SOUND SCENE IDENTIFICATION BASED ON MONAURAL AND BINAURAL FEATURES

Waldo Nogueira

Medical University Hannover and Cluster of Excellence Hearing4all
Karl-Wiechert Allee 3, 30625, Hannover
Nogueiravazquez.waldo@mh-hannover.de

ABSTRACT

This submission to the sub-task acoustic scene classification of the IEEE DCASE 2016 Challenge: Acoustic scene classification is based on a feature extraction module based on the concatenation of monaural and binaural features. Monaural features are based on Mel Frequency Cepstrums summarized using recurrence quantification analysis. Binaural features are based on the extraction of inter-aural differences (level and time) and the coherence between the two channel stereo recordings. These features are used in conjunction with a support vector-machine for the classification of the acoustic 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

A typical auditory scene contains multiple sources that change their locations. Moreover different acoustic scenes differ on the way the sound sources propagate. Some sources may reflect in walls from a room and some others are propagated in almost free field in an open space. The spectral and temporal monaural characteristics of the sound sources as well as the binaural perception of these sound sources by the human auditory system can be used to recognize a sound environment [1]. Therefore an automatic classification system may be able to use both, monaural and binaural features to improve its accuracy.

This short paper describes an acoustic scene classification system that uses both monaural and binaural features and proposes to concatenate them to improve the classification accuracy.

The selection of appropriate acoustic features is essential for successful classification. Typical monaural features for acoustic scene classification are the well known mel-frequency cepstral coefficients (MFCCs) and derived amplitude modulation [1], [2]. Moreover, it is known that binaural cues contribute to auditory scene analysis [1]. It has been shown that monaural features are not much affected by spatial configuration of sound sources, and can therefore complement binaural features. More concrete, inter-aural time difference (ITD), inter-aural level difference (ILD) and inter-aural coherence (IC) cues are proposed as features in the classification system [3].

A common practice in audio classification tasks is the integration of frame-level features over some period of time that can be used as input to state-of-the art algorithms for classification. Typically, the frame level features are summarized using the mean and the standard deviation. This approach reduces to much information about the temporal evolution and distribution of the features. For this reason Recurrence Quantification Analysis (RQA) was proposed to improve the temporal representation of frame level features leading to improvements in acoustic scene classification accuracy [4].

The system proposed in this short paper uses RQA and binaural features. These features are combined to train a support vector-machine (SVM) classifier. Figure 1 gives an overview of the whole system. The following subsections give more details on each of the processing blocks of the diagram.

2. METHODS

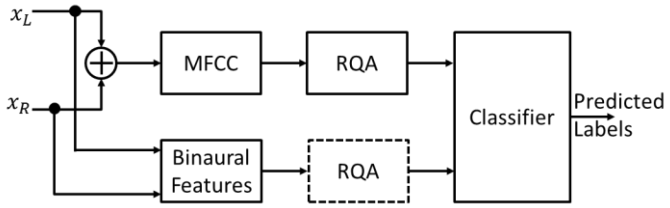

Figure 1: Block diagram of the scene classification system

2.1. Monaural Features

We extract MFCCs from the audio recordings. Our implementation uses the rastamat library [5].The stereo soundtracks formed by the left and right channels (x_L and x_R) sampled at 44.1 kHz are mixed by averaging each sample between the left and right channel and processed using a short-time Fourier magnitude spectrum calculated over a 40 ms window with a hop-size of 20 ms. The mean and standard deviation of the frame level features is computed. The mean and the standard deviation of the delta MFCC features are concatenated.

2.1.1. Recurrence Quantification Analysis

RQA is a set of techniques that have been used to classify auditory scenes and sound events [4]. The basic idea is to quantify patterns that emerge in recurrence plots. A time series, such as

MFCC features, are used to compute a distance matrix of the series. The distance matrix can be obtained using the cosine distance. Next the distance matrix is thresholded to a certain radius *r*. The radius represents the maximum distance of two observations of the series that will still be considered as belonging to the same state of the system. The resulting matrix contains ones for each pair of frame indices that are close together, and zeros for the rest. The main intuition is that diagonal lines represent periodicities in the signal, i.e. repeated (or quasi-repeated, depending on the chosen radius) sequences of frames, while vertical lines represent stationarities, i.e. the system remains in the same state. The main diagonal, or Line Of Identity (LOI) is obviously not counted. From this idea, several metrics have been developed that quantify the amount and length of lines of contiguous points in the matrix.

Because the time series obtained from sound scenes are relatively long, a windowed version was proposed [4]. This consists in computing the recurrence plots from overlapping windows of fix size. This makes it possible to analyze the temporal evolution of the features. The window length was set to 40 MFCC frames, which represent 400 ms of audio. The radius parameter was fixed to 0.03. Other parameters are the minimum line lengths for considering diagonal and vertical lines which were set to a minimum of 2 points.

