USING MAXIMUM LIKELIHOOD LINEAR REGRESSION FOR SEGMENT CLUSTERING AND SPEAKER IDENTIFICATION

Michiel Bacchiani

AT&T Labs-Research
Florham Park, NJ 07932, USA

ABSTRACT

Many adaptation scenarios rely on clustering of either the test or training data. Although consistency between the clustering and adaptation objective functions is desired, most previous approaches have not implemented such consistency. This paper shows that the statistics used in Maximum Likelihood Linear Regression (MLLR) adaptation are sufficient to cluster data with a consistent Maximum Likelihood (ML) criterion. In addition, as the algorithm uses the same statistics for both adaptation and clustering, it is computationally efficient. Clustering experiments contrasting the performance of this algorithm with the widely used text independent Gaussian mixture model approach show increased adaptation likelihoods and consistency of within-cluster speaker identity. In a speaker identification experiment the adaptation-based scoring showed improved classification performance compared to the mixture model-based scoring.

1. INTRODUCTION

Acoustic model adaptation for speech recognition has been an effective way to improve recognition accuracy. In particular, the model transformation technique Maximum Likelihood Linear Regression (MLLR) [1] is widely used. Although this technique requires fairly little adaptation data (several tens of seconds per transformation), this data requirement remains problematic if only very little adaptation data is available (rapid adaptation) or if a large number of transformations are desired. As a result, recent research efforts in acoustic model adaptation have focused on robust transformation estimation with very little available adaptation data per transform (several seconds). The developed techniques can be seen as split in two general approaches. The first approach is to incorporate additional data. In [2], the problem of rapid adaptation is addressed by using the Expectation-Maximization (EM) counts from the training data to smooth the EM counts from the adaptation data. In [3], relationships between multiple transformation classes are learned from the training set and used in testing to incorporate the adaptation data from neighboring classes for the estimation of the transform of a target class. For the Broadcast News task, clustering techniques with various distance metrics are used [4, 5, 6, 7] to find pools of adaptation data that will share transformations. Common among most of these approaches is that Text Independent Gaussian Mixture Models (TIGMMs) are used to characterize the individual adaptation data chunks (i.e. the smallest fragments of adaptation data). The second approach is to start with clustering the training data and computing cluster-dependent transformations or cluster-dependent models. Then, in the test phase, Maximum Likelihood (ML) estimates of linear combination weights are computed from the adaptation data. Using these weights, the contributions of the clusters are linearly combined providing the adapted system. As the number of free parameters of adaptation transformations or acoustic models are much larger than the cluster combination weights, the number of parameters that are to be estimated from the adaptation data is much smaller improving the robustness of the estimates. Several such approaches have recently been described varying in the part of the adapted system that is obtained by the linear combination from the clusters. In [8] an adaptation transformation is obtained by linear combination. In [9, 10, 11] the adapted model means are a linear combination of cluster model means. In [12] the likelihood of the adapted system is a linear combination of cluster model likelihoods.

The recurring element in most of the previously reported algorithms is that either the test or training data is clustered. However, the automatic clustering procedures generally do not optimize an objective function consistent with the ML criterion used in MLLR. The approaches represent the data chunks using a different model than the one used in recognition, hence the configuration that optimizes the clustering objective function does not necessarily optimize the MLLR adaptation likelihood. Furthermore, since a model is used in addition to the recognition model, all the described clustering approaches incur an additional computational cost. In [13], starting from an initial cluster configuration, MLLR likelihood was optimized. However, the described optimization process is not as efficient as the algorithm described in section 2. Here it is shown that the statistics required to compute the MLLR transformations are sufficient to cluster chunks using MLLR likelihood as the objective function. The computation of the MLLR statistics for the purpose of clustering does not pose an additional computational overhead as they will be used for adaptation in the final cluster configuration. Hence, the clustering algorithm presented here is both consistent, as it directly optimizes the MLLR likelihood, and efficient, as it uses the statistics already required for MLLR adaptation.

2. MLLR-BASED CLUSTERING

The clustering algorithm that optimizes the MLLR adaptation likelihood only requires the sufficient statistics that are collected to estimate the linear regression parameters in MLLR. First, in section 2.1, those sufficient statistics are defined. Then, in section 2.2, the two operations essential to any MLLR-based clustering approach are described. It describes how to reestimate a cluster representative from the statistics of its members and how to repartition the data across clusters, both using data likelihood as the objective function. Finally, in section 2.3, different clustering algorithms are discussed.

