Leveraging Speaker-dependent Variation of Adaptation

Arindam Mandal, Mari Ostendorf

Department of Electrical Engineering
University of Washington, Seattle, WA, USA
{marindam,mo}@ee.washington.edu

Andreas Stolcke

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, USA
stolcke@speech.sri.com

Abstract

This work introduces an automatic procedure for determining the size of regression class trees for individual speakers using an ensemble of speaker-level features to control the number of transformations, if any, that should be estimated by maximum likelihood linear regression. Experiments with a state-of-the-art speech recognition system that uses this procedure show improvements in word error rate for conversational telephone speech.

1. INTRODUCTION

An important problem in the context of speaker adaptation by maximum likelihood linear regression (MLLR) is the choice of regression class trees to cluster similar acoustic units. These acoustic units share transformations estimated from adaptation data. Automatic speech recognition (ASR) systems use a speaker-independent regression class tree with leaf node pruning based on the amount of adaptation data available. Here, we show that the data-driven pruning does not yield the best regression class tree size, and that significant gains in word error rate (WER) are achievable in the oracle case, which may involve no adaptation. To capitalize on this observation, an automatic procedure is developed to predict the best regression class tree size, if any, for a speaker by using standard statistical learning paradigms and speaker-level acoustic and recognizer features.

This paper is organized as follows: Section 2 briefly reviews MLLR and the use of regression class trees; Section 3 describes the corpus and ASR system used throughout this work; Section 4 demonstrates that significant gains in WER can be achieved by using the best regression class tree size; Section 5 details the procedure for predicting the best regression class tree size for each speaker and provides an analysis of useful features; Section 6 presents results of ASR experiments using the predicted tree size; and finally Section 7 concludes with a summary and discussion of future work.

2. MLLR REGRESSION CLASSES

MLLR [1] is a widely used model-space transformation approach for speaker and environment adaptation. In this work, we have focused on the use of MLLR for transforming the parameters of a speaker-independent acoustic model using data from an unseen speaker. The parameters that are transformed include the means and variances of the Gaussian mixture distributions that model the outputs of states of a hidden Markov model (HMM). For example, the mean vector transformation is

$$\mu_j = W_j \nu_j$$

where W_j is the $n \times (n+1)$ transformation matrix, ν_j is the extended mean vector

$$\nu_j = [1 \ \mu_{j_1} \ \cdots \ \mu_{j_n}]',$$

and j is the distribution index. The transformation matrix is estimated by maximizing the likelihood of data from the target speaker for a given word transcription. In unsupervised adaptation, as used here, the transcript is the errorful output of a previous recognition pass.

It is usually the case that the amount of adaptation data available is not sufficient to reliably estimate transformations for the output distributions of every acoustic unit. Instead, acoustic units are clustered into regression classes using a similarity measure that can be data driven as in [2, 3], or manually performed using knowledge of acoustic phonetics. There is little difference in performance between the two schemes for small tasks [4]. The regression classes are organized into a tree, referred to as a regression class tree. A transformation is computed for each regression class and shared among the acoustic units in it. The conventional approach in ASR systems is to first design a global regression class tree with a certain number of leaf nodes (classes). Next, an online complexity control strategy for the tree is used to determine the number of regression classes to use for an individual speaker and is based on the amount of adaptation data available. The regression classes used correspond to the lowest nodes in the tree that meet the minimum data criterion, i.e. data-driven pruning. The ASR system used in this work uses a regression class tree based on phonetic knowledge. A typical tree with nine clusters (leaf nodes) is illustrated in Figure 1. It is trivial to construct a tree with six clusters by pruning back some nodes in the tree.

3. CORPUS AND ASR SYSTEM

The NIST benchmark test set corpus for conversational telephone speech in English was used for all ASR ex-

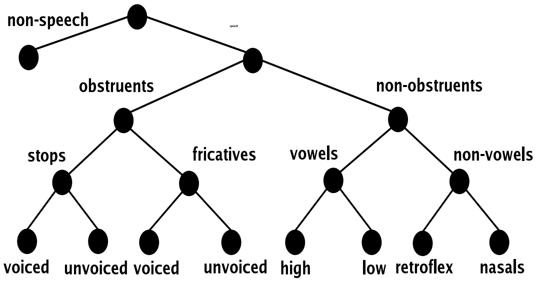

Figure 1: Regression class tree for phone clustering.

periments in this work. The corpus was split into two parts: a training set, comprising the Switchboard and Fisher speakers in benchmark test sets of 1998, 2000, 2001, 2002, and 2003 (464 speakers total) and a test set, comprising the test set of 2004 (72 speakers total).