2.2. Binaural Features

Binaural features may be useful to distinguish between different sound environments. As an example, the class train contains sound recordings inside the train and therefore the main sound belongs to the train itself. On the other hand the class metro station is composed by sounds of trains passing by. Although the temporal and spectral characteristics of the train sounds are similar on both classes, the balances between the left and right channels may be different.

The spatial feature extractor uses a model of binaural hearing [3] to estimate inter-aural time (ITD), the inter-aural level differences (ILD) and the inter-aural coherence (IC). The next paragraph provides a short description on how these features are computed, a more extensive description can be found in [3].

A gammatone filterbank is applied on the left x_L and right x_R ear signals to simulate the frequency analysis performed by the basilar membrane. The output of this analysis results in critical bands that are inserted in a model of neural transduction. The model adds internal noise to the critical bands to simulate the limited accuracy of the auditory system. After some processing the resulting nerve firing densities at the corresponding left and right ear critical bands are denoted y_L and y_R.

The ITD and IC are estimated from the normalized cross-correlation function. Given y_L and y_R for a specific center frequency f_c, at each time index n, the normalized cross-correlation function is computed as follows:

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

where

$$a_{12}(n,m) = \alpha y_L(n - max\{m,0\})y_R(n - \max\{-m,0\}) + (1-\alpha)a_{12}(n-1,m),$$

$$a_{11}(n,m) = \alpha y_L(n - max\{m,0\})y_L(n - \max\{-m,0\}) + (1-\alpha)a_{11}(n-1,m),$$

$$a_{22}(n,m) = \alpha y_R(n - max\{m,0\})y_R(n - \max\{-m,0\}) + (1-\alpha)a_{22}(n-1,m),$$

and $\alpha \in [0,1]$ determines the time constant of the exponentially decaying estimation window:

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

where f_s is the sampling frequency. The ITD (in samples) is estimated as follows:

$$\tau(n) = \underset{m}{\operatorname{argmax}}\{\beta(n,m)\},$$
$$c_{12} = max\{\beta(n,m)\}.$$

This estimate describes the coherence between the left and right ear input signals. In principle, it has a range [0,1], where 1 occurs when y_L and y_R are perfectly coherent.

The ILD is computed as

$$\Delta L = 10log10\left(\frac{L_2(n,\tau(n))}{L_1(n,\tau(n))}\right),$$

where,

$$L_1(n,m) = \alpha y_L^2\big(n - \max(m,0) + (1-\alpha)L_1(n-1,m)\big),$$
$$L_2(n,m) = \alpha y_R^2\big(n - \max(-m,0) + (1-\alpha)L_2(n-1,m)\big),$$

Finally the cue triplets $\Delta L(n), \tau(n), c_{12}(n)$ are obtained. The ILDs and ITDs where only estimated when the coherence between y_L and y_R exceeded a threshold set to 0.4.

For each sound track we extracted the triples (ITD, ILD and IC) at 250 Hz and at 2000 Hz. The mean and the standard deviation were then computed and concatenated to the monaural features.

As an example we show the estimated ITD, ILD and IC features over time at frequencies 250 Hz and 2000 Hz for a single recording of the class beach and the class office (Figure 2).

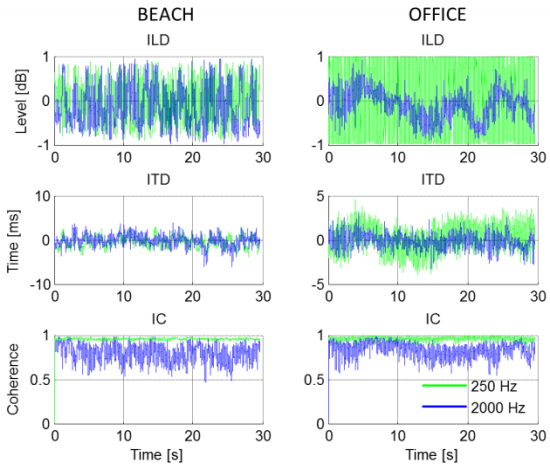

Figure 2: Example of the estimated ILD (Top), ITD (center) and IC (bottom) for the class Beach (left) and the class office (right) over time.

From Figure 2 it can be observed that ILDs at a higher frequency (2000 Hz) allow for an easier discrimination between the two classes than at low frequencies. For the ITDs however, the low frequencies seem to provide more discriminative cues. For the IC feature the differences between classes are quite small at both frequencies.

