A NEW DISTANCE MEASURE FOR PROBABILITY DISTRIBUTION FUNCTION OF MIXTURE TYPE

*Zhu Liu**

Electrical Engineering Department
Polytechnic University
Brooklyn, NY 11201
zhul@vision.poly.edu

Qian Huang

AT&T Labs - Research
100 Schulz Drive
Red Bank, NJ 07701
huang@research.att.com

ABSTRACT

Evaluating the similarity between two Probability Distribution Functions (PDF) is very important in various research problems. This paper proposes a new metric that computes the distance between two PDF's of mixture type directly from their parameters. It is posed as a linear programming problem and its theoretical properties and performance are analyzed, experimented, and compared with existing measures. In addition, as a proof of concept, we applied the new metric to the problem of audio retrieval where involved PDF's are GMM's (Gaussian Mixture Model) with 4 mixtures. Experimental results on both synthetic and real data show that this new distance measure is quite promising.

1. INTRODUCTION

Under various situations, it is necessary to measure the difference between two PDF's. For example, in text independent speaker recognition using Gaussian mixture models (GMM)[1], the classification of a given piece of speech can be done by comparing its GMM model with a set of given GMM models. Another scenario is to detect the difference among observation probabilities, often characterized by again GMM, of each state of a continuous Hidden Markov Model (HMM) [3] so that similar states can be merged to simplify the overall model in speech recognition tasks. Although needed, there is no simple way to measure the distance between two mixture PDF's. In this paper, we are particularly interested in developing such a measure and propose a general framework to compute the difference between two mixture PDF's, directly from their component model parameters.

There are three approaches to measure the difference between two PDF's, which may or may not satisfy the three well known properties of distance measure. Let $G(x)$, $F(x)$, and $H(x)$ be three PDF's. Denote $D(G,F)$ as the distance between $G(x)$ and $F(x)$, then the three properties are

$$D(G,F) \geq 0, \quad \text{and} \quad D(G,F) = 0 \quad \text{iff.} G = F \qquad (1)$$

$$D(G,F) = D(F,G) \qquad (2)$$

$$D(G,H) + D(H,F) \geq D(G,F) \qquad (3)$$

The first approach defines the distance in $\mathbf{L}^r$ space by

$$D_{L^r}(G,F) = \left(\int_{x \in \mathbf{X}} |G(x) - F(x)|^r dx \right)^{1/r},$$

where commonly used values of r may be 1 or 2. Although satisfying all three distance properties, D_{L^r} is usually computed by numerical method and the complexity can easily go beyond control with the increased dimension.

The second approach is the relative entropy or Kullback Leibler distance (KLD). It is defined as [4],

$$D_{KL}(G,F) = \int_{x \in \mathbf{X}} G(x) \log \frac{G(x)}{F(x)} dx$$

It is obvious that KLD satisfies only the first property. By extending the original KLD to $D_{KL}(G,F) + D_{KL}(F,G)$, the second condition can be met. Although the third property still does not hold, the extended KLD is popular in many applications due to the lack of other alternatives. There are practical ways to approximate $D_{KL}(G,F)$. For example, data sequences T_G and T_F can be generated from models G and F and then the average log-likelihood ratio of the sequences with respect to $G(x)$ and $F(x)$ can be used to approximate the extended KLD. That is,

$$D_{Seq}(G,F) = \frac{1}{N}(|\log \frac{p(T_G|G)}{p(T_G|F)}| + |\log \frac{p(T_F|F)}{p(T_F|G)}|),$$

where N is the length of the data sequences T_G and T_F. The performance of D_{Seq} is a function of both the value of N as well as the data generation procedure. The bigger the N is, the more reliable the approximation is. At the same time, it makes the estimation more expensive.