The ASR system used in this work was developed at SRI International [5]. The architecture of the system is shown in Figure 2 (next page). The system uses unsupervised MLLR in five decoding steps and exchanges adaptation hypotheses between the two different front ends – Mel frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) – to perform cross-system adaptation. At two stages in the system, the hypotheses from the two front ends are combined using confusion networks [6]. These two stages are henceforth referred to as the early stage and the later stage. MLLR adaptation of Gaussian distributions is performed using full mean transforms and variance scaling. The regression class tree used is shown in Figure 1, which employs the hierarchical back-off strategy described in Section 2 when less than 2 seconds of data is available for a regression class. However, a sufficient amount of data was available for the nine phonetic classes for most speakers and the back-off mechanism was rarely needed. On the NIST RT 2004 test set, the system (our baseline) achieved 18.6% WER.

4. BEST TREE SIZES

The system had five different regression class trees available each with 3, 6, 7, 9, or 11 classes. A final possibility is the use of unadapted speaker-independent models. The default setting uses nine regression classes in all stages. Choosing a different size tree can be thought of as using a more sophisticated back-off to determine the transforms in addition to data-driven pruning. To determine the best tree size for each speaker, we ran experiments with each available regression class tree for each of the steps in the system that used unsupervised MLLR. Focusing on recent tests, Table 1 shows that significant gains in WER can be achieved if the best regression class tree can be picked for each speaker.

Test Set	Default	Oracle
eval2004	18.6	17.4
eval2003	18.9	17.9

Table 1: WER(%) with oracle regression class tree sizes.

Analysis of the speakers in the training portion of the corpus provides clear evidence that the best regression class tree size for speakers varies, as shown in the histograms in Figure 3. The two graphs show the distribution of oracle regression class tree sizes in the early (left) and later (right) stages. Not surprisingly, the distribution weight for higher sizes increases in the last stage because of lower WER of the adaptation hypotheses. Also evident from these histograms is the fact that several speakers have the best WER when they are recognized using unadapted models (0 classes).

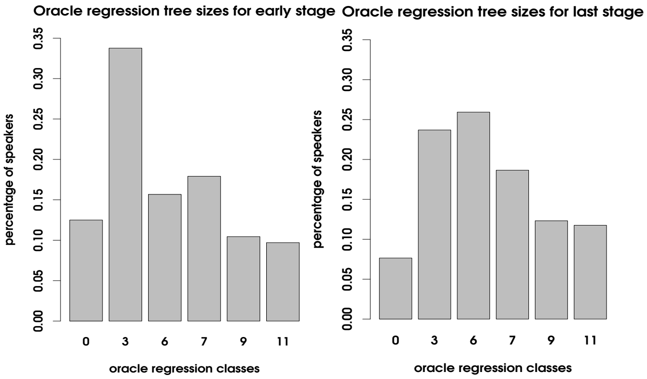

Figure 3: Distribution of oracle tree sizes.

To further understand the variability among speakers, they were clustered into groups in which speakers have similar relative gains (or losses) from different regression class trees. Each speaker was associated with a vector of relative WERs, normalizing the rate for each possible regression class tree with that obtained by the default regression class tree. They were then clustered using k-means with the number of clusters fixed at 5. Figure 4 shows the mean vector of relative WER change for the speakers in a given cluster, from the case using unadapted acoustic models through every available regression class tree. For speakers in clusters 1 and 4, the larger trees (9, 11) are best; for cluster 2, the mid-size trees (3, 6) are best; for cluster 3, there is no advantage to any size; and for cluster 5, no adaptation is the best strategy.

5. PREDICTION OF TREE SIZE

With evidence of potential performance gains from tuning the regression classes, the next step was to automatically predict the regression class tree size for each speaker, based on information observed in the adaptation data. The prediction problem could be formulated in three possible ways: *classification* (tree size is one of the possible six cases), *regression* (tree size is predicted as a real number), and *classification based on cluster membership* as shown in Figure 4. A significant number of speakers had multiple local maxima (i.e., WER change could not be ordered by tree size); hence, the regression approach was not suitable. An experiment was performed using perfect cluster classification, but it produced a gain

