AUDITORY SCENE CLASSIFICATION BASED ON THE SPECTRO-TEMPORAL STRUCTURE ANALYSIS

Tomasz Maka

Faculty of Computer Science and Information Technology,
West Pomeranian University of Technology, Szczecin
Zolnierska 49, 71-210 Szczecin, Poland
tmaka@wi.zut.edu.pl

ABSTRACT

In this report, we present a modular system for acoustic scenes classification. Our proposed system contains four modules to compute the representations describing spectro-temporal properties of audio data. The frequency components are extracted from cochleagram and low-level audio feature contours. An onset map is used to determine the properties of temporal structure, and binaural cues are additional components in the final feature space. Computed features are formed into vector and fed to random forests classifier for the purpose of classification. The results were submitted to the 2017 IEEE AASP DCASE challenge.

Index Terms— auditory scene analysis, cochleagram, onset map, feature contours, random forest

1. INTRODUCTION

Acoustic environment plays important role in many audio analysis and processing tasks, including: speech enhancement, support for multimodal systems, robotic navigation or human-machine interaction. Proper identification of acoustic scene may substantially influence on the robustness of speech and music services in real acoustic conditions. In recent years, many techniques to determine the type of acoustic scene have been proposed. A review of methods dedicated to acoustic scene classification is presented in [1]. The final classification accuracy rely on the attributes of sound and the classification strategy. In this report, we will briefly focus on the audio features and its discriminative properties.

2. SYSTEM OVERVIEW

The architecture of the system for auditory scene classification is shown in Fig. 1. The stereo input signal is fed into one module and, as a result of mixing both channels in 50/50 proportion, mono signal is used by the rest of components. Cochleagram [2] is employed by two modules and is generated using signals obtained at the outputs of gammatone filterbank with $N = 128$ bands (channels) covering frequency range from 50Hz to 8kHz.

The whole system was designed and implemented in C++ language and runs as multithreaded code. The modular architecture facilitates the design of auditory scene analysis systems, with varying level of computational constrains and various feature spaces used in the audio parameterization stage.

3. SPECTRO-TEMPORAL REPRESENTATION

The key element of acoustic analysis efficiency is the feature space. Audio features have to capture the properties of input signal in frequency domain and temporal dynamics of the signal. In this study, we have exploited four groups of audio features.

3.1. Binaural cues

Because the source audio data is recorded in two channels, we included module to calculate the binaural attributes. The common binaural cues comprise [3]: (1) interaural arrival-time difference (ITD), (2) interaural level differences (ILD) and (3) interaural coherence (IC). The stereo signal is split into 500ms long, non-overlapping frames, and for each frame ILD and IC are computed. After experimenting with binaural properties of provided data, we discovered quite low impact of ITD on the final accuracy and we did not include it in the binaural set. Using obtained contours of ILD and IC, we calculated means (μ) and standard deviations (σ) and added both pairs to feature set.

3.2. Onset map

Onset map is the representation based on the cochleagram. As the temporal properties of the audio signal are important in auditory perception [4], we decided to include such representation in our feature set. For each channel of cochleagram the onsets are detected and its attributes are calculated including the number of onsets and statistical properties of distances between adjacent onsets [5].

3.3. Low-level feature contour histograms

The next module of the system generates histograms of low-level audio feature contours. Each contour is calculated by splitting the input signal into frames (frame length = 20ms, overlapping = 50%), then for each frame a feature is computed. For obtained contour a normalization stage is performed, and histogram is estimated. We have used the following audio features [6, 7]: (1) spectral centroid, (2) spectral flatness measure, (3) dominant frequency component, (4) spectral entropy, (5) spectral sparsity, (6) spectral decrease, (7) high to low frequency ratio, (8) tonality, (9) energy in high frequencies, (10) spectral crest and 8 first linear-prediction coefficients.