Since the mixture PDF's are characterized by their component model parameters, the most desirable solution, in the third class of approach, is to compute the distance directly from their respective parameters. Ideally, it is hoped that such a method can achieve at least comparable performance with yet a precise closed form solution which consequently can lead to a more efficient computational procedure. The existing method in this category can handle only simplified cases. For example, suppose $G(m_1, \sigma_1)$ and $F(m_2, \sigma_2)$ are two Gaussians in one dimension, where m_1, m_2, σ_1, and σ_2 are their means and standard deviations. Ignoring the constant multiple, the extended KLD between G and F in this simplified case can be defined directly from the model parameters

$$D_P(G,F) = \frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2 + \left(\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 \sigma_2^2} \right) (m_1 - m_2)^2 \quad (4)$$

While the computation of D_P is simple and it can be extended to handle higher dimension Gaussian, it can not

*The work was done when the author worked at AT&T Labs - Research as an intern student.

handle multiple mixture PDF's. Even with the possibility of simplifying the models so that (4) can be applied, the outcome often indicates that it is not effective. This can be illustrated in a simple example. Consider two GMM's $G = 1/3 \times N(-2,1) + 2/3 \times N(1,1)$ and $F = 1/3 \times N(2,1) + 2/3 \times N(-1,1)$, where $N(m,\sigma)$ is Gaussian distribution with mean m and standard deviation σ. Obviously, both G and F have two components that are distributed very differently. Hence, the distance between them is clearly not zero. To apply (4), both G and F have to be simplified into one mixture Gaussian, denoted by $G'(m_G, \sigma_G)$ and $F'(m_F, \sigma_F)$, where the new mean and standard deviation can be derived as the weighted average of mean and standard deviation from their components. This yields the same mean ($m_G = m_F = 0$) and standard deviation ($\sigma_G = \sigma_F$) for both G' and F' which leads to $D_P(G', F') = 0$. Evidently, the measure derived using extended KLD from the simplified model failed to capture the obvious difference between the two original PDF's.

Therefore, there is a need to develop other alternatives that can effectively measure the difference between mixture PDF's directly from their model parameters. This paper proposes such an alternative and, through proof and experimental results, demonstrates its usefulness. In Section 2 we present our proposed metric. In section 3 we prove that the new metric satisfies the three distance properties under certain constrains. Comparison between the proposed measure and others is given in section 4. In section 5, some of the preliminary results on applying the new metric to audio based content retrieval applications. Finally section 6 concludes the paper.

2. PARAMETRIC DISTANCE METRIC FOR MIXTURE PDF

Suppose $G(x)$ and $H(x)$ are two PDF's of mixture type,

$$G(x) = \sum_{i=1}^{N} \mu_i g_i(x), \quad H(x) = \sum_{k=1}^{K} \gamma_K h_k(x), \qquad (5)$$

where $G(x)$ is a mixture of N element PDF's $g_i(x)$, $H(x)$ is a mixture of K element PDF's $h_k(x)$, and μ_i and γ_k are corresponding weights that satisfy $\sum_{i=1}^{N} \mu_i = 1$ and $\sum_{k=1}^{K} \gamma_k = 1$. For simplicity, in rest of this paper, we will not use x explicitly in other formulae. Denote the distance between any two element PDF's g_i and h_k by $d(g_i, h_k)$, the overall distance between G and H is defined as

$$D_M(G,H) = \min_{\mathbf{w}=[w_{ik}]} \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} d(g_i, h_k), \quad s.t. \qquad (6)$$

$$w_{ik} \geq 0, \quad 1 \leq i \leq N, 1 \leq k \leq K \qquad (7)$$

$$\sum_{i=1}^{N} w_{ik} = \gamma_k, 1 \leq k \leq K, \quad \sum_{k=1}^{K} w_{ik} = \mu_i, 1 \leq i \leq N \qquad (8)$$

According to the definition, any component g_i in one mixture can interact with any other component h_k in the other mixture via weighted element distance $w_{ik} d(g_i, h_k)$. The degree of interaction is inversely proportional to the element distance and proportional to the mixture weights μ_i and γ_k. The weights w_{ik} are ultimately determined through optimizing with respect to the given constraints in (7, 8). The proposed framework can be visualized in Figure 1.

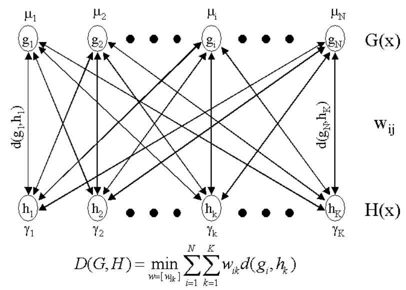

Figure 1: Distance between two mixture type PDF's.

