EVALUATION OF DISTANCE MEASURES BETWEEN GAUSSIAN MIXTURE MODELS OF MFCCS

Jesper Højvang Jensen
Aalborg University
Dept. Electron. Syst.

Daniel P.W. Ellis
Columbia University
LabROSA

Mads G. Christensen
Aalborg University
Dept. Electron. Syst.

Søren Holdt Jensen
Aalborg University
Dept. Electron. Syst.

ABSTRACT

In music similarity and in the related task of genre classification, a distance measure between Gaussian mixture models is frequently needed. We present a comparison of the Kullback-Leibler distance, the earth movers distance and the normalized L2 distance for this application. Although the normalized L2 distance was slightly inferior to the Kullback-Leibler distance with respect to classification performance, it has the advantage of obeying the triangle inequality, which allows for efficient searching.

1 INTRODUCTION

A common approach in computational music similarity is to extract mel-frequency cepstral coefficients (MFCCs) from a song, model them by a Gaussian mixture model (GMM) and use a distance measure between the GMMs as a measure of the musical distance between the songs [2, 3, 5]. Through the years, a number of distance measures between GMMs have been suggested, such as the Kullback-Leibler (KL) distance [2], optionally combined with the earth movers distance (EMD) [3]. In this article, we evaluate the performance of these two distance measures between GMMs together with the normalized L2 distance, which to our knowledge has not previously been used for this application.

2 MEASURING MUSICAL DISTANCE

In the following, we shortly describe the Gaussian mixture model and the three distance measures between GMMs we have tested. Note that if a distance measure satisfies the triangle inequality, i.e., $d(p_1, p_3) \leq d(p_1, p_2) + d(p_2, p_3)$ for all values of $p1, p2$ and $p3$, then a nearest neighbor search can be speeded up by precomputing some distances. Assume we are searching for the nearest neighbor to p, and that we have just computed the distance to p_1. If we already know the distance between p_1 and p_2, then the distance to p_2 is bounded by $d(p, p_2) \geq d(p_1, p_2) - d(p_1, p)$. If the distance to the currently best candidate is smaller than $d(p_1, p_2) - d(p_1, p)$, we can discard p_2 without computing $d(p, p_2)$.

This research was supported by the Intelligent Sound project, Danish Technical Research Council grant no. 26–04–0092, and the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences grant no. 274–06–0521.

2.1 Gaussian Mixture Models

Due to intractability, the MFCCs extracted from a song are typically not stored but are instead modelled by a GMM. A GMM is a weighted sum of multivariate Gaussians:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} c_k \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right),$$

where K is the number of mixtures. For $K = 1$, a simple closed-form expression exists for the maximum-likelihood estimate of the parameters. For $K > 1$, the k-means algorithm and optionally the expectation-maximization algorithm are used to estimate the parameters.

2.2 Kullback-Leibler Distance

The KL distance is an information-theoretic distance measure between probability density functions. It is given by $d_{\mathrm{KL}}(p_1, p_2) = \int p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} d\boldsymbol{x}$. As the KL distance is not symmetric, a symmetrized version, $d_{\mathrm{sKL}}(p_1, p_2) = d_{\mathrm{KL}}(p_1, p_2) + d_{\mathrm{KL}}(p_2, p_1)$, is usually used in music information retrieval. For Gaussian mixtures, a closed form expression for $d_{\mathrm{KL}}(p_1, p_2)$ only exists for $K = 1$. For $K > 1$, $d_{\mathrm{KL}}(p_1, p_2)$ is estimated using stochastic integration or the approximation in [4]. The KL distance does not obey the triangle inequality.

2.3 Earth Movers Distance

In this context the EMD is the minimum cost of changing one mixture into another when the cost of moving probability mass from component m in the first mixture to component n in the second mixture is given [3]. A common choice of cost is the symmetrized KL distance between the individual Gaussian components. With this cost, the EMD does not obey the triangle inequality.

2.4 Normalized L2 Distance

Let $p_i'(\boldsymbol{x}) = p_i(\boldsymbol{x}) / \sqrt{\int p_i(\boldsymbol{x})^2 d\boldsymbol{x}}$, i.e., $p_i(\boldsymbol{x})$ scaled to unit L2-norm. We then define the normalized L2 distance by $d_{\mathrm{nL2}}(p_1, p_2) = \int (p_1'(\boldsymbol{x}) - p_2'(\boldsymbol{x}))^2 d\boldsymbol{x}$. Since the ordi-

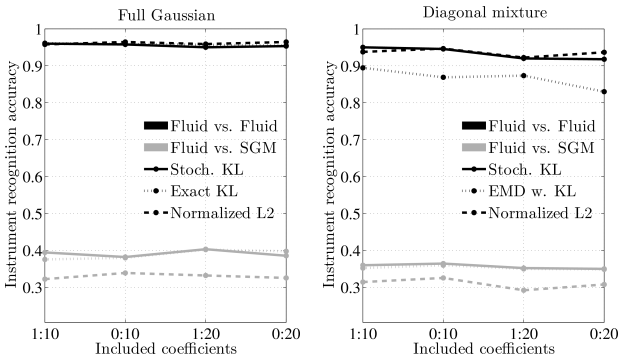

Figure 1. Instrument recognition results. Labels on the x-axis denotes the number of MFCCs retained, i.e. 0:10 means retaining the first 11 coefficients including the 0th. "Fluid" and "SGM" denotes the Fluid R3 and SGM 180 sound fonts, respectively.

