Speaker Clustered Regression-Class Trees for MLLR Adaptation

Arindam Mandal[1,2], *Mari Ostendorf*[1], *Andreas Stolcke*[2]

[1]Department of Electrical Engineering, University of Washington, Seattle, WA, USA
[2]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA

{marindam,mo}@ee.washington.edu, stolcke@speech.sri.com

Abstract

A speaker clustering algorithm is presented that is based on an eigenspace representation of Maximum Likelihood Linear Regression (MLLR) transformations and is used for training cluster-dependent regression-class trees for MLLR adaptation. It is shown that significant automatic speech recognition (ASR) system performance gains are possible by choosing the best regression-class tree structure for individual speakers. To take advantage of the potential gains, an algorithm for combining the MLLR mean transformations from cluster-specific trees is described that effectively results in a soft regression-class tree. In conversational speech recognition, only small overall improvements are obtained, but the number of speakers that have performance degradation due to adaptation is reduced by over 70%.

Index Terms: speech recognition, speaker adaptation, speaker clustering, regression class trees.

1. Introduction

In current automatic speech recognition (ASR) systems, a single speaker-independent (SI) regression-class tree is used in MLLR adaptation with run-time pruning to decide on the number of regression classes to use for each new speaker. The SI tree is often built using data-driven clustering techniques on training speakers using some criterion related to similarity of acoustic units. Hypothesizing that the transformations may be used to characterize dialect-related pronunciation variation as well as speaker-specific variation, this work investigates the use of different regression-class tree structures in speaker clustering. We present a speaker clustering technique to split a large group of training speakers into multiple clusters using an eigenspace representation of their MLLR transformations and train a separate regression-class tree for each cluster. Experiments show that different regression-class tree structures are learned for different clusters of training speakers, and the system performance on unseen test speakers varies across these trees and results in significant performance gains when the best tree structure is chosen for each speaker.

To choose the best regression-class tree structure for an unseen test speaker in an ASR system, we estimate weights for linearly combining the MLLR mean transformations from several trees using a two-step maximum likelihood procedure. Small improvements in system performance are achieved by using this approach on recent NIST conversational speech recognition test sets. In addition, it leads to improved performance on a majority of those speakers who do not benefit from MLLR adaptation using the SI regression-class tree.

The rest of the paper is organized as follows: Section 2 describes two types of regression-class trees used in this work; Section 3 describes the ASR system and corpus used; Section 4 discusses the speaker clustering procedure and oracle system performance gains; Section 5 details an algorithm for combining MLLR transformations from several regression-class trees and its performance. Finally, Section 6 concludes by summarizing the main findings.

2. Regression-Class Trees

The use of regression-class trees in MLLR adaptation [1] of HMMs has proven to be an effective strategy for adapting acoustic models (Gaussian distributions). The acoustic units are organized into a tree using either expert knowledge or by applying a data-driven clustering procedure (agglomerative or divisive) and an appropriate similarity measure for comparing the units. Given adaptation data from a test speaker, the tree is descended from the top to those nodes that satisfy a predetermined minimum count of data frames, and a transformation is estimated for each such node (regression class) to be shared by all its members, allowing for adaptation of both observed and unobserved units.

We have experimented with two divisive clustering approaches for building regression-class trees: *constrained* and *unconstrained*. In both cases, we start by estimating multivariate Gaussian distributions for each triphone state, and collect these to obtain phone-level sufficient statistics. Then, in the constrained approach, we design a decision tree to cluster the Gaussians, choosing from linguistically-motivated questions about the center phones to maximize likelihood, similar to clustering triphone states of HMMs using decision trees [2]. In the unconstrained approach, we build a binary tree by splitting the distributions at each level into two clusters, using k-means clustering and a symmetric

Kullback-Liebler distance measure. Both trees are grown to the point where all leaf nodes correspond to a single phone. The two types of trees have the same set of leaf nodes but different branching structure leading to the leaves, which leads to different adaptation results because it is often the case that the estimated transforms correspond to internal nodes, given limited adaptation data.