Clearly, the solution is posed as a linear programming problem. There are many algorithms available to solve it efficiently, such as simplex tableau method. We have a total of $N \times K$ free parameters (w_{ik}'s) and $N + K$ equality constrains, where only $N + K - 1$ of them are independent. By the optimization theory, at most $N + K - 1$ of the $N \times K$ parameters will not vanish. The above problem has solution because (1) we can easily find a feasible vector that satisfy all the constrains: $w_{ik} = \mu_i \times \gamma_k$ and (2) the upper bound for the objective function exists: $\max_{ik} d(g_i, h_k)$.

The proposed metric is defined as a general framework, constructed based on element distances. Its generality is due to the fact that the element distance measure is left unspecified. Depending on different application needs, appropriate element distance measures, which may even be non-parametric, can be plugged in and the overall distance between two mixture PDF's can be computed using the same framework. Furthermore, there is no requirement about the specific type of element distribution or that each PDF should be the same type.

3. PROOF OF DISTANCE PROPERTIES

In this section, we demonstrate that if the element distance $d(g_i, h_k)$ between two mixture components g_i and h_k satisfies the three distance metric properties (1 - 3), the overall mixture distance $D_M(G,H)$, defined earlier, also does.

The proof of the first two properties is straightforward. We here focus on the proof of the third property. For any three mixture PDF's, G, H, and F, we need to show that,

$$D_M(G,H) + D_M(H,F) \geq D_M(G,F) \qquad (9)$$

The definitions of G and H are the same as (5). F is similarly defined as $F = \sum_{j=1}^{M} \xi_j f_j$, satisfying $\sum_{j=1}^{M} \xi_j = 1$.

Applying the definition in (6) to both pairs (G,H) and (H,F), we have their distances as,

$$D_M(G,H) = \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} d(g_i, h_k)$$

$$D_M(H,F) = \sum_{k=1}^{K} \sum_{j=1}^{M} v_{kj} d(h_k, f_j)$$

where w_{ik} and v_{kj} satisfy $\sum_{i=1}^{N} w_{ik} = \sum_{j=1}^{M} v_{kj} = \gamma_k$, $\sum_{k=1}^{K} w_{ik} = \mu_i$, and $\sum_{k=1}^{K} v_{kj} = \xi_j$. Then

$$D_M(G,H) + D_M(H,F)$$

$$
\begin{aligned}
&= \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik} d(g_i, h_k) + \sum_{k=1}^{K}\sum_{j=1}^{M} v_{kj} d(h_k, f_j) \\
&= \sum_{k=1}^{K} [\sum_{i=1}^{N}\sum_{j=1}^{M} \frac{w_{ik}v_{kj}}{\gamma_K}(d(g_i, h_k) + d(h_k, f_j))] \\
&\geq \sum_{k=1}^{K} [\sum_{i=1}^{N}\sum_{j=1}^{M} \frac{w_{ik}v_{kj}}{\gamma_k} d(g_i, f_j)] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M} (\sum_{k=1}^{K} \frac{w_{ik}v_{kj}}{\gamma_k}) d(g_i, f_j) \qquad (10)
\end{aligned}
$$

Let $\alpha_{ij} = \sum_{k=1}^{K} \frac{w_{ik}v_{kj}}{\gamma_k}$ then (10) can be rewritten as,

$$D_M(G, H) + D_M(H, F) \geq \sum_{i=1}^{N}\sum_{j=1}^{M} \alpha_{ij} d(g_i, f_j) \qquad (11)$$

On the other hand, for any set α_{ij} that satisfies the equation constrains in (8), the following inequality is also true since $D_M(G, F)$ is the outcome of optimization,

$$D_M(G, F) \leq \sum_{i=1}^{N}\sum_{j=1}^{M} \alpha_{ij} d(g_i, f_j) \qquad (12)$$

Actually the variables α_{ij} indeed satisfy the required constrains.

$$\sum_{i=1}^{N} \alpha_{ij} = \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{w_{ik}v_{kj}}{\gamma_k} = \sum_{k=1}^{K} v_{jk} = \xi_j$$

Similarly we have $\sum_{j=1}^{M} \alpha_{ij} = \mu_i$. Putting (11) and (12) together, we proved (9).

4. COMPARISON WITH OTHER DISTANCE MEASURES

While we have proved that the new metric proposed possesses certain properties, we also like to demonstrate that it has similar behavior as other existing measures in experimentation. In this paper, we compare it with the two previously defined measures: D_{L^2} and D_{Seq}.