Figure 3 presents the mean and the standard deviation over time for the ILD, ITD and IC features estimated at frequencies 250 Hz and 2000 Hz for each class of the DCASE challenge 2016 dataset.

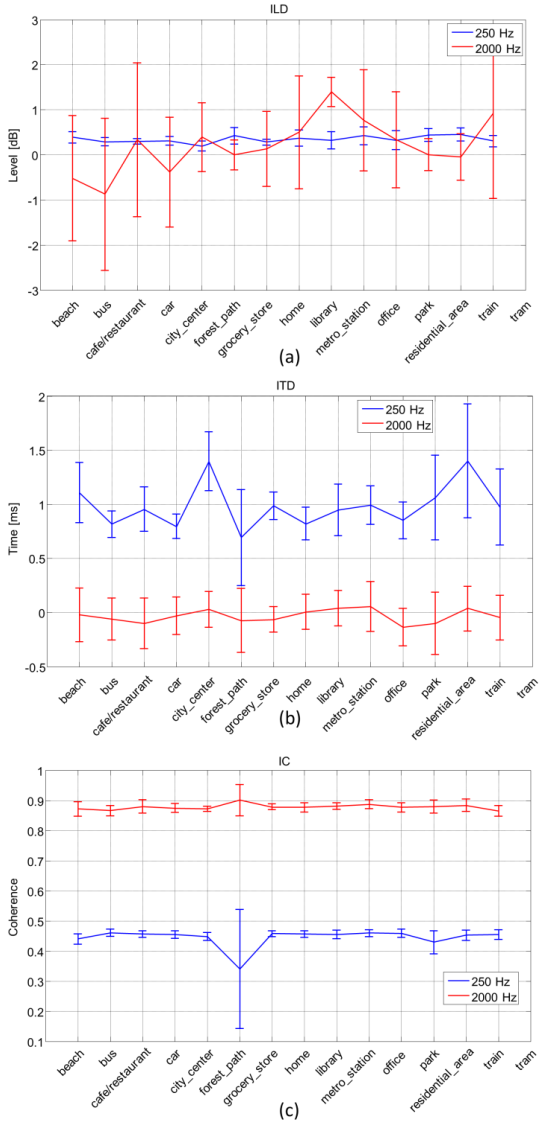

Figure 3: Mean and standard deviation for the ILD, ITD, and IC for each class of the DCASE 2016 challenge dataset.

2.2.1. Recurrence Quantification Analysis on Binaural Features

In the same manners as using monaural features, the binaural triplets (ILD, ITD and IC) at different frequencies can be stacked to form an array of binaural features in each time instant. RQA can then be used to analyze the time series of binaural features. RQA for binaural triplets (ILD, ITD and IC) at two frequencies 250 Hz and 200 Hz were estimated. The binaural triplets were averaged over 0.43 seconds before applying RQA. The same RQA parameters as in the monaural case were used. Figure 4 presents the recurrent plots based on binaural features for two environmental sounds (Urban Park and Restaurant).

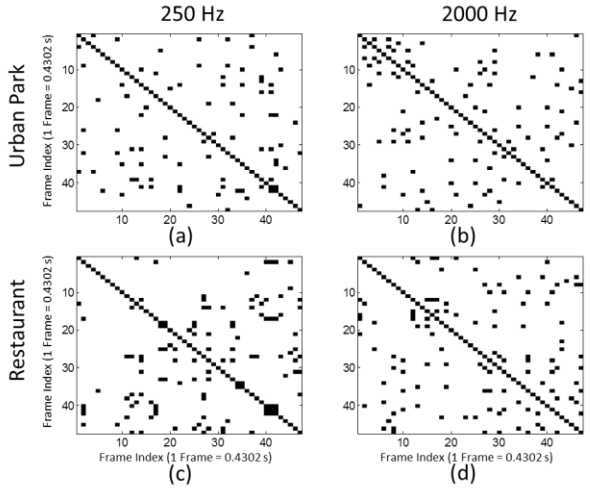

Figure 4: Recurrence plots of two classes (Urban Park and Restaurant) based on binaural features (ILD, ITD and IC) at 250 Hz and 2000 Hz.

2.3. Classifier

2.3.1. Support Vector Machine (SVM)

For classification we used a supervised learning method based on a standard SVM using an *RBF* kernel. In our method we used a linear distance between the examples to create the gram matrix

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

We used the so called slack SVM that allows a trade-off between imperfect separation 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 summarize features derived from the audio files. The SVM was implemented using the libSVM library [6].

