# A Distance measure Between GMMs Based on the Unscented Transform and its Application to Speaker Recognition

*Jacob Goldberger*

School of Engineering,
Bar-Ilan University, Israel
goldbej@eng.biu.ac.il

*Hagai Aronowitz*

Department of Computer Science,
Bar-Ilan University, Israel
aronowc@cs.biu.ac.il

## Abstract

This paper proposes a dissimilarity measure between two Gaussian mixture models (GMM). Computing a distance measure between two GMMs that were learned from speech segments is a key element in speaker verification, speaker segmentation and many other related applications. A natural measure between two distributions is the Kullback-Leibler divergence. However, it cannot be analytically computed in the case of GMM. We propose an accurate and efficiently computed approximation of the KL-divergence. The method is based on the unscented transform which is usually used to obtain a better alternative to the extended Kalman filter. The suggested distance is evaluated in an experimental setup of speakers data-set. The experimental results indicate that our proposed approximations outperform previously suggested methods.

## 1. Introduction

The Gaussian mixture distribution is a standard technique to model an acoustic speech segment. In speaker recognition or verification tasks we would like to define a distance measure based on the probabilistic representation. A similar situation occurs in speech segmentation problems where there is a need for a distance measure between any two speech segments to reflects whether the two segments are from the same speaker. Other problems where there is a need for distance measure between two acoustic segments are phoneme model clustering, speaker clustering, songs clustering and language classification. In all these problems we would like to use the Kullback-Leibler (KL) divergence which is the natural way to define a distance measure between probability distributions [9]. The KL-divergence between two distributions $f$ and $g$ is:

$$KL(f||g) = \int f(x) \log \frac{f(x)}{g(x)} \, dx$$

However, we run into difficulty due to our choice of the Gaussian mixture distribution (GMM) to model the acoustic data since there is no closed form expression for the KL-divergence between two GMMs. We can use, instead, Monte-Carlo simulations to approximate the KL-divergence between two GMMs $f$ and $g$ as follows:

$$KL(f||g) = \int f \log \frac{f}{g} \approx \frac{1}{n} \sum_{t=1}^{n} \log \frac{f(x_t)}{g(x_t)} \quad (1)$$

such that $x_1, ..., x_n$ are either the acoustic data that were used to estimate the parameters of $f$ or they are synthetic samples from $f(x)$. The symmetric version of (1) is exactly the distance measure based on the cross likelihood ratio test [11]. The drawback of the Monte-Carlo techniques is the extensive computational cost and the slow converges properties. Furthermore, due to the stochastic nature of the Monte-Carlo method, the approximations of the distance could vary in different computations. Another weakness of the Monte-Carlo approach is that in spite of the fact that we have a compact probabilistic representation, we still have to refer back to the original acoustic data.

In this study we propose a deterministic approximation for the KL-divergence between two GMMs which outperforms previously suggested methods in terms of accuracy. The organization of this paper is as follows. In the next section we review previous suggested approximations that are based on matching between the Gaussian components. In section 3 we introduce a distance measure based on the unscented transform. In section 4 we evaluate the suggested distance in an experimental setup of speakers data-set.

## 2. Matching based Approximations

Speaker recognition systems which aims to determine the identity of the talker, predominantly use Gaussian mixture models (GMMs) to represent the speaker-specific voice patterns. The main attraction of the GMM arises from its ability to provide a smooth approximation to arbitrarily shaped densities of long term spectrum. Let $f(x) = \sum_{i=1}^{n} \alpha_i f_i(x)$ and $g(x) = \sum_{j=1}^{m} \beta_j g_j(x)$ be the two Gaussian mixture densities whose KL-distance we want to compute. In many applications the two given GMMs have the same number of Gaussian components

and there is a well justified correspondence between the components. This is the case in the GMM-UBM method [10] where the two GMMs are obtained as a result of maximum a posteriori (MAP) adaptation of a universal model pre-trained using speech data from many speakers. In that case we can use the approximation:

$$KL(f||g) \approx \sum_i \alpha_i KL(f_i||g_i) \quad (2)$$

where the KL-divergence between the Gaussians $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ has the following closed form expression:

$$\frac{1}{2}(\log\frac{|\Sigma_2|}{|\Sigma_1|} + Tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^\top\Sigma_2^{-1}(\mu_1-\mu_2) - d)$$

