum *a posteriori* (MAP) adaptation [4] into N cluster-specific *1gps* SCTMs. Each *1gps* model has around 1000 codebooks, with one Gaussian per codebook. Then we scored all the speaker turns in the training data with the *1gps* SCTMs. Based on the likelihood given by the *1gps* models, we divided the training set into N clusters. Then we ran several iterations of MAP and scoring to get the final clusters and also the *1gps* SCTMs that were to be used by the speaker ID (or model selection) during recognition.

2.2. Model Adaptation

Instead of training the CDMs from scratch, we chose to adapt the general SI or Speaker-adaptive training (SAT) model [5]. The reason for doing that is because the data from each cluster on average is just a small subset of the original training data. We may not be able to train a good model with such small amount of data. On

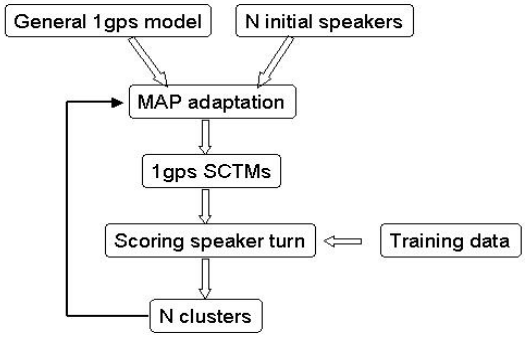

Figure 2: Diagram of 1gps-based speaker clustering

the contrary, we can take advantage of the well-trained general model and adapt it into a cluster-specific model by only changing the Gaussian parameters while keeping the same model structure.

As shown in Figure 3, the general model was first adapted with the Maximum Likelihood Linear Regression (MLLR) adaptation [7] and transformed into the i-th MLLR model for Cluster i. Then the MLLR model was adapted into the final CDM via MAP adaptation [4]. Both stages of adaptation can be iterative with the models updated in each iteration. In this work, both the means and variances for each Gaussian mixture were adapted. The mixture weights were all taken from the general model.

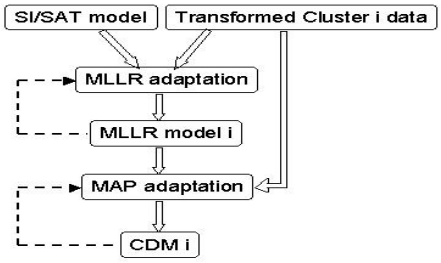

Figure 3: Diagram of adapting general model into CDM

Two sets of CDMs were trained with similar training procedures and used for unadapted decoding and adapted decoding, separately. During the training of the CDM to be used in unadapted decoding, the general SI model was adapted and the training data was transformed with a global Heteroscedastic Discriminant Analysis (HDA) transform [8] – the same transform used for training the SI model. While for adapted decoding, the general SAT model acted as the seed model. We transformed the cluster-specific data with the speaker-dependent HDA and Constrained Maximum Likelihood Linear Regression (CMLLR) transforms [6], which were used in the training of the general SAT model. With the transformed cluster-specific data, the general SAT model was adapted to the MLLR models and then the CDMs.

3. CDM SELECTION AND MERGING

3.1. Model Selection

During decoding, we first determine the most likely CDM for each test utterance or test speaker. Then we use the selected CDM to decode those data. We can see that the step of model selection is critical for the final recognition performance. The more accurate the model selection is, the better recognition performance we can achieve.

Three approaches were compared in this work. Two of them correspond to the speaker clustering techniques presented above, the online speaker clustering and the *1gps*-based speaker clustering. The third approach is the Gaussian Mixture Model (GMM) based speaker ID [9].

Single-Gaussian-Based Selection In the single-Gaussian-based model selection, we found the most likely cluster among the training clusters using a technique similar to the online speaker clustering. The decision was based on the single Gaussian covariance likelihood ratio between the Gaussian estimated on the test speaker data and those Gaussians corresponding to current clusters.

1gps-Based Selection We also tried a different model selection in adapted decoding, which is corresponding to the *1gps* speaker clustering. First we estimated the speaker-dependent feature transforms based on the 1-best hypotheses from the unadapted decoding. Then we scored each test speaker with the *1gps* SCTMs and selected the model that gave the highest likelihood for the test data. The corresponding CDM was adapted to the 1-best hypotheses with the MLLR adaptation and used to decode the data from the specific test speaker.

GMM-Based Selection The third approach we compared with is the GMM-based speaker ID. We trained a 1024-component background GMM using a 10-hour data subset randomly chosen from the training data set. Then we adapted the background GMM into multiple speaker GMMs corresponding to each CDM. During recognition, for each frame of the test utterance, the top five mixture components of the background GMM were identified first. We then evaluated the likelihood of each of the speaker GMMs by using only five corresponding mixture components. The CDM selected for an utterance is the one such that its corresponding speaker GMM produced the highest likelihood.

