INFORMED SELECTION OF FRAMES FOR MUSIC SIMILARITY COMPUTATION

Klaus Seyerlehner, Tim Pohle
Dept. of Computational Perception,
Johannes Kepler University, Linz, Austria
klaus.seyerlehner@jku.at

Gerhard Widmer, Dominik Schnitzer
Austrian Research Institute for AI
Vienna, Austria
gerhard.widmer@jku.at

ABSTRACT

In this paper we present a new method to compute frame based audio similarities, based on nearest neighbour density estimation. We do not recommend it is as a practical method for large collections because of the high runtime. Rather, we use this new method for a detailed analysis to get a deeper insight on how a bag of frames approach (BOF) determines similarities among songs, and in particular, to identify those audio frames that make two songs similar from a machine's point of view. Our analysis reveals that audio frames of very low energy, which are of course not the most salient with respect to human perception, have a surprisingly big influence on current similarity measures. Based on this observation we propose to remove these low-energy frames before computing song models and show, via classification experiments, that the proposed frame selection strategy improves the audio similarity measure.

1. INTRODUCTION

In the *iPod*-age where private music collections contain thousands of songs and commercial music catalogs consist of millions of songs, music recommender systems become more and more important especially for searching and browsing music catalogs. One can classify these recommender systems based on how the recommendations are generated. There exist three general recommendation approaches [1], namely the *collaborative filtering approach*, the *content-based approach* and *hybrid approaches*.

The focus of this paper is on content-based music recommendation. Content-based recommender systems do not rely on any kind of metadata collected from a user community, but analyze and extract descriptive information about music titles directly from the audio signals themselves. With respect to music recommendation, the so-called *bag of frames approach* (BOF) is a state-of-the art method to compute timbral similarity [2, 3]. The main idea is to extract spectral features from individual analysis frames and to model audio signals (music recordings) via the long-term statistical distribution of their local frame features. Most commonly, Mel Frequency Cepstrum Coefficients (MFCCs) are used as local spectral features, MFCC distributions are modeled as Gaussians or Gaussian mixtures, and the Kullback-Leibler (KL) Divergence is used as a measure of similarity between such distributions.

The main goal of this paper is to try to identify those frames which cause two songs to be judged similar by a BOF approach, to get a deeper insight on how the BOF approach works, and to find ways to improve it. To this end, we develop a way to analyze frame-based music similarity via nearest neighbour density estimation, and use this to find out which audio frames contribute most to the estimated similarity. That is described in section 2. In section 3 we discuss some of our findings using this analysis method, and focus on the observation that it is especially low energy frames that seem to have an unduly strong influence on the BOF similarity judgment. We propose a strategy to remove those low energy frames before computing song models, and experimentally evaluate this frame selection strategy against all-frame models and a random frame removal strategy in section 4.

2. MUSIC SIMILARITY BASED ON NEAREST NEIGHBOUR DENSITY ESTIMATION

To model the distribution of the MFCC vectors several approaches have been considered so far. In the beginning, the distributions were modeled using Gaussian Mixture Models (GMMs), which belong to the category of a semi-parametric distributions [4]. The most popular variant nowadays is to simply use a single multivariate Gaussian (SG) to model the distribution, because this turned out to be comparable to the GMM approach in terms of quality, but creating the models and especially comparing models is significantly faster [5, 6]. Most commonly the distribution models are then compared using the Kullback-Leibler (KL) divergence [7]. The KL-divergence is a measure of relative entropy between two probability distributions P and Q and is defined by equation (1).

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx \tag{1}$$

Recently also non-parametric distribution models like vector quantization (VQ) have been in the focus of research [8, 9]. In this paper we make use of an even more direct approach to modeling a distribution by applying *nearest neighbour density estimation*. Using a nearest neighbour density estimation method has several advantages compared to GMMs and SG. First of all, common parametric forms rarely fit densities actually encountered in data, in particular because all the *'classical'* parametric densities (e.g. the Gaussian distribution) are unimodal, whereas many practical problems involve multi-modal densities. A general advantage of non-parametric in contrast to parametric procedures is that they can be used with any distributions and without the assumption that the form of the underlying densities are known. The big drawback, however, is the extreme computational cost of nearest neighbour estimation compared to other density estimation methods. Therefore nearest neighbour estimation is not a practical solution for computing similarity in recommendation, but it is well suited for studying and analyzing BOF approaches.

