OPTIMAL CLUSTERING OF MULTIVARIATE NORMAL DISTRIBUTIONS USING DIVERGENCE AND ITS APPLICATION TO HMM ADAPTATION

Tor André Myrvoll[*] *and Frank K. Soong*

Bell Labs Research, Lucent Technologies
600 Mountain Ave, Murray Hill, NJ 07974

ABSTRACT

We present an optimal clustering algorithm for grouping multivariate normal distributions into clusters using the divergence, a symmetric, information-theoretic distortion measure based on the Kullback-Liebler distance. Optimal solutions for normal distributions are shown to be obtained by solving a set of Riccati matrix equations and the optimal centroids are found by alternating the mean and covariance matrix intermediate solutions. The clustering performance of the new algorithm compared favorably against the conventional, non-optimal clustering solutions of sample mean and sample covariance in its overall rate-distortion and even distributions of samples across clusters. The resultant clusters were further tested on unsupervised adaptation of HMM parameters in a framework of Structured Maximum A Posterior Linear Regression (SMAPLR). The Wall Street Journal database was used for the adaptation experiment. The recognition performance with respect to the word error rate, was significantly improved from a non-optimal centroid (sample mean and covariance) of 32.6% to 27.6% and 27.5% for the diagonal and full covariance matrix cases, respectively.

1. INTRODUCTION

The need to cluster multivariate normal distributions is often encountered when working with normal mixture density based Hidden Markov Models (HMMs) in automatic speech recognition (ASR). An indirect way to cluster distributions is frequently used when tied HMMs are constructed. Usually the clustering is performed in an agglomerative or divisive hierarchical method together with a likelihood based decision rule [1, 2]. A different approach that doesn't involve raw observations is to split the parameters of an HMM (i.e., Gaussian Kernels) directly into clusters or a hierarchical tree where each node forms its own cluster and the resultant clusters are then used in model adaptation algorithms. These algorithms include, e.g., Maximum Likelihood Linear Regression (MLLR)[3], Structural MAP adaptation (SMAP)[4] and Cluster Adaptive Training (CAT)[5]. All these approaches perform well when a logical partitioning of the HMM parameters is carried out, resulting in a richer and more structural mapping in the HMM parameter space.

Most model adaptation algorithms focus only on the mixture component mean vectors of the underlying HMMs. This is due to the fact that the state transition probabilities and mixture weights have little to no effect on the overall performance, and that the covariance matrices of the mixture components are numerically unstable to adapt and a robust estimate is difficult to obtain when the adaptation data is scarce. This simplifications enables us to focus on the clustering of the mixture components only. Ideally, we should obtain clusters of mixture components whose mean vectors are structured in such a way that any observed perturbations in a small subset should lead us to infer the adaptation direction and magnitude of the unobserved mean vectors using a simple model, e.g., an affine transformation as in the MLLR case. A simple but useful conjecture is that "similar" mixture components (i.e., Gaussian kernels) should always be grouped into the same cluster, where the term "similar" is open for an intuitively appealing and mathematically tractable definition.

One similarity measure between mixture components that has been strongly advocated and commonly used e.g., [4, 6] is the divergence measure [7]. This measure is defined for measuring the "distance" or "distortion" between two given probability density functions, f and g, as

$$d(f,g) = \int f \log \frac{f}{g} + \int g \log \frac{g}{f} \tag{1}$$

Note that the divergence does not fulfill the triangle inequality, and so it is not a distance in the sense of being a metric as defined in topology [8]. It does however fit the notion of a distortion measure as it is defined in [9].

If f and g are multivariate normal distributions, as is commonly used for modeling a continuous HMM, equation (1) has the a closed form solution as follows,

$$\begin{aligned} & d(f,g) \\ & = \frac{1}{2}\text{trace}\Big\{(\boldsymbol{\Sigma}_f^{-1} + \boldsymbol{\Sigma}_g^{-1})(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)^T \\ & \qquad + \boldsymbol{\Sigma}_f\boldsymbol{\Sigma}_g^{-1} + \boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_f^{-1} - 2\mathbf{I}\Big\}, \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the corresponding mean vectors and covariance matrices, respectively. In [6] a comparative study was performed on the use of the divergence, Euclidean distance and Bhattacharyya distance as distance measures in the construction of a hierarchy of clusters of HMM mixture components. The result of the study shows that the divergence measure gave the best adaptation results. In this work we will concentrate on how to find an optimal clustering centroid of multivariate normal distributions using the divergence as the sole similarity measure, as well as investigating its application to HMM adaptation.