2.1. MLLR Statistics

As described in [1], MLLR transforms the n dimensional means $\mu^{(m)}$ of Gaussian components m of an unadapted system to obtain the means $\hat{\mu}^{(m)}$ of an adapted system by linear regression as

$$\hat{\mu}^{(m)} = W\mu^{(m)} + b = A\xi^{(m)}. \quad (1)$$

Denoting transposition as $()^T$, $\xi^{(m)}$ is the extended mean vector $\left[\mu^{(m)T}1\right]^T$ and A is the extended $n \times (n+1)$ transformation matrix $[Wb]^T$.

Given the adaptation data set $\mathcal{O} = \{o_1, \cdots, o_T\}$ from which the transform A is to be estimated, the ML estimate of the linear regression parameters is obtained by optimizing

$$\mathcal{L}(\mathcal{M}, \hat{\mathcal{M}}) = K - \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)\left[K^{(m)} + \log(|\ \Sigma^{(m)}\ |) + (o_t - \hat{\mu}^{(m)})^T\Sigma^{(m)-1}(o_t - \hat{\mu}^{(m)})\right], \quad (2)$$

where $|\ .\ |$ denotes a matrix determinant, $\Sigma^{(m)}$ is the covariance and $K^{(m)}$ is the normalization constant of Gaussian component m, K is a constant depending only on the transition probabilities and $\gamma_m(t)$ is the posterior probability of being in Gaussian m at time t given the original model set $\mathcal{M}$. Consider the $n \times (n+1)$ matrix

$$Z = \sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)\Sigma^{(m)-1}o_t\xi^{(m)T}, \quad (3)$$

and let z_i denote the i-th row of this matrix. In addition, assuming diagonal covariances $\Sigma^{(m)} = \mathrm{diag}(\left[\sigma_1^{(m)2}, \cdots, \sigma_n^{(m)2}\right]^T)$, consider the $(n+1) \times (n+1)$ matrices,

$$G_i = \sum_{m=1}^{M}\sum_{t=1}^{T}\frac{\gamma_m(t)}{\sigma_i^{(m)2}}\xi^{(m)}\xi^{(m)T} \quad \text{for } i = 1, \cdots, n. \quad (4)$$

The matrices Z and G_i for $i = 1, \cdots, n$ are a sufficient statistic for MLLR estimation as the i-th row of the ML estimate of the transformation A can be obtained as $a_i = z_i G_i^{-1}$.

2.2. Re-estimation and Repartitioning

For clustering, assume the MLLR sufficient statistics, Z and G_i for $i = 1, \cdots, n$, are available for each data chunk that is to be clustered. Each cluster c will be represented by a transformation $A^{(c)}$ which is comprised of row vectors $a_i^{(c)}$, $i = 1, \cdots, n$. In the re-estimation step, the cluster representative $A^{(c)}$ is derived by ML estimation from the data of the cluster members. This is equivalent to the ML estimation procedure of transformation A described in section 2.1, using adaptation data set $\mathcal{O}$ equal to the union of the data sets of each cluster member. The ML estimate is therefore easily obtained using the sums of chunk sufficient statistics of the cluster members.

In a repartition step, each chunk is to be assigned to the most likely cluster. In other words, each cluster can provide an adapted model set $\hat{\mathcal{M}}_c$ by application of its representative transformation $A^{(c)}$ to the original model set $\mathcal{M}$ and each chunk is to be assigned to the cluster whose $\hat{\mathcal{M}}_c$ results in the highest chunk data likelihood. Defining the chunk sample statistics

$$\tilde{\mu}^{(m)} = \frac{\sum_{t=1}^{T}\gamma_m(t)o_t}{\sum_{t=1}^{T}\gamma_m(t)} \quad (5)$$

$$\tilde{\Sigma}^{(m)} = \frac{\sum_{t=1}^{T}\gamma_m(t)(o_t - \tilde{\mu}^{(m)})^T(o_t - \tilde{\mu}^{(m)})}{\sum_{t=1}^{T}\gamma_m(t)}, \quad (6)$$

and making the assumption that the posterior probabilities given the adapted model are equal to those given the unadapted model, the likelihood of the chunk data as function of a cluster adapted model set $\hat{\mathcal{M}}_c$ can be written as

