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 estimates 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 utilize 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 $\tilde{\mu}^{(m)}$ of an adapted system by linear regression as

$$\tilde{\mu}^{(m)} = W\mu^{(m)} + b = A_{l}g^{(m)}.$$  \hspace{1cm} (1)

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

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

$$L(M, \hat{A}) = K - \frac{1}{2} \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_{m}(t) \left[K^{(m)} + \log|\Sigma^{(m)}| + (o_t - \tilde{\mu}^{(m)})^T \Sigma^{(m)^{-1}} (o_t - \tilde{\mu}^{(m)}) \right],$$ \hspace{1cm} (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 $\gamma_t$ is the posterior probability of being in Gaussian $m$ at time $t$ given the original model set $M$. Consider the $n \times (n + 1)$ matrix

$$Z = \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_{m}(t) \Sigma^{(m)^{-1}} o_t G^{(m)} t,$$ \hspace{1cm} (3)

and let $z_{i}$ denote the $i$-th row of this matrix. In addition, assuming diagonal covariances $\Sigma^{(m)} = \text{diag}(\sigma_{1}^{(m)}, \ldots, \sigma_{n}^{(m)})^T$, consider the $(n + 1) \times (n + 1)$ matrices

$$G_{i} = \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_{m}(t) \sigma_{i}^{(m)} o_t G^{(m)} t$$ for $i = 1, \ldots, n$. \hspace{1cm} (4)

The matrices $Z$ and $G_{i}$ are a sufficient statistic for MLLR estimation as the $i$-th row of the ML estimate of the transformation $A$ can be obtained as $\sigma_{i} = z_{i}G_{i}^{-1}$.

2.2. Re-estimation and Repartitioning

For clustering, assume the MLLR sufficient statistics, $Z$ and $G_{i}$ for $i = 1, \ldots, n$, 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 $a_{i}^{(c)}$, $i = 1, \ldots, n$. 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 $O$ 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 re-partition step, each chunk is to be assigned to the most likely cluster. In other words, each cluster can provide an adapted model set $M_{c}$ by application of its representative transformation $A^{(c)}$ to the original model set $M$ and each chunk is to be assigned to the cluster whose $M_{c}$ results in the highest chunk data likelihood. Defining the chunk sample statistics as

$$\mu^{(m)} = \frac{\sum_{t=1}^{T} \gamma_{m}(t) o_t}{\sum_{t=1}^{T} \gamma_{m}(t)}$$ \hspace{1cm} (5)

$$\Sigma^{(m)} = \frac{\sum_{t=1}^{T} \gamma_{m}(t) (o_t - \tilde{\mu}^{(m)}) (o_t - \tilde{\mu}^{(m)})^T}{\sum_{t=1}^{T} \gamma_{m}(t)}$$ \hspace{1cm} (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 a function of a clustered adapted model set $M_{c}$ can be written as

$$L(M, \hat{M}_{c}) = K - \frac{1}{2} \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_{m}(t) \left[K^{(m)} + \log|\Sigma^{(m)}| + (\tilde{\mu}^{(m)} - \mu^{(m)})^T \Sigma^{(m)^{-1}} (\tilde{\mu}^{(m)} - \mu^{(m)}) \right],$$ \hspace{1cm} (7)

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

$$L'(M, \hat{M}_{c}) = \frac{1}{2} \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_{m}(t) \left[A^{(c)} \mu^{(m)} \right]^T \Sigma^{(m)^{-1}} A^{(c)} \mu^{(m)}$$ \hspace{1cm} (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

$$L'(M, \hat{M}_{c}) = \text{tr} \{ A^{(c)} Z^T \} - \frac{1}{2} \text{tr} \{ A^{(c)} Y^{(c)} A^{(c)^T} \}.$$ \hspace{1cm} (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 the 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, mergers 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 $I$ and...
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 clusters merged during the agglomeration involved two clusters with no speakers in common. After the 770 cluster mergers to form the 120 cluster configuration from the 850 messages, 102 such errors were counted for the MLLR-based approach compared to 260 for the TIGMM approach. Figure 3 shows the cumulative 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.

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 coefficients 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.
Another application of the MLLE-based clustering formulation is that it can be used to classify test messages. 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 MLLE statistics of a test message. To evaluate the performance of such a classification scheme, 208 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, MLLE transformations were computed for each speaker using training messages. Then, MLLE statistics were computed for the 208 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 transformation 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 MLLE 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 MLLE-based clustering algorithm exceeded the likelihood of the configuration defined by the supervisory speaker identity information. This can be explained by the fact that MLLE 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 MLLE-based clustering approach even in tasks where supervisory information is not available.

Evaluation of the MLLE-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 algorithm on an open task (adding the requirement of being able to reject messages of unknown speakers) was not tested.

5. REFERENCES