Approximation 2 is motivated from the following upper bound [12] that is based on the chain rule for relative entropy (see [3] page 23):

$$KL(f||g) \leq KL(\alpha||\beta) + \sum_i \alpha_i KL(f_i||g_i)$$

If no correspondence between the two GMM (with same number of) components is assumed, we can still utilize the fact that for every permutation $\pi$ defined on the set $\{1, ..., n\}$, the GMM $g_\pi = \sum_{i=1}^n \beta_{\pi(i)} g_{\pi(i)}$ is exactly the same as $g$. Hence:

$$KL(f||g) \leq \min_\pi \left( KL(\alpha||\beta_\pi) + \sum_i \alpha_i KL(f_i||g_{\pi(i)}) \right)$$

which yields the following approximation:

$$KL(f||g) \approx \min_\pi \sum_i \alpha_i KL(f_i||g_{\pi(i)}) \quad (3)$$

The Hungarian method [8] can be used to solve this minimization problem in $O(n^3)$ operations. A more efficiently computed approximation is:

$$KL(f||g) \approx \sum_{i=1}^n \alpha_i \min_{j=1}^m KL(f_i||g_j) \quad (4)$$

which can be also applied to the general case where $f$ and $g$ do not necessarily have the same number of components. Approximation (4) is based on a matching function between each element of $f$ and an element of $g$ that is most similar to it (see Figure 1). Variants of the matching based approximation of the KL-divergence between two GMMs were suggested by Vasconcelos [13] and Goldberger et al. [4]. The main difference between the methods is in the matching function between the elements of the two GMMs.

The matching based method is a good approximation for the KL-divergence if the Gaussian elements are far apart (i.e. given a sample point you can almost surely know from which Gaussian component it was sampled). However, if there is a significant overlap between the Gaussian elements, then a matching of a single component of $g(x)$ with each component of $f(x)$ can cause a significant performance degradation. In this work we aim to solve this drawback by presenting a new efficient method that is more accurate and more robust.

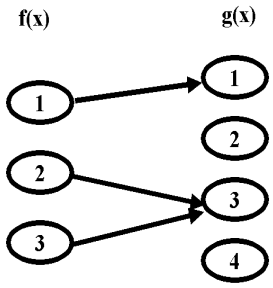


Figure 1: A matching function between the Gaussian components of two GMMs

## 3. Unscented Transform based Approximation

To handle overlapping situations we propose another approximation based on the unscented transform. The unscented transformation is a method for calculating the statistics of a random variable which undergoes a nonlinear transformation [5]. It is successfully used for nonlinear filtering. The Unscented Kalman filter (UKF) [6] is more accurate, more stable and far easier to implement than the extended Kalman filter (EKF). In cases where the process noise is Gaussian it is also better than the particle filter which is based on Monte-Carlo simulations. Unlike the EKF which uses the first order term of the Taylor expansion of the non-linear function, the UKF uses the true nonlinear function and approximates the distribution of the function output. In this section we show how we can utilize the unscented transform mechanism to obtain an approximation for the KL-divergence between two GMMs.

We shall first review the unscented transform. Let $x$ be a $d$-dimensional normal random variable $x \sim f(x) = N(\mu, \Sigma)$ and let $h(x) : R^d \rightarrow R$ be an arbitrary nonlinear function. We want to approximate the expectation of $h(x)$ which is $\int f(x)h(x)dx$. The unscented transform approach is the following. A set of $2d$ "sigma" points are chosen as follows:

$$\begin{aligned} x_k &= \mu + (\sqrt{d\Sigma})_k \qquad k = 1, ..., d \\ x_{d+k} &= \mu - (\sqrt{d\Sigma})_k \qquad k = 1, ..., d \end{aligned}$$

such that $(\sqrt{\Sigma})_k$ is the $k$-th column of the matrix square root of $\Sigma$. Let $UDU^\top$ be the singular value decom-