In the next subsections we first discuss how to derive densities using nearest neighbour estimation and then how to identify those frames which contribute most to the KL divergence of two distributions. Finally, we perform some genre classification experiments to check that this method of estimating similarities is meaningful. The results are comparable in terms of quality to an

implementation of the single Gaussian approach, indicating that our approach is correct.

2.1. Nearest Neighbour (NN) density estimation

Given a set of n points sampled from an arbitrary distribution, one can estimate the density $p(\mathbf{x})$ at or around a point $\mathbf{x}$ by counting the number of sample points that fall into a small region around $\mathbf{x}$. If k is the number of points in that region and V is the volume of the region, one can estimate the density $\hat{p}(\mathbf{x})$ at point $\mathbf{x}$ according to equation (2).

$$\hat{p}(\mathbf{x}) = \frac{k/n}{V} \tag{2}$$

This method is called *nearest neighbour density estimation* [10, 11]. Given infinitely many sample points, the radius r could be chosen infinitely small and the estimate $\hat{p}(\mathbf{x})$ would converge toward the true probability $p(\mathbf{x})$. In practice, given a finite sample set, r is a crucial application-specific parameter. In the following, we will use NN density estimation to model feature distributions. Together with an estimation of the KL divergence, this will permit us to analyze the role of individual sample points in the computation of similarities.

2.2. Estimating the KL divergence

Given n MFCC vectors of a song, we can directly estimate the density at each vector $\mathbf{x}$ by counting the number k of MFCC vectors with a distance $\leq r$ from $\mathbf{x}$, and taking k/n as an estimate of (proportional to) the density of $p(\mathbf{x})$. For two sets X_P and X_Q representing two songs, we first reduce the larger set such that both sets have equal size ($|X_P| = |X_Q|$). This can be easily achieved by randomly removing points from the larger set and should have little influence on the overall distribution. We take the usual approach to compare two distributions P and Q of two songs by computing the discrete KL divergence:

$$\hat{D}_{KL}(P||Q) = \sum_{\mathbf{x}} \hat{p}(\mathbf{x}) \log \frac{\hat{p}(\mathbf{x})}{\hat{q}(\mathbf{x})}, \tag{3}$$

where $\mathbf{x} \in X_P \cup X_Q$. Inserting the density estimates according to equation 2 we arrive at the following expression for the KL divergence, where $k_{\mathbf{x},p}$ denotes the number of points in the small region around $\mathbf{x}$ for distribution P and $k_{\mathbf{x},q}$ for distribution Q respectively. n_p is the number of sample points of distribution P and n_q the number of sample points from distribution Q.

$$\hat{D}_{KL}(P||Q) = V \sum_{\mathbf{x}} k_{\mathbf{x},p}/n_p \log \frac{k_{\mathbf{x},p}/n_p}{k_{\mathbf{x},q}/n_q} \tag{4}$$

Note that we have reduced one set of sample points such that the number of sample points n is the same for both distributions P and Q. Therefore the estimate of the KL divergence can be further reduced.

$$\hat{D}_{KL}(P||Q) = \frac{V}{n} \sum_{\mathbf{x}} k_{\mathbf{x},p} \log \frac{k_{\mathbf{x},p}}{k_{\mathbf{x},q}} \tag{5}$$

Unfortunately, this approximation is numerically unstable, because whenever the NN estimate of the probability density $\hat{q}(\mathbf{x})$ is zero (which can easily happen), $\log \frac{\hat{p}(x)}{\hat{q}(x)}$ will be undefined. We decided to circumvent this problem by increasing $k_{\mathbf{x},p}$ and $k_{\mathbf{x},q}$ by one such that the estimates $\hat{p}(\mathbf{x})$ and $\hat{q}(\mathbf{x})$ cannot become zero. We can also derive the symmetric KL divergence $D_{KL_{sym}}$.