3. ASR System and Corpus

The ASR system used for this work is SRI International's Decipher large vocabulary engine [3]. We used a version of the system that runs in five times real time and achieves competitive performance on conversational telephone speech (CTS) tasks. For first-pass decoding, the system uses word-internal triphones as HMMs, Mel-frequency cepstral coefficients (MFCCs) in the front end, a bigram language model, and phoneloop MLLR to produce lattices. Subsequently, these lattices are expanded using a 4-gram language model and also processed using confusion networks to produce higher quality hypotheses. The second pass of the system uses acoustic models based on crossword triphones and perceptual linear prediction (PLP) feature vectors as the front end. The PLP-based acoustic models also use, among other standard normalization techniques, speaker adaptive training based on constrained MLLR. The mean vectors and diagonal covariances of the Gaussian distributions of the PLP-based acoustic model are adapted to test speakers using MLLR and the hypotheses from the first stage. A full matrix transformation with an offset vector for the Gaussian means and a diagonal variance transformation vector are estimated using either type of regression-class tree described above, for each node with a minimum data count of 1700 frames, which results in an average of eight regression classes to be used for speakers in the NIST CTS test sets. This is effectively cross-system adaptation since adaptation hypotheses from MFCC-based acoustic models are used to adapt PLP-based acoustic models. For the rest of the paper, all references to MLLR adaptation are for the second stage of this system.

In this work, the corpus used for speaker clustering was comprised of conversations of speakers from the Fisher Phase 2 [4] corpus that were not used in training the ASR acoustic model and the recent NIST CTS test sets (1998-2002), which together included 1186 male speakers and 567 female speakers. Only the NIST test sets from 1998-2002 were used for the error analysis, since the transcriptions are more reliable for this data. Evaluation is on the independent Eval03 test set.

4. Cluster-Dependent Regression-Class Trees

4.1. Speaker Clustering based on MLLR Transforms

We developed a procedure that splits a large group of training speakers into multiple clusters based on an eigenspace representation of MLLR transforms. Eigenspace-based MLLR representations were found to be useful for gender classification [5] and faster speaker adaptation in ASR [6]. In [7], MLLR transformation-based representation were found to improve speaker recognition accuracy. In contrast, our approach aims to form speaker clusters using eigenspace representations of MLLR transformations and train a regression-class tree for each speaker cluster, to capture cluster-specific attributes in the structure of the trees. The idea is that choosing the correct regression-class tree structure to use with MLLR adaptation should lead to improved performance.

We first estimate MLLR transformations for a large corpus of speakers using an SI constrained regression-class tree that has R regression classes ($R = 8$). Then we vectorize the MLLR transformations (mean transform, offset vector) to produce a $d(d+1)$-length vector for a d-dimensional feature vector ($d = 39$) and normalize each dimension to have zero mean and unit variance. Next, we perform principal component analysis (PCA) using the vectorized MLLR transformations of all regression classes of all speakers, except the ones corresponding to the non-speech class. The vectorized transforms are then projected onto the first N principal components ($N = 8$), and we form an $(R-1)N$-dimensional supervector for each speaker by stacking together the PCA-reduced MLLR transforms for each of the $R-1$ classes (except the non-speech class). Finally, we use k-means clustering to partition the speakers into S clusters ($S = 4$). The supervectors capture the speaker-dependent information present in MLLR transformations, and the clustering groups together speakers that share similar transform characteristics. We then train a separate regression-class tree, both constrained and unconstrained, for each speaker cluster. To train the two kinds of cluster-specific regression-class trees, we apply the procedures described in Sec. 2 on triphone-level statistics collected for each cluster.

As expected, the cluster-dependent regression-class trees have different branching structures. Since the structure of regression-class trees describes similarities among clusters of phones (based on phone-level statistics), we conjecture that each cluster-specific tree to be representative of dialect or pronunciation patterns that are representative of its cluster. We manually compared the regression-class trees trained for each speaker cluster for any noticeable cluster-specific characteristics. Among the constrained regression trees, we found that the branches of the trees that split the acoustic units describing vowels exhibited more differences in the hierarchical structure than the branches involving the consonants, which is consistent with linguistic studies on regional variation in American English [8]. The unconstrained trees had structures that were considerably different from those of the constrained trees and also exhibited more diversity in structure details across the clusters than the con-

strained ones. We experimented with several different values of N, the number of principal components for projecting the vectorized MLLR transforms, and chose $N = 8$, since it produced the most diversity in the structure of the cluster-specific trees.

4.2. Oracle Cluster-Dependent Adaptation

To determine the potential performance gains from these trees, we computed the recognition error rates for the speakers in recent NIST CTS test sets (1998-2002) for every regression-class tree. For experiments in this section, we ignored around 1% of these speakers for whom there was no change in performance across the different regression trees, and assigned the rest of the speakers to that cluster whose tree achieved the best performance, and recomputed the overall word error rate (WER) for each new speaker cluster with every regression-class tree. The results for the unconstrained tree are shown in Table 1, where the rows represent test sets for each speaker cluster and the columns the cluster-specific regression-class trees. Obviously, the best WER is seen for all cases when the cluster-specific test set matches its target regression-class tree, i.e., the numbers along the diagonals of the tables. Additionally, the upper bound of potential gains over the SI tree are in the ranges of 0.6-0.8% absolute for the unconstrained tree. On analyzing the performance numbers for each speaker, we noticed that when the cluster-specific test set matches its target regression-class tree, the error rates for the worst performing speaker improves by 0.5-1.9% absolute, and the speaker-specific error rates have a lower standard deviation, in the range of 1.5-2.0% relative compared to the SI tree. Similar observations are made on analysis of the performance figures from the constrained trees, and are not presented here for brevity.