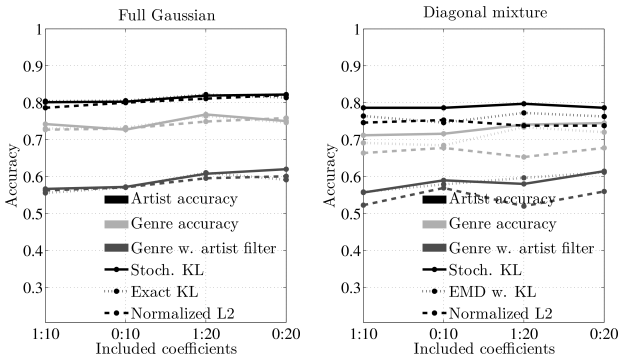

Figure 2. Genre and artist classification results for the MIREX 2004 database.

nary L2 distance obeys the triangle inequality, and since we can simply prescale all GMMs to have unit L2-norm and then consider the ordinary L2 distance between the scaled GMMs, the normalized L2 distance will also obey the triangle inequality. Also note that $d_{\text{nL2}}(p_1, p_2)$ is nothing but a continuous version of the cosine distance [6], since $d_{\text{nL2}}(p_1, p_2) = 2(1 - \int p_1'(\boldsymbol{x}) p_2'(\boldsymbol{x}) d\boldsymbol{x})$. For GMMs, closed form expressions for the normalized L2 distance can be derived for any K from [1, Eq. (5.1) and (5.2)].

3 EVALUATION

We have evaluated the symmetrized KL distance computed by stochastic integration using 100 samples, EMD with the exact, symmetrized KL distance as cost, and the normalized L2 distance. We extract the MFCCs with the ISP toolbox R1 using default options [1]. To model the MFCCs we have both used a single Gaussian with full covariance matrix and a mixture of ten Gaussians with diagonal covariance matrices. With a single Gaussian, the EMD reduces to the exact, symmetrized KL distance. Furthermore, we have used different numbers of MFCCs. As the MFCCs are timbral features and therefore are expected to model instrumentation rather than melody or rhythm, we have evaluated the distance measures in a synthetic nearest neighbor instrument classification task using 900 synthesized MIDI songs with 30 different melodies and 30 different instruments. In Figure 1, results for using a single sound font and results where the query song is synthesized by a different sound font than the songs it is compared to are shown. The former test can be considered a sanity test, and the latter test reflects generalization behaviour. Moreover, we have evaluated the distance measures using 30 s excerpts of the training songs from the MIREX 2004 genre classification contest, which consists of 729 songs from 6 genres. Results for genre classification, artist identification and genre classification with an artist filter (see [5]) are shown in Figure 2.

[1] `http://isound.kom.auc.dk/`

4 DISCUSSION

As the results show, all three distance measures perform approximately equal when using a single Gaussian with full covariance matrix, except that the normalized L2 distance performs a little worse when mixing instruments from different sound fonts. Using a mixture of ten diagonal Gaussians generally decrease recognition rates slightly, although it should be noted that [2] recommends using more than ten mixtures. For ten mixtures, the recognition rate for the Kullback-Leibler distance seems to decrease less than for the EMD and the normalized L2 distance. From these results we conclude that the cosine distance performs slightly worse than the Kullback-Leibler distance in terms of accuracy. However, with a single Gaussian having full covariance matrix this difference is negligible, and since the cosine distance obeys the triangle inequality, it might be preferable in applications with large datasets.

5 REFERENCES

[1] P. Ahrendt, "The multivariate gaussian probability distribution," Technical University of Denmark, Tech. Rep., 2005.

[2] J.-J. Aucouturier, "Ten experiments on the modelling of polyphonic timbre," Ph.D. dissertation, University of Paris 6, France, 2006.

[3] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2001, pp. 745 – 748.

[4] E. Pampalk, "Speeding up music similarity," in *2nd Annual Music Information Retrieval eXchange*, London, 2005.

[5] ——, "Computational models of music similarity and their application to music information retrieval," Ph.D. dissertation, Vienna University of Technology, Austria, 2006.

[6] J. R. Smith, "Integrated spatial and feature image systems: Retrieval, analysis and compression," Ph.D. dissertation, Columbia University, New York, 1997.