$$\begin{aligned}
\hat{D}_{KL_{sym}}(P||Q) &= \hat{D}_{KL}(P||Q) + \hat{D}_{KL}(Q||P) \\
&= \sum_{\mathbf{x}} \hat{p}(\mathbf{x}) \log \frac{\hat{p}(\mathbf{x})}{\hat{q}(\mathbf{x})} + \sum_{\mathbf{x}} \hat{q}(\mathbf{x}) \log \frac{\hat{q}(\mathbf{x})}{\hat{p}(\mathbf{x})} \\
&= \sum_{\mathbf{x}} \hat{p}(\mathbf{x}) \log \frac{\hat{p}(\mathbf{x})}{\hat{q}(\mathbf{x})} + \hat{q}(\mathbf{x}) \log \frac{\hat{q}(\mathbf{x})}{\hat{p}(\mathbf{x})} \\
&= \sum_{\mathbf{x}} (\hat{p}(\mathbf{x}) - \hat{q}(\mathbf{x})) \log \frac{\hat{p}(\mathbf{x})}{\hat{q}(\mathbf{x})} \\
&= \frac{V}{n} (\sum_{\mathbf{x}} (k_{\mathbf{x},p} - k_{\mathbf{x},q}) \log \frac{k_{\mathbf{x},p}}{k_{\mathbf{x},q}})
\end{aligned} \tag{6}$$

Since we use the symmetric KL divergence as a distance measure and V/n is a constant factor, we can safely neglect V/n for our purpose. For the NN approach the KL divergence is a sum over all sample points. Therefore we can easily figure out how much an individual vector $\mathbf{x}$ contributes to the overall KL divergence. We have developed a little tool that loads two audio files, computes their spectral representation, and calculates the similarity value based on nearest neighbour density estimation[1]. To analyze which frames make two songs similar from a machine's point of view, it sorts and visualizes the audio frames according to their contribution to the overall distance (see equation 6). It is worth mentioning that in our implementation we do not compute the real number of neighbours within the radius r, but only compute the approximate nearest neighbours using *Locality Sensitive Hashing* because of performance reasons, since computing the exact nearest neighbours would require to compute the euclidean distance between all frames.

2.3. Validation of the Approach

To validate the nearest neighbour density approach and our implementation thereof, we compare it to a standard Single Gaussian (SG) method with KL-divergence, on a genre classification task (as is common in MIR). The validation dataset contains 10 genres each consisting of 10 songs. The main reason for this small classification dataset is the high computational cost of comparing two distributions using the nearest neighbour density method. To prevent artist effects there are no two songs by the same artist in this collection. On this validation dataset we get a 10-NN classification accuracy of **45.4%** for the single Gaussian model, and a 10-NN classification accuracy of **42.1%** for the nearest neighbour density approach. For NN density estimation, we chose a radius of $r = 32$. This setting seemed plausible after analyzing some songs (from a different collection) using the developed tool, however we did not optimize this parameter in any way. It is interesting that the NN approach performs about 3 percentage points worse than the SG approach. Choosing a more appropriate radius r might reduce the difference in classification accuracy. But that is not the point of this experiment. The important aspect of this experiment is that the NN approach itself and the implementation thereof is correct, because the classification accuracy is far above the baseline and close to the result obtained using the SG approach. Thus we can make use of this approach to analyze and understand the BOF in more detail.

[1] http://www.cp.jku.at/people/seyerlehner/nn_density/kl.html

3. A SIMPLE FRAME SELECTION STRATEGY

Using the developed tool we have analyzed many pairs of songs to find out which frames make them appear similar to the machine. What jumps to the eye is that frames with rather low energy tend to contribute quite a lot to the similarity judgment. This is especially interesting as those frames are of course not the most salient ones with respect to human perception. From a technical point of view frames with low energy will of course have a low euclidean distance to each other, which implies that the density of low energy frames will be high. Thus the KL divergence for those frames will only be low if the other song has an equal amount of e.g. silent or almost silent frames. Consequently songs having low energy frames will more likely match songs that have low energy frames as well, although from a human point of view the similarity of two songs is surely related to the dynamic parts of a song and is not too much influenced by the silent parts of a song. We therefore decided to investigate whether removing those low energy frames before building a model would improve audio similarity measures. To select frames of high energy, we determine the energy of each frame after mapping onto the mel-scale by summing across the mel bands. Then all the frames are sorted according to their energy and a given percentage of frames is dropped (e.g 50%). We then compute the model as usual, but for the remaining frames only. This frame selection strategy can be applied to any BOF approach. In the next section we present extensive classification experiments for the single Gaussian model illustrating the improvement in quality that can be achieved with this simple frame selection strategy.