[*] This work was conducted while pursuing the degree Doktor Ingeniør from the Department of Telecommuncations, Norwegian University of Science and Technology.

0-7803-7663-3/03/$17.00 ©2003 IEEE

2. THE EXPECTATION CENTROID

Although the divergence measure has been proposed to measure the similarity between multivariate normal densities used in a context of k-means clustering e.g., [10], no solution has been published, to the authors' best knowledge, of the the true centroid of a cluster of multivariate Gaussian densities. The centroid we are interested in is a multivariate normal density that minimizes the total distortions in a cluster. Formally, a centroid c is defined as,

$$c = \underset{c'}{\operatorname{argmin}} \sum_{n=1}^{N} d(x_n, c'), \tag{3}$$

where N is the number of cluster members, and x_n is the nth cluster member.

In [4] a centroid which we shall refer as the *expectation centroid* is defined as a density whose mean and covariance are the sample mean and the sample covariance,

$$\begin{aligned} \boldsymbol{\mu}_c &= \frac{1}{N}\sum_{n=1}^{N} E[x_n] \\ &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\mu}_n, \end{aligned} \tag{4}$$

$$\begin{aligned} \boldsymbol{\Sigma}_c &= \frac{1}{N}\sum_{n=1}^{N} E[(x_n - \boldsymbol{\mu}_c)(x_n - \boldsymbol{\mu}_c)^T] \\ &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\Sigma}_n + \boldsymbol{\mu}_n\boldsymbol{\mu}_n^T - \boldsymbol{\mu}_c\boldsymbol{\mu}_c^T, \end{aligned} \tag{5}$$

where x_n refers to a random variable distributed according to the nth cluster member. Later experiments show that this centroid has good convergence properties and reasonable clustering performance. However, it is not the optimal centroid which minimizes the total divergence of a cluster. Next we will present our solution of the optimal centroid.

3. THE OPTIMAL CENTROID

The centroid we need to find is a multivariate normal distribution which solves the following minimization problem,

$$\{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\} = \underset{\boldsymbol{\mu}'_c, \boldsymbol{\Sigma}'_c}{\operatorname{argmin}} \sum_{n=1}^{N} d(f_{c'}, f_n). \tag{6}$$

Here $d(\cdot,\cdot)$ is the Kullback-Liebler divergence measure, $f_{c'}$ is a multivariate normal distribution having mean vector $\boldsymbol{\mu}'_c$ and covariance matrix $\boldsymbol{\Sigma}'_c$, and f_n refers to the nth member in the cluster.

The mean vector of the centroid, $\boldsymbol{\mu}_c$, can be found simply by setting the gradient of the object function in equation (6) with respect to $\boldsymbol{\mu}'_c$ to zero. This yields the following solution

$$\boldsymbol{\mu}_c = \left[\sum_{n=1}^{N}(\boldsymbol{\Sigma}_n^{-1} + \boldsymbol{\Sigma}_c^{-1})\right]^{-1} \left[\sum_{n=1}^{N}(\boldsymbol{\Sigma}_n^{-1} + \boldsymbol{\Sigma}_c^{-1})\boldsymbol{\mu}_n\right]. \tag{7}$$

The procedure to find the covariance matrix of the centroid is a bit more involved. Let us first state the following useful lemma from functional analysis [11]:

Lemma 3.1 *If* $\mathbf{A}$ *is an element in a unital Banach algebra* $\mathfrak{A}$ *with* $\|\mathbf{A}\| < 1$, *then* $\mathbf{I} - \mathbf{A} \in \mathbf{GL}(\mathfrak{A})$, *the set of all elements having an inverse, and*

$$(\mathbf{I} - \mathbf{A})^{-1} = \sum_{i=0}^{\infty} \mathbf{A}^i. \tag{8}$$

For our purpose it is sufficient to note that the vector space of all finite dimensional $n \times n$ matrices is an unital Banach algebra under matrix multiplication and the operator norm.

Let $\boldsymbol{\Sigma}_c + \varepsilon\mathbf{R}\mathbf{R}^T$ be a perturbation of the optimal covariance matrix, where $\mathbf{R}$ is any real matrix. The inverse is then given by lemma 3.1, provided that ε is chosen sufficiently small:

$$\begin{aligned} (\boldsymbol{\Sigma}_c + \varepsilon\mathbf{R}\mathbf{R}^T)^{-1} &= \left(\boldsymbol{\Sigma}_c\left[\mathbf{I} - (-\varepsilon\boldsymbol{\Sigma}_c^{-1}\mathbf{R}\mathbf{R}^T)\right]\right)^{-1} \\ &= \left(\mathbf{I} - (-\varepsilon\boldsymbol{\Sigma}_c^{-1}\mathbf{R}\mathbf{R}^T)\right)^{-1}\boldsymbol{\Sigma}_c^{-1} \\ &= \left\{\sum_{n=0}^{\infty}(-\varepsilon\boldsymbol{\Sigma}_c^{-1}\mathbf{R}\mathbf{R}^T)^n\right\}\boldsymbol{\Sigma}_c^{-1}. \end{aligned} \tag{9}$$