position of $\Sigma$, such that $U = \{U_1, ..., U_d\}$ and $D = \text{diag}\{\lambda_1, ..., \lambda_d\}$ then $(\sqrt{\Sigma})_k = \sqrt{\lambda_k} U_k$. These sample points completely capture the true mean and variance of the normal distribution $f(x)$ (see Figure 2). The uniform distribution over the points $\{x_k\}_{k=1}^{2d}$ has mean $\mu$ and covariance matrix $\Sigma$. Given the "sigma" points, we define the following approximation:

$$\int f(x)h(x)dx \approx \frac{1}{2d}\sum_{k=1}^{2d} h(x_k). \qquad (5)$$

Although this approximation algorithm resembles a Monte-Carlo method, no random sampling is used thus only a small number of points are required. It can be verified that if $h(x)$ is a linear or even a quadratic function then the approximation is precise. The basic unscented method can be generalized. The mean of the Gaussian distribution $\mu$ can be also included in the set of sigma points. Scaling parameters can provide an extra degree of freedom to control the scaling of the sigma points further or towards $\mu$ [7].

The unscented transform can be used to approximate he KL-divergence between the following two GMMs:

$$f = \sum_{i=1}^{n} \alpha_i f_i = \sum_{i=1}^{n} \alpha_i N(\mu_i, \Sigma_i) \quad \text{and} \quad g = \sum_{j=1}^{m} \beta_j g_j$$

Since $KL(f||g) = \int f \log f - \int f \log g$, it is sufficient to show how we can approximate $\int f \log g$. The linearity of the construction of $f$ from its components yields:

$$\int f \log g = \sum_{i=1}^{n} \alpha_i \int f_i \log g = \sum_{i=1}^{n} \alpha_i E_{f_i}(\log g)$$

Assume that $x$ is a Gaussian random variable $x \sim f_i$ then $E_{f_i}(x) = \mu_i$ and $E_{f_i}(\log g(x))$ is the mean of the non-linear function $\log g(x)$ which can be approximated using the unscented transform. Hence:

$$\int f \log g \approx \frac{1}{2d}\sum_{i=1}^{n} \alpha_i \sum_{k=1}^{2d} \log g(x_{i,k})$$

such that:

$$\begin{aligned} x_{i,k} &= \mu_i + (\sqrt{d\Sigma_i})_k \qquad k = 1, ..., d, \quad (6) \\ x_{i,d+k} &= \mu_i - (\sqrt{d\Sigma_i})_k \qquad k = 1, ..., d. \end{aligned}$$

If the covariance matrices of the two GMMs are diagonal (as it usually is case in speech segment modelling) the computational complexity of the unscented approximation is significantly reduced. Assume the covariance matrices of the components of $f$ have the following form:

$$\Sigma_i = diag(\sigma_{i,1}^2, ..., \sigma_{i,d}^2) \qquad i = 1, .., n$$

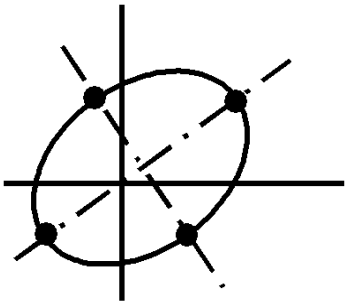

Figure 2: The sigma points of the unscented transform

then the sigma points are simply:

$$\mu_i \pm \sqrt{d}\,\sigma_{i,k} \qquad k = 1, ..., d$$

The diagonal structure of the covariance matrices of the components of $g$ can be utilized to further reduce the complexity of computing the value $g$ on the $2nd$ sigma points.

## 4. Experimental Setup and Results

The proposed dissimilarity measure was evaluated in a task of speaker recognition. The front-end feature extraction is based on Mel-frequency cepstrum processing (MFCC). For each frame the front-end produces a feature vector consisting of 13 Mel-cepstral coefficients. The first time derivatives of the elements are appended to the feature vector. An energy based voice activity detector is used to remove silence. The GMMs were trained by adapting universal background models (UBM) as in [10]. We used the T-norm [1] technique in order to normalize the dissimilarity scores. We trained the universal background models using the SPIDRE corpus [14] which consists of telephone conversational speech. We used the NIST-2004 speaker evaluation data set [15] for evaluation of the proposed dissimilarity measure. The primary data set was used for selecting both target speakers and test data. The data set consists of 616 1-sided single conversations for training 616 target models, and 1174 1-sided test conversations. All conversations are about 5 minutes long and originate from various channels and handset types. In order to increase the number of trials, every target model was tested against every test session. We evaluated the proposed dissimilarly measure in 3 testing conditions. In the first test condition we tested full conversations which were modeled by mixtures of 2048 Gaussians. In the second test condition we tested 30 seconds speech segments (after silence removal) which were modeled by mixtures of 128 Gaussians. In the third test condition we tested 3 seconds speech segments (after silence removal) which were also modeled by mixtures of 128 Gaussians. Table 1 summarizes the identification results. We use two performance measures: EER and minimal detection cost