3.4. Dominant bands histogram

An analysis of energy distribution over cochleagram's frequency bands is executed in the last module. At the output, a histogram

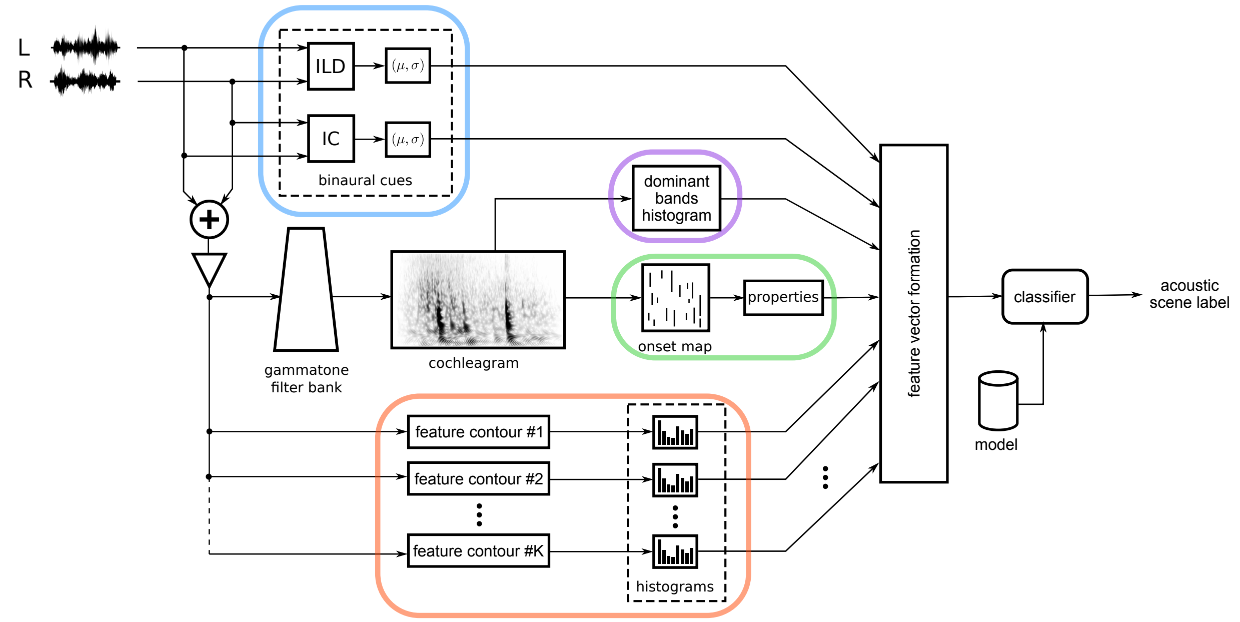

Figure 1: Proposed system architecture.

of dominant bands is provided using M dominant energies in each frame. Algorithm 1 presents a scheme to compute such histogram.

```
Algorithm 1: Dominant bands histogram calculation
  Inputs : $\Phi$ – cochleagram representation of audio signal,
           $M$ – number of dominant bands, $N$ – total number
           of frequency bands, $M < N$
  Output : $\Gamma^{(N)}$ – frequency bands histogram
  $\forall k \in [1, N]$ initialize $\Gamma_k^{(N)} = 0$
  for each frame of $\Phi$ do
      • Compute energies $E_n$ for each band of $\Phi$,
        $n = 1, \ldots, N$.
      • Sort the band numbers in ascending order by $E_n$.
      • Select first $M$ band numbers with the highest energies
        $E_m, m = 1, \ldots, M$.
      • For each selected number $m$ update histogram
        by $\Gamma_m^{(N)} = \Gamma_m^{(N)} + 1$
  end
  return $\Gamma^{(N)}$
```

4. EXPERIMENTAL RESULTS

All experiments were performed on development set of the DCASE'2017 challenge. Description of dataset and task details can be found in [8]. In the classification stage we employed random forest classifier [9]. This classifier performs bagging procedure to reduce the variance by averaging trees and uses majority voting for the final decision. The number of trees exploited in experiments was equal to 850.

In the first experiment, an influence of the number of dominant bands (M) on the classification accuracy using $\Gamma^{(128)}$ histogram representation was executed. The result is depicted in Fig. 2. Next, the classification was conducted using separate feature extraction modules of the development set. Obtained scores of both separate and joint folds is shown in Tab. 1. Finally, the classification accuracy of the development set for all four modules used in the parameterization is depicted in Tab. 2. The cases where we obtained better results than baseline approach are marked with boxes.