	Clust 1	Clust 2	Clust 3	Clust 4	SI
Clust 1	**20.5**	21.2	21.2	21.3	21.3
Clust 2	22.0	**21.3**	22.1	22.1	21.9
Clust 3	24.1	24.4	**23.6**	24.0	24.3
Clust 4	21.6	21.8	21.6	**20.9**	21.7

Table 1: WER(%) with unconstrained trees.

Test Set	Clust 1	Clust 2	Clust 3	Clust 4	SI
Clust 1	21.2	21.1	21.0	21.1	21.3
Clust 2	21.9	21.9	22.0	21.9	21.9
Clust 3	23.8	23.8	23.9	24.1	24.3
Clust 4	21.4	21.4	21.4	21.4	21.7

Table 2: WER(%) with retrained unconstrained trees.

To gain a better understanding of the performance of cluster-specific regression class trees for individual speakers, we used the new assignment of speakers to that cluster whose tree produced the lowest WER, retrained the unconstrained regression-class trees for each cluster using its phone-level statistics and the procedure in Sec. 2, and performed the error analysis just described. However, the results from this analysis, shown in Table 2, do not exhibit patterns similar to those in Table 1, which indicates the existence of a more complex relationship between speaker cluster membership and performance obtained from cluster-specific regression-class trees.

5. Soft Regression-Class Trees

Based on the evidence presented in the previous section, we developed a procedure to incorporate the cluster-specific regression-class trees into the MLLR framework, using a linear combination of the transforms learned using the individual trees. Since the structure of the regression-class trees differs across speaker clusters, each tree will specify a different transformation tying scheme, and the weighted combination approach can be viewed as a "soft" assignment of acoustic units to regression classes. The expectation is that the resulting "smoothed" transform should be more robust than that from a single tree in the case where the oracle cluster assignment is not known.

5.1. Weight Estimation

In this work, we have focused on only linearly combining the mean transformations by estimating weights that maximize the likelihood of a speaker's adaptation data. Previous work on combining transformations using weights has been reported in [1, 9]. The work in [1] estimates SI or speaker-dependent (SD) weights for transformations for different regression classes of the same tree, while that of [9] involves combining transformations that are representative of a speaker of a given speaker cluster. The key difference in our approach is that for an unseen test speaker we estimate weights to combine transformations, of the same speaker, that have been estimated using regression classes representative of different speaker clusters.

Define the transformed mean vector of the m-th Gaussian as

$$\hat{\mu}_m = \hat{\mathbf{M}}_m \hat{\alpha}^{(l)}$$

where

$$\hat{\mathbf{M}}_m = [\hat{\mu}_m^{(1)} \cdots \hat{\mu}_m^{(S)}], \quad \hat{\mu}_m^{(s)} = \hat{\mathbf{W}}^{(s,r)} \xi_m,$$

where $\hat{\mathbf{W}}^{(s,r)}$ is the transformation associated with the r-th regression class of the s-th speaker cluster on the extended mean vector ξ_m, and

$$\hat{\alpha}^{(l)} = [\hat{\alpha}_1^{(l)} \cdots \hat{\alpha}_S^{(l)}]^T$$

where $\hat{\alpha}_s^{(l)}$ are the weights for the mean transformation at the l-th node of the s-th regression-class tree. Using a procedure similar to [1], the weights $\alpha_s^{(l)}$ can be estimated by solving

$$\Big[\sum_{r=1}^{R}\sum_{c=1}^{C_r}\sum_{m=1}^{M_c}\sum_{\tau=1}^{T}\gamma_m(\tau)\hat{\mu}_m^{(s)T}\Sigma_m^{-1}\hat{\mathbf{M}}_m\Big]\hat{\alpha}^{(l)} = \sum_{r=1}^{R}\sum_{c=1}^{C_r}\sum_{m=1}^{M_c}\sum_{\tau=1}^{T}\gamma_m(\tau)\hat{\mu}_m^{(s)T}\Sigma_m^{-1}\mathbf{o}(\tau),$$

or

$$\mathbf{Z}^{(l)}\alpha^{(l)} = \mathbf{V}^{(l)} \quad (1)$$

where R is the number of regression classes containing C_r mixture Gaussian distributions, each of which has M_c component Gaussian distributions; $\mathbf{o}(\tau)$ is the observation vector at time τ and $\gamma_m(\tau)$, $\hat{\mu}_m$ and Σ_m^{-1} are the occupation probability at time τ, mean vector, and inverse covariance of the of the mth Gaussian distribution.

