MEL-BAND FEATURES FOR DCASE 2016 ACOUSTIC SCENE CLASSIFICATION TASK

Juliano Henrique Foleiss

Universidade Tecnologica Federal do Parana
Campo Mourao, PR - Brasil
julianofoleiss@utfpr.edu.br

Tiago Fernandes Tavares

Universidade Estadual de Campinas
Campinas, SP - Brasil
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 [1]. 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 [3] 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.

Index Terms— Spectral Features, Timbral Features, Acoustic Scene Detection

1. INTRODUCTION

Acoustic scene classification is an important audio signal processing application that has several potential uses in scenarios such as security, surveillance and context-aware consumer applications. In a sense this task is an instance of a broader audio classification problem, in which a particular sound signal is associated with a semantic label. Usually for this problem a set of features is calculated for each instance in a given dataset. It is expected that features calculated from semantically-related instances yield similar values, as long as the features chosen to represent the audio signal are correlated with the classes they represent.

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 [1]. 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 [3] 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), variance, first and second derivatives 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` [6].

4. EXPERIMENTS AND RESULTS

For the 2016 DCASE Challenge, an extensive audio dataset is being used for evaluating acoustic scene classification system [7]. This dataset is comprised of 15 acoustic scenes:

- Bus
- Cafe / restaurant
- Car
- City center
- Forest path
- Grocery store
- Home
- Lakeside beach
- Library
- Metro station
- Office
- Residential area
- Train
- Tram
- Urban park

The DCASE 2016 Acoustic scenes dataset is partitioned into two subsets: the development and the evaluation dataset. The development dataset consists in 78 segments of 30 seconds of audio for each one of the acoustic scenes. The evaluation dataset has 26 segments of 30 seconds of audio for each acoustic scene. The development dataset was released as soon as the challenge was announced. The evaluation dataset was released shortly before submission, although the ground truth will only be released afterwards.

For comparison purposes, the development dataset was divided into a cross-validation setup. Four folds were defined and released to the public. Table 1 summarizes the results for each fold:

Fold	Accuracy	F1-Score
1	71	67
2	68	67
3	77	74
4	71	66
Average	71.75	68.5

Table 1: Results for Development Dataset (in %)

Table 2 presents the average results for each acoustic scene across the 4 folds. When comparing to the baseline system provided, we achieve better results for 7 scenes: beach, car, city center, grocery store, library, park and train.

We have also calculated class predictions for the evaluation dataset. For this task, we used the entire development dataset to train and tune our SVM classifier.

Scene	Precision	F1-Score
beach	89.25	72.25
bus	78.00	84.75
cafe/restaurant	53.50	42.25
car	90.00	81.75
city_center	87.75	91.75
forest_path	72.00	81.25
grocery_store	71.25	83.25
home	68.00	71.25
library	69.00	69.75
metro_station	54.50	57.75
office	72.25	65.50
park	55.00	44.00
residential_area	68.25	68.50
train	79.75	49.25
tram	60.25	68.25

Table 2: Results per class across all 4 folds (in %)

5. REFERENCES

[1] J. R. Zapata and E. Gómez, "Comparative evaluation and combination of audio tempo estimation approaches," 2011. [Online]. Available: http://mtg.upf.edu/system/files/publications/Audio_tempo_comparison_Zapata_Gomez.pdf

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

[3] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features," *Trans. Multi.*, vol. 11, no. 4, pp. 670–682, June 2009. [Online]. Available: http://dx.doi.org/10.1109/TMM.2009.2017635

[4] 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.

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

[6] T. F. Tavares and J. H. Foleiss, "pymir3 – Python Music Information Retrieval Reproducible Research Framework," https://github.com/pymir3/pymir3, 2016.

[7] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, 2016.