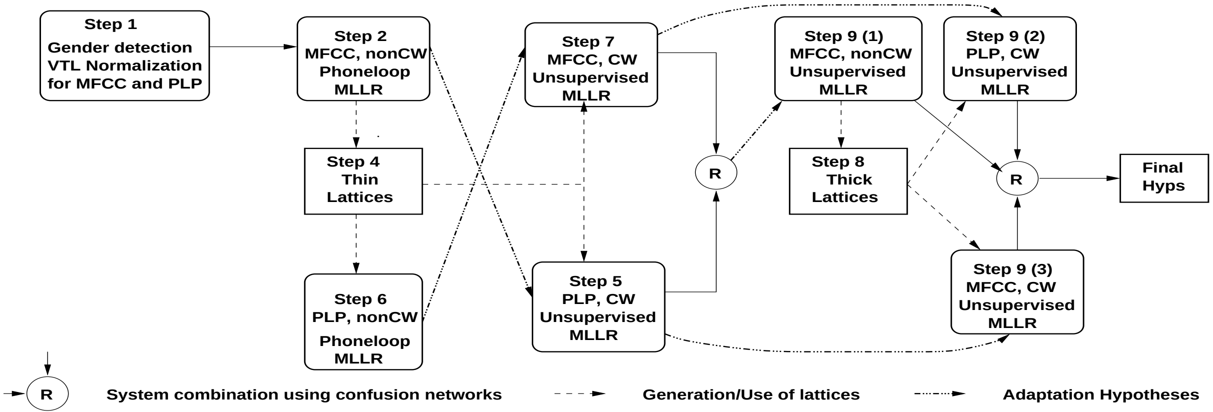

Figure 2: System architecture of a 20xRT (real-time) ASR engine.

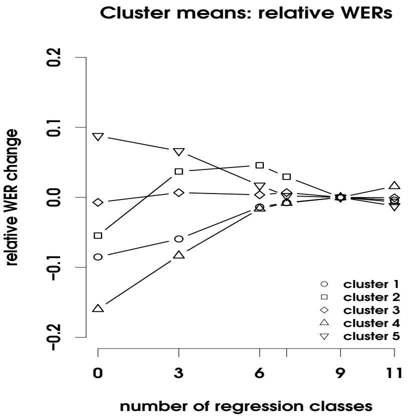

Figure 4: Mean relative change in WER (compared to default) for speaker clusters over different tree sizes.

of only 0.2% in WER compared to around 1%, which can be achieved by using perfect regression class tree size. Based on this evidence, we decided to classify each speaker into one of six possible regression class tree sizes shown in Figure 3.

Several speaker-level features computed from the adaptation data were investigated: acoustic scores per cluster in regression class trees, seconds of speech per cluster, average word-based confidence scores from system combination, normalized energy measure (per frame), vocal tract length normalization factor for MFCC and PLP front ends, and rate of speech (in phones per second). For each of the per-cluster features, there were 11 scores, one for each node in the 11-leaf tree. The seconds of speech per cluster feature was used since it is the standard back-off criterion in the case of insufficient adaptation data. Word-level confidence features have been found to be useful in improving performance of MLLR-based adaptation in [7]. This provided the motivation for using confidence scores in this work. However, we averaged the word-level confidence scores over a speaker to compute a speaker-level confidence measure.

After the raw features are extracted from adaptation data, they were processed as follows. Each training sample represented a speaker and was labeled with the best regression class tree size for that speaker. To compensate for the small number of training samples available, only 464 speakers, the second best regression class tree size was added to the set of training labels, if the WER was not significantly worse than the best regression class tree. Thus, each training sample could have as many as two training labels. In addition, to counter the lack of training data, bagging with replacement was used for selecting training samples. Since the number of samples from each class was the same in the bagged training set, the final classifier posteriors were renormalized with the priors seen in the examples in the training set. Next, we performed dimensionality reduction on the training feature vectors using PCA followed by LDA and used the resulting features to train standard statistical learners.

We used a 4-fold cross-validation training paradigm (1998+2000, 2001, 2002, 2003), designing classifiers on three sets and tuning parameters on the fourth validation set. For each of the classifiers, the number of samples to use from each class for bagging and the number of PCA components were decided by the accuracy on the validation set. The best configuration to use was chosen by varying the bag size per class from 25 to 75 in steps of 25 and the number of PCA components from 10 to 35. The LDA transformation always produced a 5-dimensional feature vector, since the number of possible classes was six. The ensemble of classifiers from the cross-validation partitions were combined to form a stacked learner. The posteriors from each of the classifiers in the stacked learner were then averaged to obtain the final class posteriors.

Several statistical learning paradigms were explored including decision trees, support vector machines, k-nearest neighbor and multinomial neural networks. Decision trees were found to perform best, although the performance of any of the other methods was not significantly different. The overall classification error rate obtained was in the range of 55%-64% for each of the held out sets. The relative reduction in classification error rate compared to chance error was in the range of 4%-20%.