$$\mathcal{L}(\mathcal{M}, \hat{\mathcal{M}}_c) = K - \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)\left[K^{(m)} + \log(|\ \Sigma^{(m)}\ |) + \mathrm{tr}\{\Sigma^{(m)-1}\tilde{\Sigma}^{(m)}\} + (\tilde{\mu}^{(m)} - \hat{\mu}^{(m)})^T\Sigma^{(m)-1}(\tilde{\mu}^{(m)} - \hat{\mu}^{(m)})\right], \quad (7)$$

where $\mathrm{tr}\{\}$ denotes a matrix trace. Then, considering only those terms dependent on the transformation $A^{(c)}$ a modified likelihood is defined as

$$\mathcal{L}'(\mathcal{M}, \hat{\mathcal{M}}_c) = \sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)\left(A^{(c)}\xi^{(m)}\right)^T\Sigma^{(m)-1}\tilde{\mu}^{(m)} - \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)\left(A^{(c)}\xi^{(m)}\right)^T\Sigma^{(m)-1}A^{(c)}\xi^{(m)} \quad (8)$$

Finally, let $Y^{(c)}$ be the $n \times (n+1)$ matrix with the i-th row equal to $a_i^{(c)}G_i$, then the modified likelihood can be expressed in terms of the MLLR sufficient statistics as

$$\mathcal{L}'(\mathcal{M}, \hat{\mathcal{M}}_c) = \mathrm{tr}\{A^{(c)}Z^T\} - \frac{1}{2}\mathrm{tr}\{A^{(c)T}Y^{(c)}\}. \quad (9)$$

This shows that the statistics sufficient for MLLR transform estimation are also sufficient for performing the clustering steps with an ML criterion.

Note that this derivation easily extends to the case of R regression classes with $R > 1$. This divides the adaptation data in R disjoint observation sets so there will be R sets of sufficient statistics, each cluster will be represented by R transformations and the modified likelihood, defined in equation 9, will be a sum over the different regression classes.

2.3. Clustering Approaches

In pilot experiments, several clustering approaches were implemented using the ML re-estimation and repartitioning expressions derived in section 2.2. It was observed that both binary divisive and K-means approaches showed many local optima, possibly explained by the fact that initial clusters will be highly inconsistent (in terms of transformation characteristics), resulting in near identity transformations for every cluster. The agglomerative approach, used in all subsequent experiments, showed more desirable convergence properties. In this approach, each chunk was initially considered a cluster with a single chunk occupancy. Then, merges were evaluated using a likelihood ratio test for all possible cluster pairs and the merge resulting in the smallest likelihood loss was applied. This process was repeated until the empirically desired number of clusters was reached or until the likelihood loss exceeded an empirically set threshold. For each candidate merge of clusters l and

r, the representative of the merged cluster s was computed first and the likelihood loss computed as

$$\mathcal{L}'(\mathcal{M}, \hat{\mathcal{M}}_l) + \mathcal{L}'(\mathcal{M}, \hat{\mathcal{M}}_r) - \mathcal{L}'(\mathcal{M}, \hat{\mathcal{M}}_s), \qquad (10)$$

using the expression given in equation 9.

3. EXPERIMENTS

Experiments were conducted on a database of approximately 100 hours of voicemail recordings obtained from the mailboxes of 140 employees at AT&T. The messages were recorded from a variety of telephones including regular handsets and cellular phones. The speech signals were retrieved from the voicemail store which stored their digitized representation using an 8 kHz sampling rate and 8-bit μ-law samples.

The base HMM model set was trained on 39 dimensional observation vectors consisting of 12 Mel-warped cepstral coefficients, an energy component and their first and second order time derivatives. The models are Gender Dependent (GD) and reported experimental results are only on male data using the male system. Similar results were obtained from the female system. The training set used to train the HMM system consisted of approximately 60 hours, equally divided among genders. The HMM system was a state-clustered triphone system using decision tree clustering to determine the state tying. The male system had 4016 tied states, each modeled by a 12 component Gaussian mixture distribution. The mixture distributions were trained using Baum Welch (BW) re-estimation. Starting from single component mixture distributions for the tied states, the mixture distributions were obtained by incrementally increasing the number of mixture components, using the N component estimates to initialize training of the $N + 1$ component densities. Word boundaries were retained at fixed locations during BW training and were adjusted using Viterbi alignment with the 6, 8 and 10 component intermediate complexity systems. The recognition performance of this system on the 40 hours not used in training (using a 20k vocabulary and a trigram language model trained on 652k words from the training reference transcripts) was 33.3% WER.