By replacing the covariance matrix $\boldsymbol{\Sigma}$ in equation (6) with a perturbed one, we can now compute the Gâteaux variation (see [12]) and set it to zero. In other words, we need to solve the following equations:

$$\left.\frac{\partial}{\partial\varepsilon}\right|_{\varepsilon=0} \sum_{n=1}^{N} d(f_c, f_n) = 0, \tag{10}$$

for $\boldsymbol{\Sigma}_c$, to obtain the optimal centroid.

It turns out that solving equation (10) is equivalent to solving the following Riccati matrix equation,

$$\mathbf{A} + \mathbf{B}\mathbf{X} + \mathbf{X}\mathbf{B}^* - \mathbf{X}\mathbf{C}\mathbf{X} = \mathbf{0}, \tag{11}$$

where

$$\mathbf{A} = \sum_{n=1}^{N}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_c)(\boldsymbol{\mu}_n - \boldsymbol{\mu}_c)^T + \boldsymbol{\Sigma}_n, \tag{12}$$

$$\mathbf{B} = \mathbf{0}, \tag{13}$$

$$\mathbf{C} = \sum_{n=1}^{M} \boldsymbol{\Sigma}_n^{-1}, \tag{14}$$

$$\mathbf{X} = \boldsymbol{\Sigma}_c. \tag{15}$$

Equation (11) can be written in a block matrix form as,

$$\begin{bmatrix} \mathbf{I} & -\mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{B} & \mathbf{A} \\ \mathbf{C} & -\mathbf{B}^* \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{I} \end{bmatrix} = \mathbf{0}. \tag{16}$$

Let us denote the square block matrix as $\mathbf{M}$. We now apply the following theorem to find the optimal covariance matrix [13]:

Theorem 3.1 *Assume that* $\mathbf{A}$, $\mathbf{C}$ *are positive semidefinite Hermitian, and let* $\mathbf{v}_1, \ldots, \mathbf{v}_d$ *be eigenvectors of* $\mathbf{M}$ *corresponding to the eigenvalues* $\lambda_1, \ldots, \lambda_d$. *Further, if* $\mathbf{v}$ *is an eigenvector of* $\mathbf{M}$, *we will write*

$$\mathbf{v} = \begin{bmatrix} \mathbf{u} \\ \mathbf{w} \end{bmatrix},$$

where $\mathbf{u}$ *and* $\mathbf{w}$ *are the upper and lower halves of* $\mathbf{v}$ *respectively. Then, if* $\lambda_1, \ldots, \lambda_d$ *has positive real parts and* $[\mathbf{w}_1, \ldots, \mathbf{w}_d]$ *is nonsingular, we have that*

$$\mathbf{X} = [\mathbf{u}_1, \ldots, \mathbf{u}_d][\mathbf{w}_1, \ldots, \mathbf{w}_d]^{-1}, \tag{17}$$

is a solution to equation (16) *as well as positive semidefinite.*

For a proof of the existence of at least d positive eigenvalues, $\lambda_1, \ldots, \lambda_d$, as well as the non-singularity of $[\mathbf{w}_1, \ldots, \mathbf{w}_d]$, we refer to [14]. In the same reference it is also shown that the objective function is convex in both $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$, hence guaranteeing a global minimum.

In the special case where all the distributions have a diagonal covariance matrix we can constrain the covariance of the centroid to be diagonal, yielding the following simple expressions for the ith elements of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ respectively,

$$\boldsymbol{\mu}_c(i) = \frac{\sum_{n=1}^{N} \left(\boldsymbol{\Sigma}_c^{-1}(i) + \boldsymbol{\Sigma}_n^{-1}(i)\right) \boldsymbol{\mu}_n(i)}{\sum_{k=1}^{N} \left(\boldsymbol{\Sigma}_c^{-1}(i) + \boldsymbol{\Sigma}_k^{-1}(i)\right)}, \qquad (18)$$

$$\boldsymbol{\Sigma}_c(i) = \sqrt{\frac{\sum_{n=1}^{N} \boldsymbol{\Sigma}_n(i) + (\boldsymbol{\mu}_c(i) - \boldsymbol{\mu}_n(i))^2}{\sum_{k=1}^{N} \boldsymbol{\Sigma}_k^{-1}(i)}} \qquad (19)$$

Finally it should be noted that the solutions for $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are dependent on each other. No joint solution is readily obtainable, and we compute the $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ alternatively, starting from the expectation centroid.