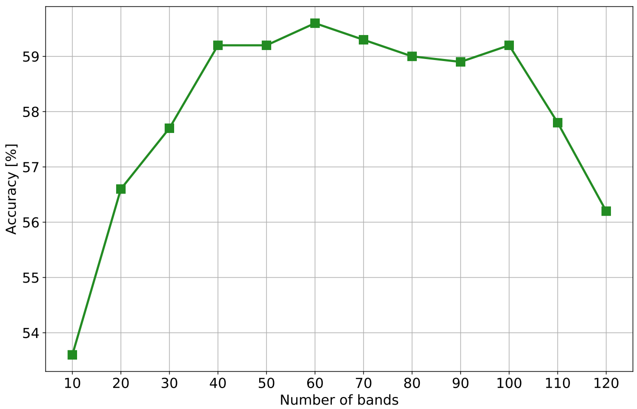

Figure 2: Classification accuracy for feature space generated by Algorithm 1.

5. CONCLUSIONS

The presented approach exploits the spectro-temporal properties of sound data to determine the type of auditory scene. The modules used in this study are selected components of our developed system

Table 1: Classification accuracy of separate representations.

Representation	**Accuracy [%]**				
	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Average**
Dominant bands histogram (M=50)	61.4	63.4	53.6	58.4	59.2
Onset map	38.8	34.0	35.9	33.3	35.5
Binaural cues	39.0	39.3	36.2	36.2	37.7
Low-level feature contours (K=18)	52.7	55.1	55.3	58.8	55.5

Table 2: Class-wise accuracy of the proposed system.

Acoustic scene	**Accuracy [%]**				
	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Average**
Beach	$\boxed{88.5}$	61.5	85.9	53.8	72.4
Bus	85.9	$\boxed{78.2}$	$\boxed{87.2}$	$\boxed{94.9}$	$\boxed{86.5}$
Cafe/Restaurant	46.2	51.3	$\boxed{80.8}$	46.2	56.1
Car	80.8	93.6	$\boxed{96.2}$	94.9	91.3
City center	78.2	85.9	78.2	85.9	82.1
Forest path	$\boxed{98.7}$	94.9	75.6	82.1	87.8
Grocery store	$\boxed{88.5}$	$\boxed{80.8}$	$\boxed{75.6}$	$\boxed{94.9}$	$\boxed{84.9}$
Home	$\boxed{84.6}$	$\boxed{76.5}$	50.6	55.1	66.7
Library	34.6	$\boxed{69.2}$	47.4	$\boxed{79.5}$	$\boxed{57.7}$
Metro station	53.8	64.1	62.8	76.9	64.4
Office	82.1	92.3	61.5	93.6	82.4
Park	74.4	$\boxed{74.4}$	$\boxed{61.5}$	$\boxed{80.8}$	$\boxed{72.8}$
Residental area	44.9	59.0	53.8	41.0	49.7
Train	37.2	30.8	24.4	44.9	34.3
Tram	73.1	56.4	79.5	69.2	69.6
Average	70.1	71.3	68.1	72.9	70.6

for real-time auditory scene analysis. In the presented form, the effectiveness can be still improved by the careful selection of low-level audio features. However, the acoustic similarities between classes and a short length of signals deteriorate the final accuracy. Despite the rather average results, the proposed solution is easily extensible, has low computational needs, and can be run in real-time.

6. REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, “Acoustic scene classification: Classifying environments from the sounds they produce,” *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May 2015.

[2] M. Slaney and R. F. Lyon, “On the importance of time – a temporal representation of sound,” in *Visual representations of speech signals*, M. Cooke, S. Beet, and M. Crawford, Eds. John Wiley & Sons, Inc., 1993, pp. 95–116.

[3] J. Blauert, *The Technology of Binaural Listening*, 1st ed. Springer, 2013.

[4] R. F. Lyon, *Human and Machine Hearing*. Cambridge University Press, May 2017.

[5] T. Maka, “Audio content analysis based on density of peaks in amplitude envelope,” in *39th International Conference on Telecommunications and Signal Processing – TSP’2016*, Vienna, Austria, June 27–29 2016, pp. 331–334.

[6] D. Mitrovic, M. Zeppelzauer, and C. Breiteneder, “Features for content-based audio retrieval,” in *Advances in Computers, Volume 78: Improving the Web*. Elsevier, 2010, vol. 78, pp. 71–150.

[7] A. Lerch, *An Introduction to Audio Content Analysis – Applications in Signal Processing and Music Informatics*. John Wiley & Sons, Inc., 2012.

[8] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, R. Badlani, E. Vincent, B. Raj, and T. Virtanen, “DCASE 2017 challenge setup: Tasks, datasets and baseline system,” in *Detection and Classification of Acoustic Scenes and Events 2017*, Munich, Germany, November 16 2017.

[9] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, October 2001.