function (minDCF) [15].

| | EER (%) | min-DCF |
|---|---|---|
| Matching based algorithm - full conversations | 13.2 | 0.049 |
| Proposed algorithm - full conversations | 13.2 | 0.048 |
| Matching based algorithm - 30sec | 15.8 | 0.056 |
| Proposed algorithm - 30sec | 15.2 | 0.055 |
| Matching based algorithm - 3sec | 17.8 | 0.068 |
| Proposed algorithm - 3sec | 17.6 | 0.066 |

Table 1: Comparison of speaker recognition results using matching based approximation and unscented based approximation on several testing conditions.

## 5. Conclusions

In this paper we described a method for approximating the KL-divergence between Gaussian mixture densities. The performance of this method was demonstrated on a standard speaker recognition task. In all the experiments conducted, the unscented approximation achieves the best results. We expect that utilizing the proposed method, we can achieve even more performance improvement in tasks where each speaker signal is modeled with few Gaussians learned specifically for the speaker without the adaptation phase.

## 6. References


[1] R. Auckenthaler, H. Carey M., and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems", Digital Signal Processing, vol. 10, pp. 42-54, 2000.

[2] H. Beigi, S. Maes and J. Sorensen, "A distance measure between collections of distributions and its application to speaker recognition", ICASSP, 1998.

[3] T. Cover and J.Thomas, "Elements of information theory", *Wiley Series in Telecommunications*, Jhon Wiley and Sons, New-York, USA, 1991.

[4] J. Goldberger, H. Greenspan and S. Gordon, "An efficient similarity measure based on approximations of KL-divergence between two Gaussian mixtures", *International Conference on Computer Vision (ICCV)*, 2003.

[5] S. Julier and J. K. Uhlmann, "A general method for approximating nonlinear transfromations of probability distributions", *Technical report*, RRG, Dept. of Engineering Science, University of Oxford, 1996.

[6] S. Julier and J. K. Uhlmann, "A new extension of the Kalman filter to non-linear systems", *Proc of AeroSense: The 11th International Symposium on Aerospace/Defence Sensing, Simulation and Control*, Florida, 1997.

[7] S. Julier, "The scaled unscented transformation", to appear in Automatica, 2000.

[8] H. W. Kuhn, "The Hungarian method for the assignment problem", *Naval Research Logistics Quarterly*, Vol., pp 83-97,1955.

[9] S. Kullback, "Information theory and statistics", Dover Publications, New York, 1968.

[10] D. Reynolds, T. Quatieri and R. Dunn, "Speaker verification using adapted gaussian mixture models", *Digital Sugnal Processing*, 10:19-41, 2000.

[11] D. Reynolds, E. Singer, B, Carlson G. O'Leary, J. Mclaughlin and M. Zissman, "Blind clustering of speech uttrances based on speaker and language characteristics", Proc ICSLP, pp. 3193-3196, 1998.

[12] Y. Singer and M. K. Warmuth "Batch and on-line parameter estimation of Gaussian mixtures based on the joint entropy", *Advances in Neural Information processsing Systems (NIPS)*, pp 578-584, 1998.

[13] N. Vasconcelos , "On the complexity of probabilistic Image Retrieval", *Proc. of the Int. Conference on Computer Vision*, 2001.

[14] Linguistic Data Consortium, SPIDRE documentation file, http://www.ldc.upenn.edu/Catalog /readme_files/spidre.readme.html.

[15] "The NIST Year 2004 Speaker Recognition Evaluation Plan", http://www.nist.gov/speech/ tests/spk/2004/.