4. EXPERIMENTAL RESULTS

In this section we present two sets of experimental results. The first experiment is clustering performance where we compare the true centroids, in both forms of full and diagonal covariance matrix, with the expectation centroid defined in equations (4) and (5). In the second experiment we use the true centroid and the expectation centroid to build regression trees and compare the their performance in SMAPLR adaptation [15].

The multivariate normal distributions used in these experiments are from HMMs that is trained using the speech data from 84 speakers in the Wall Street Journal Corpus. The model contains 37,786 mixture components – all in the form of multivariate normal distribution with diagonal covariance matrix. The model is constructed using feature vectors of 12 mel-frequency cepstral coefficients together with the normalized log energy plus the Δ- and Δ^2-coefficients, 39 features altogether.

4.1. Clustering Performance

We will now use the clustering algorithm with the divergence measure to group HMM mixture components. The set of mixture components was to be grouped into 10 clusters. To account for the fact that the initial conditions change the final clustering performance, we repeated the experiment five times using different initializations while the same initializations were used for all the three centroid finding procedures.

The results can be seen in figure 1. As expected, the two optimal centroids clearly outperforms the expectation centroid in terms of the total divergence, with the full covariance centroid being slightly better than the diagonally constrained centroid. Also the expectation based centroid exhibits a non-monotonic decrease of distortion, which is to be expected as the computation of a new set of centroids is not guaranteed to lower the total distortions, as opposed to the true centroids.

The true centroids also yield a more evenly distributed clusters of the mixture components, which can be seen from the example in figure 2. A more formal evaluation of the flatness is given in table 1, where the Shannon entropy is used to measure the flatness of the cluster distribution for the five different experiments.

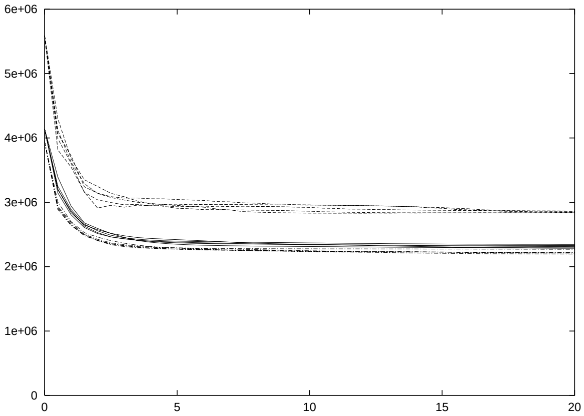

Fig. 1. Convergence performance of three different centroids. The figure shows the total divergence on the y-axis versus the iteration number on x-axis. The dashed lines at the top corresponds to the expectation centroids, the drawn lines, in the middle group, to the optimal diagonally constrained centroids, and the dash-dotted lines, which give the lowest overall distortion, to the optimal centroids with full covariances.

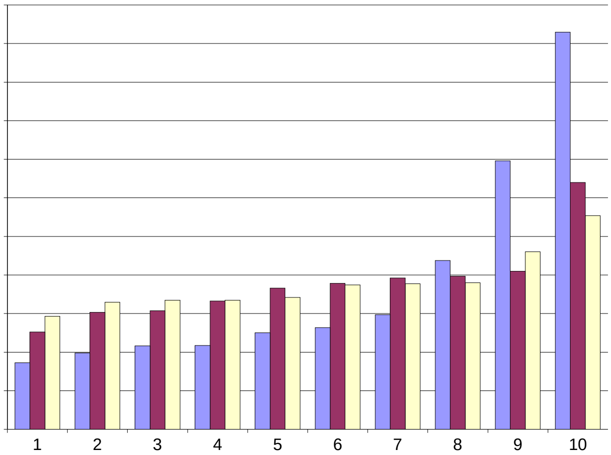

Fig. 2. Distribution of mixture components across clusters. Left bar corresponds to expectation centroid, middle to diagonally constrained centroid and right to full covariance centroid.

Expectation	Diagonal	Full
3.0363	3.2766	3.2972
2.9978	3.3028	3.2964
3.0114	3.2782	3.2332
3.0136	3.2599	3.2791
3.0002	3.2724	3.2768
3.0119	3.2780	3.2765

Table 1. The flatness of the mixture components distribution across clusters, as measured by the entropy of the normalized bin counts. Five different experiments are presented as well as the average over these experiments.