Table 2 shows the percentage of times a particular class of features was used by the decision trees in training classifiers when the allowed subsets of features were recognizer-independent (seconds of speech, VTL, energy, and rate of speech measures), recognizer-dependent (acoustic scores and confidence measures), or all features. Acoustic scores and seconds of speech per cluster are used more frequently than other features. This observation could also be explained by the fact that the number of dimensions representing these two classes of features was much higher than the others, most of which were represented by one dimension.

	% of questions		
Features	Rec. indep	Rec. dep	All
Acoustic Scores	-	85.7	45.2
Seconds of speech	78.2	-	37.2
Confidence	-	14.3	8.6
VTL	10.4	-	5.4
Energy	5.7	-	2.2
Rate of speech	5.7	-	1.4

Table 2: Features usage in the decision trees that were trained on different feature subsets.

6. RECOGNITION EXPERIMENTS

Various combinations were experimented with in predicting the regression class tree size for individual speakers in the NIST RT 2004 test set using the decision tree described in Section 5. In Table 3, column 2 is the case where a single prediction of regression class tree sizes is used throughout the system (auto1); column 3 is the case where two different sets of predicted regression class tree sizes are used: one each for the early and later stages (auto2); and column 4 is the case where predicted regression class tree sizes are used only in the later stages (auto3). The results of recognition experiments are shown in rows 10-12 in Table 3 for different feature subsets, which correspond to columns 2-4 in Table 2. Compared to the baseline (18.6% WER), gains are seen for all cases where predicted regression class tree sizes are used. However, compared to the case when oracle regression class tree sizes are used (17.4% WER), there is substantial room for further improvement. The strategy that uses different predicted regression class tree sizes for the early and later stages (auto2) produces the best improvement in WER, an absolute 0.4% compared to the baseline, using recognizer-dependent features, which is statistically significant at the level $p = 0.002$ according to the matched pair sentence segment test. However, differences relative to using a single tree prediction are not significant. Contrary to trends in the oracle case, the average predicted regression class tree size was 7 for the early stage and 3 for the later stage.

	auto1	auto2	auto3
step 5	P2+6	P2+6	9
step 7	P2+6	P2+6	9
step 5+7	P2+6	P2+6	9
step 9(1)	P2+6	P5+7	P5+7
step 9(2)	P2+6	P5+7	P5+7
step 9(3)	P2+6	P5+7	P5+7
	WER(%) for Eval2004		
Default	18.6	18.6	18.6
Rec. indep	18.3	18.3	18.4
Rec. dep	18.3	**18.2**	18.5
All	18.3	18.3	18.4
Oracle	17.4	17.4	17.4

Table 3: Results of using predicted regression class tree sizes with features from steps X and Y (PX+Y)

7. CONCLUSIONS

This work shows that significant improvement in WER can be achieved by selecting the correct size of regression class trees for individual speakers, including the possibility of no adaptation. Initial efforts at developing an automatic procedure to predict the regression class tree sizes have yielded modest improvements in WER. Analysis of the features used for prediction shows that acoustic scores of adaptation data along with amount of adaptation data available are the most useful features. Future work will be directed toward investigation of better features for this task and predicting different tree structures for speaker types.

8. Acknowledgments

The authors thank Ramana Rao Gadde and Jing Zheng for feedback on implementation. This work was funded by DARPA under contract No. MDA972-02-C-0038. Distribution is unlimited. The views herein are those of the authors and do not reflect the views of the funding agencies.

9. References

[1] C. Leggetter and P. Woodland, “Maximum likelihood linear regression for speaker adaptation of HMMs,” *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

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

[3] C. Leggetter and P. Woodland, “Flexible speaker adaptation using maximum likelihood linear regression,” in *Proc. ARPA Spoken Language Technology Workshop*, 1995, pp. 104–109.

[4] C. Leggetter, “Improved acoustic modelling for HMMs using linear transformations,” Ph.D. dissertation, University of Cambridge, 1995.

[5] A. Stolcke and et al., “SRI system description,” in *EARS RT04 Workshop*, Palisades, 2004.

[6] L. Mangu, E. Brill, and A. Stolcke, “Finding consensus in speech recognition: Word error minimization and other applications of confusion networks,” *Computer Speech and Language*, vol. 14, pp. 373–400, 2000.

[7] L. Ubel and P. Woodland, “Improvements in linear transform based speaker adaptation,” in *Proc. of ICASSP*, 2001.