2.4. Databases

The DCASE 2016 [7] datasets for environmental sound detection was used. The DCASE 2106 dataset was recorded using a Soundman OKM II Klassik/studio A3, electret binaural microphone and a Roland Edirol R-09 wave recorder using 44.1 kHz sampling rate and 24 bit resolution. The microphones are specifically made to look like headphones, being worn in the ears. As an effect of this, the recorded audio is very similar to the sound that reaches the human auditory system of the person wearing the equipment. Given this recording system some acoustic scenes may differ from each other based on their binaural characteristics.

2.5. Baseline System

The baseline system consists of a classical mel frequency cepstral coefficient (MFCC) and a support vector machine (SVM) based classifier. MFCCs were calculated for all audio using 40 ms frames with Hamming window and 50% overlap and 40 mel bands. The number of cepstral coefficients was set to 40 coefficients including the 0th order coefficient. Delta and acceleration coefficients were also calculated using a window length of 9 frames, resulting in a frame-based feature vector of dimension 160. Each acoustic scene was used to train a SVM as described in section 2.3.

3. RESULTS

We evaluated our approaches (our baseline and the new developments based on monaural, monaural + RQA, RQA and binaural features and RQA based on monaural and binaural features) using the development and the evaluation dataset provided by the IEEE DCASE 2016 challenge organizers. The development dataset consists of 78 segments of 30 seconds of audio for each acoustic scene. The development dataset was divided into a fur fold cross-validation setup. At each fold, SVM classifiers for each concept were trained on 80% of the development data, tuned on 20% and then tested on the 20% of development dataset not used for training. Classification performance was measured using accuracy: the number of correctly classified segments among the total number of test segments.

Figure 5 presents the accuracy for each class using monaural features (40 mean MFCC + 40 std MFCC + 40 mean delta MFCC +40 std delta MFCC) + monaural RQA features (11 features) + binaural features (mean and std for each triplet at two 250 Hz and 2000 Hz resulting in 12 features) + RQA binaural features (11 RQA for 250 Hz and 11 RQA for 2000 Hz). The analysis resulted in a set 205 features for each sound scene. Figure 5 presents the accuracy results obtained for each class. The mean accuracy of our pilot experiment after 4 folds was 75.96% for the development dataset.

The accuracy of our baseline system (MFCCs + SVM) was 71.32%. Note that the accuracy of the DCASE 2016 baseline system, based on MFCC features and a GMM classifier was 72.5 %.

The evaluation dataset was released shortly before submission. The results of the system here presented based on MFCC+monaural RQA+binaural features+RQA binaural obtained an accuracy of 81% whereas the DCASE 2016 baseline system obtained an accuracy of 77.2%.

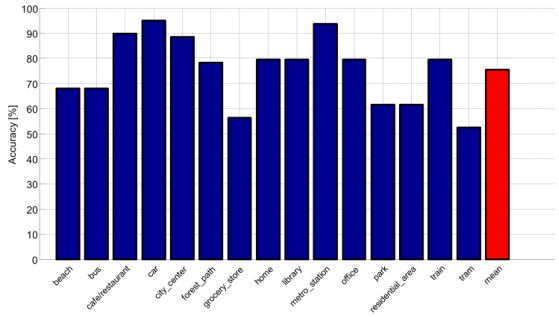

Figure 5: Accuracy for each class using the development dataset. The mean accuracy using the evaluation dataset was 75.96%.

4. CONCLUSIONS

This work has presented a method to classify sound scenes based on monaural and binaural features combined with RQA and a SVM classifier. The combination of monaural and binaural features improves the classification accuracy with respect to a baseline based solely on monaural features. The results here presented demonstrate that adding RQA with binaural features improve the accuracy of a sound scene classifier based on monaural features and a SVM by around 4 % based on the DCASe 2016 development dataset.

5. ACKNOWLEDGMENT

This work was supported by the DFG Cluster of Excellence EXC 1077/1 “Hearing4all”.

6. REFERENCES

[1] Y. Jiang, D. Wang, R. Liu, Z. Feng, “Binaural Classification for Reverberant Speech Segregation Using Deep Neural Networks”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, December, 2014.

[2] W. Nogueira, G. Roma, P. Herrera, “Sound scene identification based on MFCC, binaural features and a support vector machine classifier”, IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events, 2013.

[3] 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.

[4] G. Roma, W. Nogueira, P. Herrera, “Recurrence quantification analysis features for environmental sound recognition”, 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 1-4, 2013.

[5]http://www.ee.columbia.edu/˜dpwe/resources/matlab/rastamat/mfccs.html.

[6] 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

[7] A. Mesaros, T. Heittola, T. Virtanen, “TUT Database for Acoustic Scene Classification and Sound Event Detection”, in 24rd European Signal Processing Conference 2016 (EUSIPCO 2016). Budapest, Hungary, 2016.