To evaluate the performance of the clustering algorithm, 890 messages from 120 male speakers were clustered into 120 clusters. The available data per speaker ranged in duration from 91 seconds to 1635 seconds. The duration of the messages ranged from 1.95 seconds to 184 seconds. In the clustering experiments, all non-silence data from a single message constituted a data chunk (890 chunks in total). The MLLR-based clustering experiments used a single regression class with a full transform.

To compare the performance of the MLLR-based clustering algorithm, an approach similar to the one described in [6] was implemented. In this approach 64 component, diagonal covariance TIGMMs were estimated using EM re-estimation for all chunks. The features used for these experiments were 12 dimensional linear predictive coding-based cepstral coefficient and their derivatives. The chunks were then clustered agglomeratively using a likelihood-based distance metric and the furthest-neighbor algorithm. The 64 component mixture densities were trained similarly to the HMM system in terms of the incremental increase of the number of mixture components. To avoid data sparsity problems on very short messages, the covariances of mixture components were tied during the incremental mixture component increase if fewer than 100 training frames were available for a mixture component that was to be split.

Consider the MLLR likelihood range from a lower bound (all chunks share a single MLLR transformation) to an upper bound (each chunk has its own MLLR transformation). The 120 cluster configuration found by the MLLR-based clustering algorithm reached 45.9% of this range in comparison to 38.9% using the TIGMM approach. Grouping the messages in 120 clusters using the supervisory information about speaker identity resulted in 43.2% of this range.

Using the evaluation criterion reported in [5], cluster purity can be computed as the percentage of messages in a cluster that are from the most frequently represented speaker (dominating speaker) in that cluster. Figure 1 shows this metric for the 120 cluster configuration found by the TIGMM approach, figure 2 for the MLLR-based approach.

Another evaluation of the clustering performance was to measure how many cluster merges during the agglomeration involved two clusters with no speakers in common. After the 770 cluster merges to form the 120 cluster configuration from the 890 messages, 102 such errors were counted for the MLLR-based approach compared to 260 for the TIGMM approach. Figure 3 shows the accumulative number of merge errors for the MLLR-based approach next to the likelihood losses (equation 10) of those merges. It shows there is a correlation between the increase of the likelihood losses of subsequent merges and the number of merge errors indicating that the likelihood loss increase can be used as a complexity control parameter. The merge costs in the TIGMM approach did not show such a correlation to merge errors.

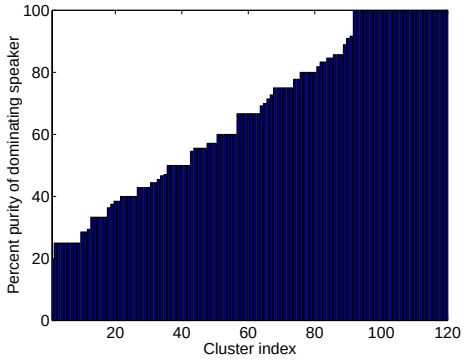

Figure 1: Cluster purity of the 120 cluster configuration found by the TIGMM-based clustering approach.

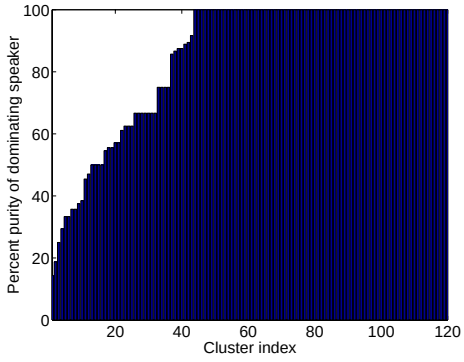

Figure 2: Cluster purity of the 120 cluster configuration found by the MLLR-based clustering approach.

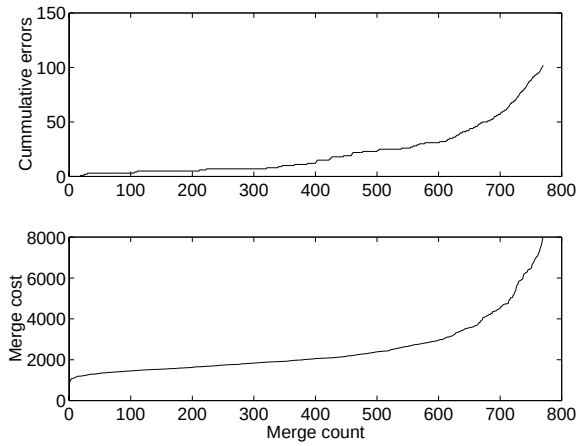

Figure 3: Cumulative merge errors and merge costs of the MLLR-based clustering approach.