For simplicity, we perform the comparison on 2 dimensional GMM's F and G, each with two mixtures. The element distance used is KLD defined in (4). Specifically, F is

$$F = 0.5N\left(\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right) + 0.5N\left(\begin{bmatrix}-1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right)$$

where $N(\mu, \delta)$ is a 2-D gaussian with mean vector μ and diagonal covariance δ. The comparison is conducted in four settings, in each of which, by perturbing the model parameters in G we observe how the three different measures (D_{L^2}, D_{Seq}, and D_M) react to the changes.

In setting one, G has exactly the same component Gaussians as F with yet variable mixture weights,

$$G = \mu N\left(\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right) + (1-\mu)N\left(\begin{bmatrix}-1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right)$$

where μ varies between 0 and 0.5.

In setting two, the two component Gaussians of G have the same weights and covariances as those of F but with variable mean vectors, changed along a circle of radius one.

$$G = 0.5N\left(\begin{bmatrix}\cos\alpha\\\sin\alpha\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right) + 0.5N\left(-\begin{bmatrix}cos\alpha\\\sin\alpha\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right)$$

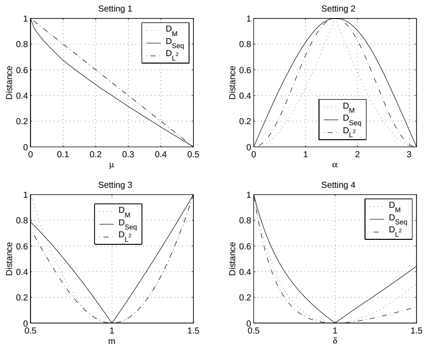

Figure 2: Behaviors of three different measures under four testing settings.

where α is in the range 0 to π.

Setting three is similar to setting two except we vary the mean vectors of G symmetrically in the first dimension.

$$G = 0.5N\left(\begin{bmatrix}m\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right) + 0.5N\left(-\begin{bmatrix}m\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right)$$

where m is from 0.5 to 1.5.

In setting four, G has the same weights and mean vectors for both components but with the covariance changing along both dimension simultaneously.

$$G = 0.5N\left(\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}\delta\\\delta\end{bmatrix}\right) + 0.5N\left(\begin{bmatrix}-1\\0\end{bmatrix}, \begin{bmatrix}\delta\\\delta\end{bmatrix}\right)$$

where δ ranges from 0.5 to 1.5.

Figure 2 shows the plotted behavior of the three measures under four different settings. All the curves are normalized so that the maximum distance is 1. From these plots, one can see that the overall behaviors of all three are consistent in all settings. D_M curve overlaps with D_{L^2} in setting one and part of setting three. In setting four, D_M falls between D_{L^2} and D_{Seq}. These plots show that the proposed new metric behaves similarly in different scenarios as the existing measures that have been widely used in practice. But the proposed metric is obviously more efficient in terms of computation. In addition, with this metric, there is no need to store or to generate data points in order to compare the difference between two PDF's. This is significant, particularly in content based search and retrieval where large amounts of data is pre-indexed, stored, preferably, in a succinct parametric form, and searched in realtime. For example, to retrieve the speech segments of a particular speaker given as an query example, all the pre-stored speaker segments in a database have to be matched against the given query sample. In this case, having a measure that can compare the similarity directly from the speaker model parameters will be much more efficient than the ones that require to generate the data points from the models first and then compare, especially when the search range is large, a realistic scenario in almost all content based retrieval tasks.

5. ENABLING MORE EFFICIENT AUDIO BASED CONTENT RETRIEVAL

To further demonstrate the usefulness of the proposed measure in practical applications, we apply it to the real problem of audio based query-by-example. Given a database of audio events, the task is to search and retrieve some given type of audio event specified by a query example. Each audio event in the database is stored as a set of parameters of a 4 mixture GMM with diagonal covariance matrix. By choosing appropriate element distance measures, we illustrate that the query/retrieval task using our proposed metric yielded comparable performance with yet more efficient computation.

In our experiment, a database containing 278 audio events is constructed from 7 hours of NBC Nightly News programs. Each event is an acoustically homogeneous segment such as a segment of speech from a particular speaker or a piece of music. A set of acoustic features (Root Mean Square energy and 12 Mel-Frequency Cepstral Coefficients) are extracted from the audio signal and are fitted by a 4 mixture GMM model (see [2] for details), whose parameters are stored in the database. During query, an audio segment is provided by users as the query example and the retrieval process is to find all the audio segments in the database that have the similar acoustic properties as the query example. For example, if the query sample is a piece of speech from president Clinton, the task is to find all Clinton speech segments from the database.