The maximum likelihood solutions for the weights do not have constraints on them. To handle instances when numerical instability leads to bad estimates of weights, we introduce Lagrange multipliers into the objective function, to constrain the weights to be non-negative and sum to one. It is straightforward to solve for the Lagrange multipliers, and the details are not provided here.

5.2. Performance of Soft Regression Trees

For a given test speaker, we first estimate the mean and diagonal variance MLLR transformations for every cluster-specific regression-class tree from HMM state occupation statistics collected using the speaker's adaptation data and the unadapted SI acoustic model and, we also determine the tree that produces the highest gain in likelihood on the adaptation data. Using the same statistics, we next estimate the mean transformation smoothing weights, without any inequality constraints (as described in the Sec. 5.1) and its corresponding diagonal variance transformation, and determine its likelihood gain on the adaptation data. If the gain is less than the best gain from the individual cluster-specific trees, we estimate the smoothing weights with inequality constraints, and its corresponding diagonal variance transformation. Finally, the SI acoustic model is adapted using the appropriate set of transformations.

We also experimented with tying the mean transformation smoothing weights at the root (global weights) and at the leaves of the regression-class trees, similar to the tying of regression classes in the tree. In Table 3, we report results of using soft regression-class trees on the NIST CTS test set for year 2003. The results in Table 3 show small improvements over the baseline. On analyzing the performance of individual speakers with the soft regression class trees, we found that about 74% of speakers use weights estimated without constraints. Additionally, we noticed that, while for about 15% of the speakers the WER increases after MLLR adaptation using the baseline system (with SI regression-class tree), a majority of these speakers (about 70%), benefit from using the soft regression-class trees.

	WER(%) on Eval 2003	
Configuration	Constrained	Unconstrained
Baseline	21.4	21.5
+ML weights(root)	21.3	21.4
+ML weights(leaves)	21.3	21.3

Table 3: Results with soft regression-class trees

6. Conclusions

The two main contributions of this work are a speaker clustering algorithm for building cluster-specific regression-class trees, and a framework for using soft regression-class trees for unseen test speakers by estimating weights for a linear combination of MLLR mean transformations corresponding to cluster-specific regression-class trees. The trees built for the different speaker clusters reveal partitions of acoustic units that are possibly indicative of cluster-specific characteristics that can be used for automatic dialect clustering. The soft regression-class trees are able to achieve only a small overall improvement in performance. Since the oracle clustering results hold promise of bigger performance wins, we plan to investigate further refinements to the weight estimation procedure.

Acknowledgments

The authors thank V. R. R. Gadde and Jing Zheng for their helpful suggestions with MLLR implementation details in the Decipher system. This work was funded by DARPA under contract No. HR0011-06-C-0023. Distribution is unlimited. The views herein are those of the authors and do not reflect the views of the funding agencies.

7. References

[1] M.J.F. Gales, "The generation and use of regression class trees for MLLR adaptation," Tech. Rep. CUED/F-INFENG/TR263, Cambridge University, 1996.

[2] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree based state tying for high accuracy modelling," in *Proc. ARPA Spoken Language Technology Workshop*, 1994.

[3] A. Stolcke et al., "Recent innovations in speech-to-text transcription at SRI-ICSI-UW," in *IEEE Transactions on Speech, Audio and Language Processing*, to appear.

[4] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *LREC*, May 2004.

[5] C. Huang, T. Chen, S. Li, E. Chang, and J. Zhou, "Analysis of speaker variability," in *Proc. of Eurospeech*, 2001.

[6] K. T. Chen, W. W. Liau, H. M. Wang, and L. S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Proc. of ICSLP*, 2000.

[7] A. Stolcke et al., "MLLR transforms as features in speaker recognition," in *Proc. of Eurospeech*, September 2005.

[8] W. Labov, "The organization of dialect diversity in North America," in *ICSLP4*, Philadelphia, 1996.

[9] C. Boulis, V. Diakoloukas, and V. Digalakis, "Maximum likelihood stochastic transformations adaptation for medium and small data sets," *Computer Speech and Language*, vol. 15, no. 3, pp. 257–287, July 2001.