Evaluation of Distance Measures Between Gaussian Mixture Models of MFCCs

Jesper Højvang Jensen[1], Daniel P.W. Ellis[2], Mads G. Christensen[1] and Søren Holdt Jensen[1]

[1] Dept. Electron. Syst., Aalborg University
[2] LabROSA, Columbia University

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. A Statistical Timbre Model

We use the common approach of extracting 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, 4, 6].

1.1 Mel-Frequency Cepstral Coefficients

MFCCs are a compact, perceptually based representation of speech frames [3]. They are computed as follows:

1. Estimate the log-amplitude or log-power spectrum of 20–30 ms of speech.
2. Sum the contents of neighboring frequency bins in overlapping bands distributed according to the mel-scale.
3. Compute the discrete cosine transform of the bands.
4. Discard high frequency coefficients from the cosine transform.

1.2 Gaussian Mixture Models

We model the MFCCs from each song by a Gaussian mixture model (GMM):

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

where K is the number of components. For $K = 1$, a 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 needed.

2. Distance Measures Between GMMs

As distance measure between the GMMs, we have evaluated the symmetrized Kullback-Leibler distance, the earth movers distance and the normalized L2 distance.

2.1 Kullback-Leibler Distance

The KL distance is given by

$$d_{\mathrm{KL}}(p_1, p_2) = \int p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} d\boldsymbol{x}. \tag{1}$$

As the KL distance is not symmetric, we use a symmetrized version,

$$d_{\mathrm{sKL}}(p_1, p_2) = d_{\mathrm{KL}}(p_1, p_2) + d_{\mathrm{KL}}(p_2, p_1). \tag{2}$$

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 [5].

2.2 Earth Movers Distance

The earth movers distance (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, c_{mn}, is given [7, 4]. Let a_{1k} be the weights of the Gaussians in $p_1(\boldsymbol{x})$, and a_{2k} the weights of $p_2(\boldsymbol{x})$, then $d_{\mathrm{EMD}}(p_1, p_2)$ is given by

$$d_{\mathrm{EMD}}(p_1, p_2) = \min \sum_m \sum_n c_{mn} f_{mn} \tag{3}$$

subject to

$$f_{ij} \geq 0 \tag{4}$$

$$\sum_i f_{ij} = a_{1j} \tag{5}$$

$$\sum_j f_{ij} = a_{2j} \tag{6}$$

As cost we use the symmetrized KL distance between the individual Gaussian components.

2.3 Normalized L2 Distance

Let $p_i'(\boldsymbol{x})$ be $p_i(\boldsymbol{x})$ scaled to unit L2-norm, i.e.,

$$p_i'(\boldsymbol{x}) = p_i(\boldsymbol{x}) / \sqrt{\int p_i(\boldsymbol{x})^2 d\boldsymbol{x}}. \tag{7}$$

We then define the normalized L2 distance as

$$d_{\mathrm{nL2}}(p_1, p_2) = \int (p_1'(\boldsymbol{x}) - p_2'(\boldsymbol{x}))^2 d\boldsymbol{x}. \tag{8}$$

This distance measure obeys the triangle inequality. 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

The three distance measures have been evaluated using a proof of concept data set and the MIREX 2004 Genre classification training set.

3.1 Proof of Concept

As MFCCs are supposed to model the spectral envelope, which mostly depends on instrumentation, a synthetic proof of concept data set is generated by 900 MIDI songs from 30 melodies and 30 instruments. The songs have been synthesized using two different sound fonts to give an idea of the generalization behaviour.

Figure 1: *Instrument recognition results. X-axis denotes the number of MFCCs, 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.*

3.2 MIREX 2004 Genre

The MIREX 2004 Genre classification training set consists of 729 songs from 6 genres. The GMMs are trained on 30 s excerpts from the middle of each song.

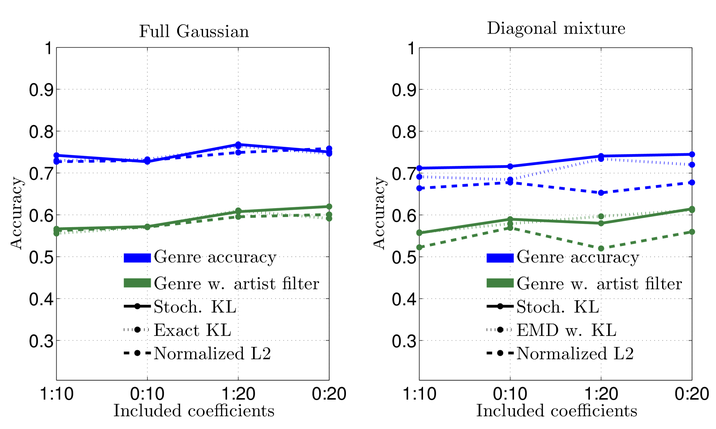

Figure 2: *Genre and artist classification results for the MIREX 2004 database.*

4. Conclusion

- All three distance measures perform approximately equally when using a single Gaussian having a 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.
- For ten mixtures, the recognition rate for the Kullback-Leibler distance seems to decrease less than for the EMD and the normalized L2 distance.

Name	Closed form	Speed	Triangle inequality	Accuracy
Sym. KL, $K = 1$	Yes	Fast	No	Highest
Sym. KL, $K > 1$	No	Slow	No	High
EMD w. sym. KL	No	Medium	No	Medium
Normalized L2	Yes	Slow	Yes	Lowest

Table 1: *Overview of the tested distance measures between Gaussian mixture models. K denotes the number of components in the GMM.*

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] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366, 1980.

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

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

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

[7] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proc. IEEE Int. Conf. Computer Vision*, 1998, pp. 59–66.