A SPECTRAL BANDWISE FEATURE-BASED SYSTEM FOR THE MIREX 2016 TRAIN/TEST TASK

Juliano Henrique Foleiss
Universidade Tecnológica Federal do Paraná
julianofoleiss@utfpr.edu.br

Tiago Fernandes Tavares
Universidade Estadual de Campinas
tavares@dca.fee.unicamp.br

ABSTRACT

In this work we propose to separately calculate spectral low-level features in each frequency band, as it is commonly done in the problem of beat tracking and tempo estimation [6]. We based this assumption in the same auditory models that inspired the use of Mel-Frequency Cepstral Coefficients (MFCCs) [2] or energy through a filter bank [1] for audio genre classification. They rely on a model for the cochlea in which similar regions of the inner ear are stimulated by similar frequencies, and are processed independently. Both the MFCCs and the energy through filterbank approaches only generate an energy spectrum. In our approach, we expand this idea to incorporate other perceptually-inspired features.

1. INTRODUCTION

Automatic Musical Genre Classification is the problem of determining the musical genre (e.g., rock, jazz, classical...) of an audio track. One possible solution for this problem is to estimate a set of features from each audio track and yielding them to machine learning algorithms [4]. This process does not depend on human interference, but is only as effective as the feature set is correlated to the desired musical genres.

Tzanetakis and Cook [4] proposed a widely used feature estimation process. First, digital audio is broken into short-time (around 23ms) frames. Descriptive features are calculated from each frame, generating feature tracks. After that, statistics are calculated from each feature track in 1s-long frames, called *texture windows*. Last, the system estimates statistics from the texture window statistics, generating a vector representation related to the audio track.

The assumption in this context is that frame-wise features are correlated to perceptual audio characteristics. Therefore, audio tracks that sound similar tend to have more similar vector representations. This property allows audio files to be classified using their vector representation as basis.

In our system, we propose to separately calculate low-level features in each frequency band, as it is commonly done in the problem of beat tracking and tempo estimation [6]. We based this assumption in the same auditory models that inspired the use of Mel-Frequency Cepstral Coefficients (MFCCs) [2] or energy through a filter bank [1] for audio genre classification. They rely on a model for the cochlea in which similar regions of the inner ear are stimulated by similar frequencies, and are processed independently.

Both the MFCCs and the energy through filterbank approaches only generate an energy spectrum. In our approach, we expand this idea to incorporate other perceptually-inspired features from the literature. By calculating spectral features over different frequency bands we expect to get a richer audio description, thus being able to distinguish better among classes composed of similar timbres.

2. OUR SYSTEM

Our system is based on a traditional machine learning feature extraction ⇒ model training ⇒ model testing pipeline for audio signal classification. The following sections present our approach in detail.

2.1 Feature Extraction

In the feature extraction phase all audio files are transformed into the frequency domain through a 1024-sample STFT with 50% overlap. In our approach, the spectrum is divided into 50 mel-spaced bands, and the following spectral features are extracted for each band:

- Flatness
- Roll-off
- Centroid
- Flux
- Energy
- Low Energy

Other non-bandwise features were also used:

- First 20 MFCC coefficients
- Time-domain zero-crossings

Statistics such as expectation (mean) and variances are computed to aggregate all time frames into a smaller set of values representing each of features for every mel-band.

Once the features are computed for every file in the dataset, our system uses a fairly standard approach to machine learning. A support vector machine (SVM) [5] is trained to model the feature space. Grid search is used to tune the hyper-parameters of the SVM using the training data. ANOVA feature selection is used to lower the dimensionality of the vector representation.

Once the SVM is trained, a class prediction of all files in the test set is obtained thru classical SVM procedures.

3. IMPLEMENTATION

Our system was implemented in Python 2.7 using standard scientific computing libraries such as `numpy`, `scipy`, and `multiprocess`. For feature extraction we used our own MIR framework, called `pymir3` [3].

4. REFERENCES

[1] Chang-Hsing Lee, Jau-Ling Shih, Kun-Ming Yu, and Hwai-San Lin. Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. *Trans. Multi.*, 11(4):670–682, June 2009.

[2] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*, 2000.

[3] Juliano H. Foleiss Tiago F. Tavares. pymir3 – Python Music Information Retrieval Reproducible Research Framework. https://github.com/pymir3/pymir3, 2016.

[4] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, Jul 2002.

[5] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[6] Jose R. Zapata and Emilia Gómez. Comparative evaluation and combination of audio tempo estimation approaches. 2011.