Another application of the MLLR-based clustering formulation is that it can be used to classify test messages. Given a set of transformations, one for each speaker in a training set, a speaker identification experiment can be conducted by finding the most likely training speaker for the MLLR statistics of a test message. To evaluate the performance of such a classification scheme, 268 messages were selected from the 40 hour test set. The speaker identities of these messages were overlapping with the 120 speakers used in the clustering experiments. First, MLLR transformations were computed for the 120 speakers using their training messages. Then, MLLR statistics were computed for the 268 test messages based on the recognizer transcripts (33.3%WER). The correct classification rate of this scheme was 81% compared to 65% using a similar scheme with TIGMMs.

All clustering and speaker identification experiments were repeated with 42 regression classes (one per center phone) with diagonal plus shift transformations. For this set of experiments, the performance was slightly worse than the full transform setup but still better than the TIGMM approach for all evaluation metrics.

4. CONCLUSIONS

Many of the adaptation approaches intended for use with little adaptation data require either the training or test data to be clustered. As this clustering is intended to be used for adaptation, a consistent approach is to use adaptation likelihood as the clustering objective function. This paper shows how the MLLR statistics are sufficient for a very efficient implementation of such a clustering scheme. Clustering results on a voicemail database show that the algorithm finds cluster configurations that result in both higher adaptation likelihood as well as more consistent within-cluster speaker identity than the more widely used approach based on TIGMMs. The experiments showed that the likelihood of the cluster configuration found by the MLLR-based clustering algorithm exceeded the likelihood of the configuration defined by the supervisory speaker identity information. This can be explained by the fact that MLLR adaptation compensates for both channel and speaker characteristics and both vary in the voicemail data used in these experiments. This illustrates the usefulness of the MLLR-based clustering approach even in tasks where supervisory information is available.

Evaluation of the MLLR-based approach in a speaker identification type of application showed improved performance of the adaptation-based classification compared to use of TIGMMs. The performance of the algorithms on an open task (adding the requirement of being able to reject messages of unknown speakers) was not tested.

5. REFERENCES

1. C. J. Legetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language,* pp. 171-185, 1995.
2. W. Byrne and A. Gunawardana, "Discounted Likelihood Linear Regression for Rapid Adaptation," In *Proceedings European Conference on Speech Communication and Technology,* Vol. 1, pp. 203-206, 1999.
3. S.-J. Doh and R. Stern, "Inter-Class MLLR for Speaker Adaptation," In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* , Vol. 3, pp.1543-1546, 2000.
4. P. C. Woodland M. J. F. Gales, D. Pye and S. J. Young, "The Development of the 1996 Broadcast News Transcription System," *Proc. DARPA Speech Recognition Workshop,* pp. 73-78, 1997.
5. S. Chen and P. S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," *Proc. DARPA Speech Recognition Workshop,* pp. 127-132, 1998.
6. J.-L. Gauvain *et al.*, "The LIMSI 1998 Hub-4E Transcription System," *Proc. DARPA Speech Recognition Workshop,* pp. 99-104, 1999.
7. P. Beyerlein *et al.*, "Automatic Transcription of English Broadcast News," *Proc. DARPA Speech Recognition Workshop,* pp. 85-90, 1998.
8. M. Gales, "Transformation Smoothing for Speaker and Environmental Adaptation," In *Proceedings European Conference on Speech Communication and Technology,* pp. 2067-2070, 1997.
9. M. Gales, "Cluster Adaptive Training for Speech Recognition," In *Proceedings of the International Conference on Spoken Language Processing,* Vol. 5, pp. 1783-1786, 1998.
10. M. Padmanabhan, *et al.*, "Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition Systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 6, No. 1, pp. 71-77, 1998.
11. Y. Gao, *et al.*, "Speaker Adaptation based on Pre-clustering Training Speakers," In *Proceedings European Conference on Speech Communication and Technology,* pp. 2091-2094, 1997.
12. C. Boulis and V. Digalakis, "Fast Speaker Adaptation of Large Vocabulary Continuous Density HMM Speech Recognizer using a Basis Transform Approach," In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* Vol. 2, pp. 989-992, 2000.
13. S. E. Johnson and P. C. Woodland, "Speaker Clustering Using Direct Maximisation of the MLLR-Adapted Likelihood," In *Proceedings of the International Conference on Spoken Language Processing,* Vol. 5, pp. 1775-1778, 1998.