4. CLASSIFICATION EXPERIMENTS

For lack of reliable ground truth w.r.t perceived audio similarity, we follow the standard procedure in MIR research and evaluate the frame selection strategy in an indirect way, via music genre classification. To prevent collection specific effects we use two completely different genre classification datasets to evaluate the proposed strategy.

4.1. Dataset-1: *1517artists*

The *'1517artists'* genre classification dataset consists of freely available songs from *download.com*[2] and has already been used in [9]. To ensure reasonable track quality approximately the 190 most popular songs (according to the number of total listens) were selected for each genre. Altogether there are 3180 tracks from 1517 different artists distributed over 19 genres in this dataset. It is worth mentioning that this collection has an almost uniform genre distribution and contains tracks from a large number of different artists.

4.2. Dataset-2: *103artists*

The *'103artists'* dataset consists of 2445 commercial songs from 103 artists organized in albums. These 103 artists are divided into 21 genres and the number of songs per genre varies depending on the genre. As the number of artists is rather low — there are at most 9 artists per genre and in average about 5 artists per genre — we expect to observe a high artist related classification effect on this dataset.

[2]Used with permission from CBS Interactive, Inc., Copyright 2008. All rights reserved.

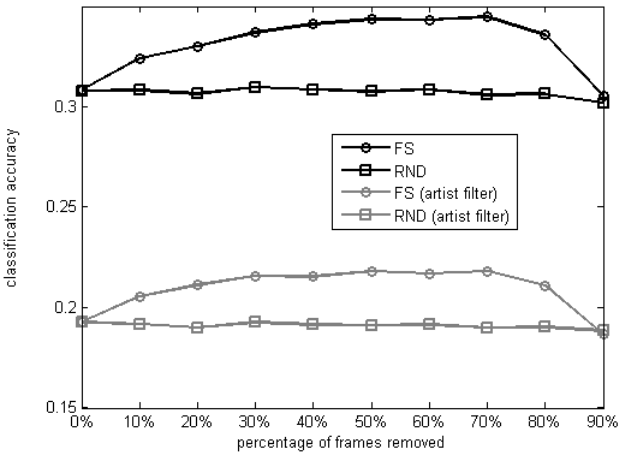

Figure 1: *Classification accuracies of the proposed frame selection strategy and of a random selection baseline-strategy for dataset '1517artists'.*

4.3. Evaluation Procedure

We evaluate the proposed frame selection strategy using the single Gaussian (SG) component of the similarity algorithm proposed by Pohle and Schnitzer [12], which took the first rank in the MIREX 2007 Audio Music Similarity and Retrieval task. First we compute the classification accuracy using all frames (the original algorithm). Then we systematically remove 10%, 20%, ..., 80% and 90% of the frames according to their energy as described in section 3. This strategy is called *Frame Selection* (**FS**). Additionally, we compare this strategy against a random frame selection strategy (**RND**), where we remove the same percentage of frames by choosing the frames to be removed randomly.

For all our classification experiments we report the k-NN classification accuracy. The k-NN classification accuracy is a quality indicator that is related to a query scenario, where the algorithm is asked to return a set of songs that *sound like* the query song. For such a result set one counts the number of *correctly* returned songs. A song in the result set is assumed to be correct w.r.t. our evaluation if the genre is the same as the genre of the query song. Additionally we also report *artist-filtered* k-NN classification accuracies, where only songs are considered that do not belong to the same artist as the query song; this is to quantify artist-related or production effects. In all our evaluations we present results for the 4-NN accuracy, because some of the genres of the *'103artists'* dataset do not contain more than 4 different artists, and measuring the accuracy for more than the top 4 recommendations would return invalid classification results in combination with an artist filter.