3.2. Model Merging

Besides using the top 1 CDM identified via speaker ID for decoding, we also tried merging the several top CDMs to enhance the robustness. The corresponding Gaussians in these CDMs were merged together with certain weights to build a new model for each test speaker. The reason we can do that is because all the CDMs were originally derived from the same seed model. The merged model was further adapted to the test data using MLLR before being used for decoding. There are various ways of setting the merging weights. They can either be uniform weights or be estimated based on the likelihood from speaker ID.

Two types of unequal weights were proposed in this work. The first was based on the log likelihood ratio between the top models, as

$$w_i = \frac{logp_i - logp_{m+1}}{\sum_{j=1}^{m}(logp_j - logp_{m+1})}, \quad (1)$$

where p_i is the likelihood from the i-th most likely model and m is the number of models to merge. The log likelihood from the $(m+1)$-th most likely model serves as a bottom line here.

Another weighting was based on the *a posteriori* probability,

$$\hat{w}_i = \frac{p_i}{\sum_{j=1}^{m} p_j}. \quad (2)$$

4. EXPERIMENTAL RESULTS

4.1. Corpus and Recognizer

The acoustic training corpus used in this work is the 843-hour BN data set. It includes the 141-hour Hub-4 acoustic training data and 702 hours of automatically selected TDT data [1]. All of the speakers in this data set were clustered into 80 clusters to be used in the training of 80 CDMs. The number of clusters was determined such that the average amount of training data for each cluster is around 10 hours, which we think is appropriate for the large aoustic model used in this work (around 700k Gaussians). The test material is the EARS BN 2003 development test set. It consists of 3 hours of speech from 6 broadcasting sources. We also validated our results on another 3-hour test set, the EARS BN 2004 development set.

In our Byblos BN transcription system, the decoding process contains two stages, the unadapted decoding and adapted decoding. First, the SI model (or CDM derived from SI model) generated hypotheses for unsupervised adaptation. Then, the decoding was repeated but with the SAT model (or CDM derived from SAT model) that has been adapted to the hypotheses generated in the first stage.

The original 60-dimensional feature vector consists of 14 PLP cepstral coefficients [10], energy and their first, second and third derivatives. It is furthered transformed into a 46-dimensional feature vector with a global HDA transform or speaker-dependent HDA and CMLLR transform. The frame rate is 10 ms.

4.2. Unadapted Decoding

The word error rates (WER) of the unadapted decoding stage are listed in Table 1. As we can see, the WER was reduced for all shows. The absolute reduction on each show ranges from 0.2% to 3.1%. Overall, we obtained 1.2% reduction (12.6% vs. 13.8%) compared to the baseline using the SI model.

Model	ABC	CNN	MSN	NBC	PRI	VOA	All
SI	12.6	19.9	10.5	10.5	9.6	19.8	13.8
CDM	12.1	18.0	9.4	10.1	9.4	16.7	12.6

Table 1: Unadapted decoding results for the SI baseline and the CDM

We also investigated the effects of model adaptation during the *training* of the CDM. As shown in Table 2, when using only MAP adaptation, the overall WER decreased to 13.2%, or 0.6% absolute compared to the baseline 13.8% in Table 1. When using only MLLR adaptation, the overall WER decreased from 13.8% to 12.8%, or 1.0% absolute. Adding one iteration of MAP adaptation after two iterations of MLLR adaptation provided another 0.2% gain.

The single-Gaussian (SG) speaker clustering is also compared with the GMM-based speaker ID in Table 2. There is no difference between these two in terms of the final WER.

MLLR iter	MAP iter	Speaker ID	WER
0	1	SG	13.2
2	0	SG	12.8
2	1	SG	12.6
2	1	GMM	12.7

Table 2: Comparison of various transforms and speaker ID (SG: Single-Gaussian-based speaker ID)

In order to understand the effect of the CDMs, we carried out a cheating experiment in which each test utterance was decoded 80 times using 80 different CDMs. As expected, the overall WER for the entire test set when using one of the 80 CDMs was higher than that of the SI baseline model. The WERs ranged from 14.9% to 20.7% as shown in Table 3. However, if we selected only the hypothesis with the lowest error rate among the 80 hypotheses for each test utterance, the WER for the entire test set would be 8.7%. In other words, this is the WER we would have achieved if we could select the right CDM for every utterance. This Oracle WER implies that there is plenty of room to improve our model selection procedure. Similarly, the result of the cheating selection at the speaker level is listed in Table 3. It is higher than the Oracle WER selected at the utterance level. This is an indication that the automatically grouping of utterances into a speaker is not optimal.