4.2. SMAPLR Adaptation

The clustering performance results are definitely encouraging, but it is yet to be confirmed that the optimal centroids can improve the recognition performance, say, in a model adaptation task. The three centroids from the previous clustering experiments were used to build three sets of hierarchical clusterings of HMM mixture components in a form of regression tree. These trees were then used with the SMAPLR approach on the Wall Street Journal Spoke III task, consisting of ten non-native speakers of American English. The original, speaker independent model is the same as the one that we used for our clustering experiments in the previous section. The un-adapted model gives us the baseline performance of a word error rate (WER) of 29.2%. Only one adaptation utterance was used to adapt the model in this adaptation experiment. Earlier experiment has shown that we need to be very conservative in adapting the HMM parameters when data is scarce, as the performance can degrade significantly when a poor mapping is obtained. Here we try to be more more ambitious such that we do not end up using only one global transformation, something that would render the experiment meaningless. Just like the clustering experiments we repeated the tree building procedure five times using different initializations. The results are presented in table 2.

Expectation	Diagonal	Full
31.6	27.1	27.9
33.4	27.5	27.2
34.7	27.5	27.9
31.7	27.8	26.7
31.7	27.9	27.7
32.6	27.6	27.5

Table 2. The word error rate (WER) on the Wall Street Journal Spoke III task for three different centroids based hierarchical clustering. Results for five separate experiments are presented, as well as their averages. The baseline WER of the speaker independent model is 29.2%.

The results clearly indicates that the true centroids based hierarchical tree clusterings are better suited for adaptation than the trees built using expectation centroids. While it is still difficult to pinpoint the reason why the adaptation performance of of the expectation centroid based trees is inferior to that of the trees built using the optimal centroids, we do feel that a more consistent minimization of the total distortion and the true optimal centroids thus derived should provide more compact and homogeneous clusters in the information theoretic sense, hence leading to better regression trees for adaptation. However, whether the improvement should be attributed to a more uniform cluster size, or to a more homogeneous clustering is yet to be confirmed via further study.

5. CONCLUSION

In this work we present a novel clustering algorithm for finding the optimal centroids of multivariate normal distributions using the Kullback-Liebler divergence measure. It is shown that the clusterings obtained using the optimal centroid yield significantly lower overall distortion than the centroid based on the sample mean and the sample covariance found in the literature. We also observe that the multivariate normal distributions are more evenly distributed across resultant clusters. The optimal centroids were further used to construct hierachical regression trees and tested for adapting HMM parameters. The adaptation result shows a clear improvement in WER when compared to trees built with the non-optimal centroids.

6. REFERENCES

[1] A. Kannan, M. Ostendorf, and J. R. Rohlicek, "Maximum likelihood clustering of gaussians for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, Jul. 1994.

[2] J. J. Odell, P. C. Woodland, and S. J. Young, "Tree-based state clustering for large vocabulary speech recognition," in *Int. Symp. on Speech, Image Proc. and Neural Networks*, Hong Kong, Apr. 1994, pp. 690–693.

[3] C. J. Leggetter and P. C. Woodland, "Speaker adaptation of continuous density HMMs using multivariate linear regression," in *Proc. ICSLP*, Yokahama, Japan, Sep. 1994.

[4] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Transactions on Speech and Audio Processing*, 2000.

[5] Mark. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, Jul. 2000.

[6] Jen-Tzung Chien, "Online hierarchical transformation of hidden Markov models for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 656–667, Nov. 1999.

[7] Solomon Kullback, *Information Theory and Statistics*, Dover Publications, 1997.

[8] H. L. Royden, *Real Analysis*, Macmillan Publishing Company, 3 edition, 1988.

[9] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, Inc, 2 edition, 2001.

[11] Gert K. Pedersen, *Analysis Now*, Spinger-Verlag New Yourk, Inc, 2 edition, 1995.

[12] John L. Troutman, *Variational Calculus and Optimal Control*, Springer-Verlag New York, Inc., 1996.

[13] James E. Potter, "Matrix quadratic solutions," *SIAM J. Appl. Math*, vol. 14, no. 3, pp. 496–501, 1966.

[14] Tor André Myrvoll, *Adaptation of Hidden Markov Models using Maximum a Posteriori Linear Regression with Hierarchical Priors*, Ph.D. thesis, Norwegian University of Science and Technology, 2002.

[15] Olivier Siohan, Tor André Myrvoll, and Chin-Hui Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech and Language*, vol. 16, no. 1, pp. 5–25, Jan. 2002.