Using the given query example, a 4 mixture GMM is built and compared with all other GMM's stored in the database. Two categories of measures are used to perform the comparison of GMM models. One is the distance measure by sequence D_{Seq} and the other category is the distance measure proposed in this paper. Since the proposed distance measure uses element distance measure as building block, we choose, in this experiment, two types of element distance measures to show that the proposed framework has the flexibility of adapt to different application needs. One element distance measure is L_1 norm and the other is L_2 norm, both satisfy all three distance properties. Formally the distance between f and g can be written as,

$$d_{L_r}(f,g) = \left(\sum_{i=1}^{N} |\mu_i^f - \mu_i^g|^r + \sum_{i=1}^{N} |\sigma_i^f - \sigma_i^g|^r \right)^{1/r}, r = 1, 2$$

where N is feature dimension, μ_i^f, μ_i^g, σ_i^f and σ_i^g are the the $i-th$ means and standard deviations of f and g.

Even though the mean and standard deviation may have very different dynamic ranges, the choice is reasonable for this application because when one range is much larger than the other, the impact from the smaller one is negligible in the overall distance value. Plugging in the two chosen element distance measures, it yields two measures, denoted by D_{ME1} and D_{ME2}. Using each of the three measures, we compute distance between the given query example and each of the audio event in the database. When the distance is smaller than a threshold (can be set by user), the corresponding audio event is considered as a hit.

To evaluate the retrieval performance, we use Recall Rate (RR) and False detection Rate (FR) to plot the curve in Figure 3. Specifically, they are defined as follows. Assume that there is a total of T recorded events in database. Given a query example, there are Q events in the database that are the true match. If the retrieval process return R events as query results, among which C events are the correct match, then RR is defined as C/Q, and FR is defined as $(R-C)/(T-Q)$. Similar to the Receiver Operating Characteristic (ROC) in classical detection theory [5], we can plot a 2D graphs (similar to the PF-PD graph in detection theory) in Figure 3 to visualize the retrieval performance.

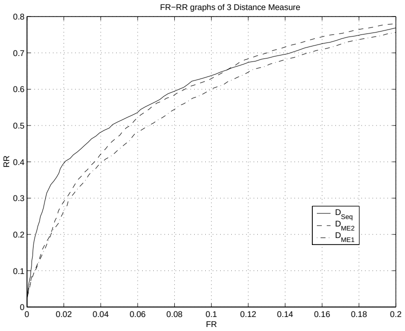

Figure 3: Performance comparison using FR-RR curves.

The query is for a particular speaker, the anchor of NBC Nightly News, Tom Brokaw. In the database, there are 55 segments that are Tom Brokaw's speeches. We used each of them as a query example and compute the corresponding FR-RR graph. Figure 3 shows the average FR-RR graph of all the query performance. As it can be seen from the figure, D_{ME1} and D_{ME2} display similar performance as D_{Seq}. When $FR < 0.11$, D_{ME2} is slightly worse than D_{Seq} and D_{ME2} is slightly better than D_{Seq} when $FR > 0.11$. While computing D_{Seq}, we choose the length of the testing sequence as 5000. For each query, the computation of D_{Seq} costs 25 times as those of D_{M1} and D_{M2}. Taking into account the significant reduction in computation, the proposed new metric outperforms the existing ones.

6. CONCLUDING REMARKS

A new distance measure based on parameters for mixture type PDF's is proposed in this paper. The new distance measure satisfies the three properties of a distance metric under certain condition. Comparison of the new distance measure with existing ones is performed using both synthetic and real data sets. Both theoretical and empirical demonstrations show that the proposed parameter based distance measure is quite promising. Furthermore, the proposed framework can be adapted to different types of applications by choosing appropriate element distance measure as building block.

7. REFERENCES

[1] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gausian Mixture Speaker Models," *IEEE Trans. on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72-83, 1995.

[2] Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, B. Shahraray, "Automatic Content Hierarchy Generation By Integrating Audio, Video and Text Information," *Proc. of IEEE ICASSP'99*, Phoenix, March, 1999.

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

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory,* John Wiley & Sons, 1991.

[5] H. L. Van Trees, *Detection, Estimation, and Modulation Theory,* John Wiley & Sons, 1967.