SOUND SCENE IDENTIFICATION BASED ON MFCC, BINAURAL FEATURES AND A SUPPORT VECTOR MACHINE CLASSIFIER

Waldo Nogueira, Gerard Roma and Perfecto Herrera

Music Technology Group
Universitat Pompeu Fabra
Roc de Boronat 138,
08018, Barcelona
waldo.nogueira@upf.edu

ABSTRACT

This submission to the sub-task scene classification of the IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events is based on a feature extraction module in three dimensions (spectral, temporal and spatial). Spectral features are based on Mel frequency cepstrums coefficients, temporal features are based on an event density extractor and the spatial features are based on the extraction of inter-aural differences (level and temporal) and the coherence between the two channels of stereo recordings. After feature selection, the features are used in conjunction with a support-vector-machine for the classification of the sound scenes. In this short paper the impact of different features is analyzed.

***Index Terms*—** mfcc, support vector machine, sound scene, machine learning, binaural

1. INTRODUCTION

This short paper describes a submission for the scene analysis's challenge. The system designed has a feature selection unit composed by different feature extractors in three dimension (temporal, spectral and spatial). The different features are combined and the most statistically relevant features are selected to train a support-vector-machine (SVM) classifier. Figure **??** gives an overview of the whole system. The following subsections give more details on each of processing block of the diagram.

2. METHODS

2.1. Spectral Features

The spectral features are based on the commonly used mfcc and we used the rastamat implementation [4] .The stereo soundtracks formed by the left and right channels (x_1 and x_2) sampled at 44.1 kHz are mixed by taking the average of each sample between the left and right channel and processed using a short-time Fourier magnitude spectrum calculated over a 20-ms window every 10 ms. The spectrum of each window is converted into to the Mel frequency scale, and the log of the Mel bands is computed. Finally, discrete cosine transform (DCT) is applied to the Mel bands to decorrelate the data. For each audio file we took 13 mfcc.

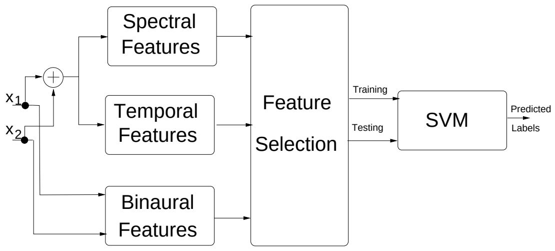

Figure 1: Block diagram of the scene classification system.

Two sets of mfcc's were extracted, the first subset takes 40 Mel bands for the frequency range comprised between 0Hz and 900 Hz. The mean and standard deviation of the 13 cepstral coefficients were extracted for each sound file. Additionally the mean and standard deviation of 12 delta coefficients, where the delta was computed in the cepstral dimension (and not in the common temporal dimension) were extracted. We will term these features as ledtas instead of deltas following the name convention spectrum-cepstrum.

The second subset decomposes each sound file into 40 Mel bands for the frequency range comprised between 900Hz until 10 kHz. Next, only the mean of 13 cepstral coefficients is extracted.

In total 13+13+12+12+13= spectral features were extracted.

2.2. Temporal Features

Two types of temporal features are extracted. On the one hand we wanted to describe the temporal evolution of the mfcc's. To achieve this new mfcc's consisting of 13 cepstral coefficients out of 40 Mel bands on the frequency range from 0 Hz until 10 kHz were calculated. Next the amplitude variation of the DCT coefficients at 4 different modulation rate ranges (0-2Hz, 2-5Hz, 5-10Hz and 10-20Hz). The 4 modulation rates were estimated for each of the 13 mfcc coefficients forming a total of 13x4 features for each sound file. An additional feature extracting the total standard deviation of the 13 mfcc was computed.

Second we wanted to estimate the amount of events happening in each sound file. The hypothesis is that some acoustic scenes tend to contain more acoustic events than others. For this we used the event density estimator implemented in the Mirtoolbox [3].

In total 13x4(mfcc 4-band modulation) + 1(mfcc global sd) +

1(event density) = 54 temporal features were extracted.

2.3. Spatial Features

Because the audio data set provided has two channel stereo information, and motivated by the fact that some scene classes might differ from each other by their binaural characteristics, we decided to extract binaural features form the scene. As an example, the class train might be dominated by the sound emitted by the train where the microphone was placed whereas the class tube station might be dominated by the sound of trains passing by. We expect the sound from both classes (train and tube station) to have similar temporal and spectral features but the balances between the left and right chanels might be different for each class.

The spatial feature extractor uses a model of binaural hearing [2] to estimate intereaural time differences (ITD) and interaural level difference (ILD). Additionally the interaural coherence (IC) is extracted.

The ITD and IC are estimated from the normalized cross-correlation function. Given x_1 and x_2 for a specific center frequency f_c, at the index of each sample n, a running normalized cross-correlation function is computed according to:

$$\gamma(n,m) = \frac{a_{12}(n,m)}{\sqrt{a_{11}(n,m)a_{22}(n,m)}}, \tag{1}$$

where

$$\begin{aligned} a_{12}(n,m) = \alpha x_1(n - max\{m,0\})x_2(n - max\{-m,0\}) \\ +(1-\alpha)a_{12}(n-1,m), \end{aligned} \tag{2}$$

$$\begin{aligned} a_{11}(n,m) = \alpha x_1(n - max\{m,0\})x_1(n - max\{m,0\}) \\ +(1-\alpha)a_{11}(n-1,m), \end{aligned} \tag{3}$$

