USING TIMBRE MODELS FOR AUDIO CLASSIFICATION

Franz de Leon
University of Southampton
fadl1d09@ecs.soton.ac.uk

Kirk Martinez
University of Southampton
km@ecs.soton.ac.uk

ABSTRACT

In this submission, audio features that approximate timbre are used for genre classification and music similarity estimation. This abstract describes the feature set, distance computation method, and classifier model used for the submitted algorithms.

1. INTRODUCTION

In our system, audio data are modeled as long-term accumulative distribution of frame-based spectral features. This is also known as the "bag-of-frames" (BOF) approach wherein audio data are treated as a global distribution of frame occurrences. This approach is widely used in MIREX submissions. For MIREX 2010 genre classification and audio similarity estimation task, the BOF approach was used for some of the top performing systems[1][2] .

The features that are extracted from audio files are approximations of timbre. The feature extraction, distance computation and classification algorithms are implemented in MATLAB®.

2. FEATURE EXTRACTION

This section describes the processes involved in feature extraction. More detailed explanation can be found on the cited references.

2.1 Audio Preprocessing

The input signal is assumed to be sampled at 22050 Hz, as specified in MIREX wiki[1]. The audio signal is normalized and preprocessed to remove inaudible parts. The signal is then cut into frames with a window size 512 samples (~23 msec.) and hop size 512 samples.

2.2 Timbre Component

To compute timbre similarities it is necessary to extract *features* or *descriptors* from the audio signal. The extracted features should be able to capture the salient attributes of timbre as the audio retrieval system can only be as good as the features it employs. In this work, we extract the following features: 1) Mel-frequency Cepstral Coefficients to model the spectral envelope [3], 2) spectral contrast to describe the range between tonal and noise-like character [4], 3) sub-band flux to describe the temporal unfolding and shaping of sound spectra [5] and, 4) other spectral features commonly used in the literature.

2.2.1 Mel-frequency Cepstral Coefficients

The normalized audio signals signal is divided into frames with a window size and hop size of 512 samples (~23 msec.). The length of the segment ensures that the segmented signal is pseudo-stationary while the hop size keeps the continuity of the segments. Next, a window function (e.g. Hanning window) is applied to each segment. This is necessary to reduce spectral leakage. The following steps are then performed to each segment:

1. Calculate the power spectrum using FFT.
2. Transform the power spectrum to Mel-scale using a filter bank consisting of triangular filters.
3. Get the sum of the frequency contents of each band.
4. Take the logarithm of each sum.
5. Compute the discrete cosine transform (DCT) of the logarithms.

2.2.2 Spectral Contrast

The spectral contrast algorithm published in [4] is very similar to the MFCC algorithm. In our implementation, we still use the Mel-scale filters instead of octave-based filterbank to optimize the system since the output of this block is used for MFCC and spectral contrast. We use DCT to decorrelate the coefficients.

The raw spectral contrast features estimate the strength of spectral peaks, valleys and their differences in each sub-band. The strength of the peaks and valleys are estimated by the average value in the small neighborhood around maximum and minimum value respectively, instead of the exact maximum and minimum value themselves.

2.2.3 Sub-band Flux

A set of features called sub-band flux has been proposed by Alluri and Toivianen to represent the fluctuation of frequency content in octave-scaled bands of the spectrum [5].

[1] http://www.music-ir.org/mirex/wiki/2011:Audio_Music_Similarity_and_Retrieval

The division into sub-bands was obtained using a 10-channel filterbank of octave-scaled fourth-order butterworth filters. The sub-bands are defined as follows: {0 ~ 25Hz, 25 ~ 50Hz, 50 ~ 100Hz, 100 ~ 200Hz, 200 ~ 400Hz, 400 ~ 800Hz, 800 ~ 1600Hz, 1600 ~ 3200Hz, 3200 ~ 6400Hz, 6400 ~ 11025Hz} where the sample rate is 22050Hz. For each of the channels the spectral flux was computed using Euclidean distance between successive magnitude spectra.