System	WER
SI	13.8
Single CDM	14.9 - 20.7
Speaker ID + CDM	12.6
Oracle (speaker-level)	11.5
Oracle (utterance-level)	8.7

Table 3: Real error rates compared to oracle error rates

4.3. Adapted Decoding

The recognition WERs in the adapted decoding stage, when using the SAT-CDMs trained under different setups, are shown in Table 4. The first row displays the baseline WER when using the standard SAT model. In the remaining four experiments using the SAT-CDMs, we tried adapting only the means or both the means and the variances of the mixture densities. Adapting both the means and variances helped a little bit. For model selection, we tried all of the three approaches, the SG-based, GMM-based or *1gps*-based speaker ID. They resulted in the same performance. The lowest WER that the SAT-CDMs achieved was 10.8%, or 0.2% absolute reduction in comparison to the baseline result. This is much smaller than what we obtained in the unadapted decoding stage.

To understand why we got smaller gain in adapted decoding, we calculated the *Kullback-Leibler* (K-L) divergence between the Gaussians of multiple *1gps* SCTMs. As shown in Table 5, the K-L divergence was reduced by almost a factor of three after applying the speaker-dependent transforms. So the CDMs trained in the speaker-dependent-transformed feature space are closer to each other than those trained in the global-HDA-transformed feature space. This could be part of the reasons for the small gain

Model	Adapted	Speaker ID	WER
SAT	N/A	N/A	11.0
CDM	Mean	SG	10.9
CDM	Mean,Var	SG	10.8
CDM	Mean,Var	GMM	10.8
CDM	Mean,Var	1gps	10.8

Table 4: Adapted decoding results

from the CDM in adapted decoding.

Transform	K-L Divergence
Global HDA	12.9
Speaker-dependent transform	4.7

Table 5: K-L divergence after different feature transforms

4.4. Model Merging

In order to achieve further improvement, we tried model merging. The top 3 or 5 CDMs selected by the speaker ID were merged into a new model. As shown in Table 6, with equal weights, the merging of the top 3 CDMs gave 0.3% gain compared to the baseline. Another 0.1% gain was obtained from merging the top 5 CDMs. No further gain was observed when merging the top 8 CDMs. With weights w_i, determined by the log likelihood ratio between the top five models and the the sixth model, we achieved 0.5% absolute gain compared to the baseline. The weights $\hat{w}_i$ which were based on the *a posteriori* probability didn't outperform the weights w_i.

Model	No. of Models	Weights	WER
SAT	1	1	11.0
CDM	3	1/3	10.7
CDM	5	1/5	10.6
CDM	8	1/8	10.7
CDM	5	w_i	10.5
CDM	5	$\hat{w}_i$	10.7

Table 6: Results of model merging

4.5. Validation on Another Test Set

To validate the results we obtained from the CDM, we also tested on another test set, the EARS BN 2004 development set. As shown in Table 7, CDM provided 1.3% absolute gain in unadapted decoding and 0.6% gain in adapted decoding, similar with what we obtained on the previous test set.

5. CONCLUSION

As reported in this paper, CDMs gave more than 1% absolute gain in unadapted decoding and 0.6% gain in adapted decoding on the EARS BN development test sets. We also noticed that the gap in performance between unadapted and adapted decoding was reduced. Different speaker clustering and speaker ID techniques did not bring much difference in terms of the final WER. Model merging and system combination provided further gains.

System	Unadapted decoding	Adapted decoding
Baseline	16.0	12.9
CDM	14.7	12.3

Table 7: Results on EARS BN 2004 development test set

All the general models used in this work were trained under maximum likelihood (ML) criterion. Currently we are working on the Maximum Mutual Information (MMI) based CDMs by adapting the general MMI model into multiple MMI-CDMs via MMI-MAP [11]. We also saw that the WER is still high when compared to the Oracle WER. This indicated a large room for improving the model selection, which is part of our future work. We will also apply the CDMs to other speech recognition task, such as recognition of the conversational telephone speech.

References

1. L. Nguyen and B. Xiang, "Light supervision in acoustic model training," *Proc. ICASSP*, Montreal, Canada, May 2004.
2. D. Liu and F. Kubala, "Online speaker clustering," *Proc. ICASSP*, Apr. 2003.
3. L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz and J. Makhoul, "Progress in transcription of broadcast news using Byblos," *Speech Communication*, 38, pp. 213-230, 2002.
4. J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291-298, Apr. 1994.
5. T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," *Proc. ICSLP,* Philadelphia, PA, Oct. 1996.
6. S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections," *IEEE ASRU Workshop*, St. Thomas, Nov. 2003.
7. C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171-186, 1995.
8. N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication,* vol. 26, no. 4, Dec. 1998.
9. D. A. Reynolds, T. Quatieri and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.
10. H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738-1752, April 1990.
11. D. Povey, P. C. Woodland and M. J. F. Gales, "Discriminative MAP for acoustic model adaptation," *Proc. ICASSP*, Apr. 2003.