$$\begin{aligned} a_{22}(n,m) = \alpha x_2(n - max\{-m,0\})x_2(n - max\{-m,0\}) \\ +(1-\alpha)a_{22}(n-1,m), \end{aligned} \tag{4}$$

and α [0,1] determines the time constant of the exponentially decaying estimation window

$$T = \frac{1}{\alpha f_s}, \tag{5}$$

where f_s is the sampling frequency, $\gamma(n,m)$. The ITD (in samples) is estimated as follows:

$$\tau(n) = \underset{x}{argmax}\{\gamma(n,m)\}. \tag{6}$$

$$c_{12} = \underset{m}{max}\{\gamma(n,m)\}. \tag{7}$$

This estimate describes the coherence of the left and right ear input signals. In principle, it has a range [0,1], where 1 occurs for perfectly coherent x_1 and x_2.

The ILD is computed as

$$\Delta L = 10log_{10}(\frac{L_2(n,\tau(n)}{L_1(n,\tau(n))}), \tag{8}$$

where

$$L_1(n,m) = \alpha x_1^2(n - maxm, 0 + (1-\alpha)L_1(n-1,m)), \tag{9}$$

$$L_2(n,m) = \alpha x_2^2(n - max{-}m, 0) + (1-\alpha)L_2(n-1,m), \tag{10}$$

Finally the cue triplets $\Delta L(n), \tau(n), c_{12}(n)$ are obtained. Following different considerations the same method as specified in **??** was used to select ITD and ILDs cues.

For each sound track and for 3 different frequencies (250Hz, 500Hz and 1000Hz) we the mean and the standard deviation for $\Delta L(n), \tau(n), c_{12}(n)$. This gives a total of 3 frequencies x 3 triplets x 2 (mean and std) = 18 binaural features.

2.4. Feature Selection

All the features were combined forming a feature array of 63 + 54 + 18 = 135 dimensions. Because of the relative large number of features selected and to try to avoid over-fit we used feature selection. We selected a subset of 35 features that delivered the highest Fisher score [6].

2.5. Support Vector Machines (SVMs)

For classification we used a state-of-the-art supervised learning method based on SVM. As it is well known, data items are projected into a high dimensional feature space, and the SVM finds a separating hyperplane in that space that maximizes the margin between sets of positive and negative training example. Instead of working in the high-dimensional space directly, the SVM requires only the matrix of inner products between all training points in that space, also known as the kernel or gram matrix. In our method we used a linear distance between the examples to create the gram matrix $K(f,g)$

$$K(f,g) = e^{-\gamma D(f,g)}. \tag{11}$$

We use the so-called slack SVM that allows a trade-off between imperfect separation of training examples and smoothness of the classification boundary, controlled by a constant C that we vary in the range $10^1, 10^2, ..., 10^{10}$. Both tunable parameters γ and C were chosen to maximize the classification accuracy over a held-out set of validation data. After training an independent SVM model for each concept, we apply the classifiers to summary features derived from the audio files.

We evaluated the classifiers by calculating the confusion matrix and obtaining the accuracy of the predicted classes.

3. EVALUATION

We evaluated our approaches using a fivefold cross validation on the labeled collection of 100 audio files provided by the IEEE ASP challenge organizers. At each fold, SVM classifiers for each concept were trained on 80% of the data, tuned on 20% and then tested on the remaining 20%.

Next table presents the confusion matrix obtained with system described in previous sections. The rows and columns correspond in order with the following classes: 1 bus, 2 busystreet, 3 office, 4 openairmarket, 5 park, 6 quietstreet, 7 restaurant, 8 supermarket, 9 tube and 10 tubestation. The mean accuracy of our pilot experiment after 5 folds was 69.47%.

5	0	0	0	0	0	0	0	0	0
0	10	0	0	0	0	0	0	0	0
0	0	8	0	1	0	0	0	1	0
1	0	0	7	0	0	0	1	1	0
1	0	0	0	7	2	0	0	0	0
0	1	0	1	2	6	0	0	0	0
0	0	0	0	0	0	8	2	0	0
0	1	0	1	0	1	1	6	0	1
0	2	1	0	0	0	2	0	1	4
0	0	0	0	0	0	1	0	1	8

4. CONCLUSIONS

In this short paper we have described a method to classify sound scenes based on spectral, temporal and binaural features. The relative large amount of features has been reduced by selecting some of the using the Fisher's method. This features have been used to train a support-vector-machine-classifiers.

5. REFERENCES

[1] K. Lee and D. P. Ellis, and E. Roberts, "Audio-Based Semantic Concept Classification for Consumer Video," *IEEE Trans. Audio, Speech and Language Process.*, vol. 18, no. 6, pp. 1406–1416, Aug. 2010.

[2] C. Faller, J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *Journal of the Acoustical Society of America.*, vol 16, no. 5, pp. 3075–2089, November 2004

[3] O. Lartillot, P. Toiviainen, T. Eerola, "A Matlab Toolbox for Music Information Retrieval", in C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (Eds.), *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, 2008.

[4] `http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/mfccs.html`.

[5] Chang, Chih-Chung and Lin, Chih-Jen, "LIBSVM: A library for support vector machines", *ACM Transactions on Intelligent Systems and Technology*, vol. 2, issue. 3, 2011, pp. 27:1–27:27, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[6] Quanquan Gu, Zhenhui Li, Jiawei Han, "Generalized Fisher Schore for Feature Selection", CoRR, abs1202.3725,2012 `http://arxiv.org/abs/1202.3725`,2012.