2.2.4 Spectral Distribution Descriptors

In order to enhance the timbre model, a number of spectral features are also derived. These features are based on the short time Fourier transform (STFT) and are calculated for every frame of sound. The use of spectral descriptors were analyzed by [6] to describe *timbral texture*. The following features are also derived in our system:

1. *Spectral centroid* – defined as the center of gravity of the magnitude spectrum
2. *Spectral spread* – defines the dispersion or spread of the magnitude spectrum
3. *Spectral skewness* – describes the symmetry of the magnitude spectrum
4. *Spectral kurtosis* – measures how “Gaussian” the magnitude spectrum looks like
5. *Spectral flatness* – indicates whether the magnitude spectrum is flat or “spiky”
6. *Spectral flux* – measures the amount of spectral leakage
7. *Spectral rolloff* – defined as the frequency below which 85% of the magnitude distribution is concentrated
8. *Spectral brightness* – measures the amount of energy above the cut-off frequency of 1500 Hz
9. *Spectral entropy* – indicates the magnitude spectrum has predominant peaks or not

2.2.5 Timbre Modelling

We then model the distribution of the MFCCs for the audio file using a Gaussian mixture model (GMM). In this work, we use a single Gaussian represented by its mean μ and covariance matrix Σ [7]. The feature vectors are also mapped to Euclidean space using a modified Fastmap algorithm [8].

3. AUDIO CLASSIFICATION

Audio similarity estimation is done in two stages. In the first stage, the nearest neighbors to the query item are returned based on the Euclidean distances of the mapped vectors. In the second stage, the nearest neighbors are ranked according to the aggregate of feature distances.

The feature distances are calculated separately. Before they are combined, each distance component is normalized by removing the mean and dividing by the standard deviation of all the distances. Symmetry is obtained by summing up the distances in both directions for each pair of tracks [9].

Distances between timbres are computed by comparing the GMM models. We use symmetric Kullback-Leibler (SKL) distance between two models [7]. The SKL distances are transformed into metric by getting the root of the logarithm for each distance measure.

A direct approach to combine the feature distances is to compute a weighted sum of the individual distances. Each distance component is normalized by removing the mean and dividing by the standard deviation of all the distances. The system is then optimized by determining the appropriate weights for each distance component. Finally, the genre of the nearest track based on weighted distances of the features is used to label the untagged track (1-NN).

4. RESULTS

This submission is an updated version of the algorithm submitted to MIREX 2011 Train/Test task. From the MIREX 2011 data, the classification accuracy obtained was 58.51%.

5. ACKNOWLEDGMENTS

Mr. Franz de Leon is supported by the Engineering Research and Development for Technology Faculty Development Program of the University of the Philippines, and DOST.

6. REFERENCES

[1] K. Seyerlehner, M. Schedl, T. Pohle, and P. Knees, “Using Block-Level Features for Genre Classification , Tag Classification and Music Similarity Estimation,” in *Submission to Audio Music Similarity and Retrieval Task of MIREX 2010*, 2010.

[2] K. Seyerlehner, G. Widmer, M. Schedl, and P. Knees, “Automatic Music Tag Classification Based on Block-Level,” in *Proceedings of Sound and Music Computing 2010*, 2010.

[3] B. Logan, “Mel Frequency Cepstral Coefficients for Music Modeling,” in *Proceedings of International Society for Music Information Retrieval Conference*, 2000.

[4] D.-N. Jiang, L. Lu, H. Zhang, J. Tao, and L. Cui, “Music type classification by spectral contrast feature,” in *Proceedings. IEEE International*

Conference on Multimedia and Expo, 2002, pp. 113–116.

[5] V. (University O. J. Alluri and P. (University O. J. Toiviainen, “Exploring Perceptual and Acoustical Correlates of Polyphonic Timbre,” *Music Perception*, vol. 27, no. 3, pp. 223–242, 2009.

[6] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[7] M. I. Mandel and D. P. W. Ellis, “Song-level Features and Support Vector Machines for Music Classification,” in *Submission to MIREX 2005*, 2005, pp. 594–599.

[8] D. Schnitzer, A. Flexer, and G. Widmer, “A Filter-and-Refine Method for Fast Similarity Search in Millions of Tracks,” in *ISMIR 2009*, 2009, no. April, pp. 537–542.

[9] E. Pampalk, “Audio-Based Music Similarity and Retrieval : Combining a Spectral Similarity Model with Information Extracted from Fluctuation Patterns,” in *Submission to MIREX 2006*, 2006.