4.4. Results and Conclusions

Figures 1 and 2 show the genre classification results for the *'1517 artists'* and *'103artists'* datasets, respectively. A look at the results for the random frame selection strategy shows that the single Gaussian model is remarkably stable. Even with up to 90% of all audio frames removed, the classification accuracy decreases

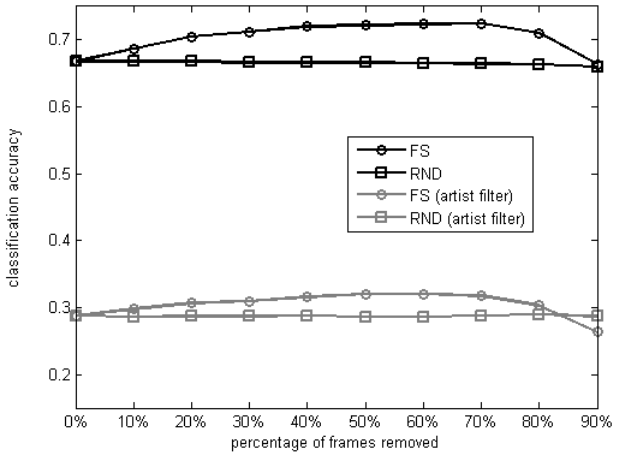

Figure 2: *Classification accuracies of the proposed frame selection strategy and of a random selection baseline-strategy for dataset '103artists'.*

only marginally. Furthermore for both datasets we can observe the same effect, an improvement in classification accuracy when we remove low energy frames using the proposed frame selection strategy. For both collections we can reach a classification optimum if we remove between 50% and 70% of all frames, giving an improvement of about 3 percentage points in classification accuracy on our datasets. For the *'103artists'* dataset the classification results with and without artist filter differ extremely. For the *'1517artists'* dataset the artist effect is not as extreme, which is in line with our expectations.

An interesting finding is that the symmetric KL divergence can be interpreted as the difference of the densities $\hat{p}(\mathbf{x})$ and $\hat{q}(\mathbf{x})$ times the logarithm of the ratio of $\hat{p}(\mathbf{x})$ and $\hat{q}(\mathbf{x})$ (see equation 6). This immediately raises the question if we can define some more intuitive distance measures between two distributions. Based on the nearest neighbour density estimation approach other measures e.g. the euclidean or the squared euclidean distance of the densities could be used to define a measure for the similarity of two distributions.

Another important conclusion of our analysis is that BOF approaches are based on the euclidean distances between spectral frames, which obviously does not correspond to the human similarity perception in some cases. For example a human being would not claim that two song are similar because they share some silent parts, although these silent frames will have a low euclidean distance to each other. It turns out that BOF approaches are rather limited in identifying musical similarities among songs in general and only match songs that have very similar spectra, which is in line with recent findings [4]. While we have identified one case in which the BOF approach does not correspond to the human perception, we will further investigate if one can identify other such discrepancies. This will allow us to better understand what the limits of such spectral low level representations are and what will be important for the development of higher level representations of music.

5. ACKNOWLEDGMENTS

This research was supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF) under grant L511-N15.

6. REFERENCES

[1] Z. Huang, W. Chung, and H. Chen, "A graph model for e-commerce recommender systems," *J. Am. Soc. Inf. Sci. Technol.*, vol. 55, no. 3, pp. 259–274, 2004.

[2] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[3] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *In proceedings IEEE International Conference on Multimedia and Expo (ICME)*, Tokyo, Japan, August 2001.

[4] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high's the sky?," *J. Negative Results Speech Audio Sci.*, vol. 1, no. 1, 2004.

[5] M. Mandel and D. Ellis, "Song-level features and svms for music classification," in *In Proceedings of the 6th International Conference on Music Information Retrieval, ISMIR'05*, London, UK, 11-15th September 2005.

[6] M. Levy and M. Sandler, "Lightweight measures for timbral similarity of musical audio," in *AMCMM '06: Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, Santa Barbara, California, USA, 2006, pp. 27–36.

[7] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[8] M. Hoffman, D. Blei, and P. Cook, "Content-based musical similarity computation using the hierarchical dirichlet process," in *In Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, Philadelphia, USA, Sept. 14-18, 2008, pp. 349–354.

[9] K. Seyerlehner, G. Widmer, and P. Knees, "Frame level audio similarity - a codebook approach," in *In Proceedings of the 11th International Conference on Digital Audio Effects (DAFx'08)*, Espoo, Finland, Sept. 1-4, 2008, pp. 349–354.

[10] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, August 2006.

[11] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification (2nd Edition)*, Wiley-Interscience, November 2000.

[12] T. Pohle and D. Schnitzer, "Striving for an improved audio similarity measure," in *Third Music Information Retrieval Evaluation eXchange, MIREX'07*, Vienna, Austria, 2007.