CLASSIFICATION OF ACOUSTIC SCENES BASED ON MODULATION SPECTRA AND THE CEPSTRUM OF THE CROSS CORRELATION BETWEEN BINARUAL AUDIO CHANNELS

Technical Report

Arturo Paniagua, Rubén Fraile, Juana M. Gutiérrez-Arriola, Nicolás Sáenz-Lechón, Víctor J. Osma-Ruiz

Research Center on Software Technologies and Multimedia Systems for Sustainability (CITSEM)
Universidad Politécnica de Madrid, Madrid, Spain
arturo.paniagua.santamaria@alumnos.upm.es, r.fraile@upm.es, juana.gutierrez.arriola@upm.es, nicolas.saenz@upm.es, v.osma@upm.es

ABSTRACT

A system for the automatic classification of acoustic scenes is proposed that uses one audio channel for calculating the spectral distribution of energy across auditory-relevant frequency bands, and some descriptors of the envelope modulation spectrum (EMS) obtained by means of the discrete cosine transform. When the stereophonic signal captured by a binaural microphone is available, this parameter set is augmented by including the first coefficients of the cepstrum of the cross-correlation between both audio channels. This cross-correlation contains information on the angular distribution of acoustic sources. These three types of features (energy spectrum, EMS and cepstrum of cross-correlation) are used as inputs for a multilayer perceptron with two hidden layers and a number of adjustable parameters below 15,000.

Index Terms— Acoustic scene classification, modulation spectrum, cepstrum, Multilayer Perceptrons

1. INTRODUCTION

This submission consists of a system for the classification of acoustic scenes based on a combination of features obtained from the envelope modulation spectrum (EMS) [1] calculated using a gammatone filter-bank [2], and from the cepstrum claculated from the cross-correlation function of the left and right channels, in case these are available. The EMS is calculated from both audio channels. These features are used as inputs for a standard Multilayer Perceptron (MLP) with only two hidden layers [3].

2. MATERIALS

Audio recordings correspond to the datasets specified for DCASE 202 tasks 1A and 1B [4]. For the 1A task, audios are taken from the TAU Urban Acoustic Scenes 2020 Mobile dataset [5]. This dataset consists of recordings captured at distinct locations in 12 European cities with four different devices, and split into 10-second segments. The duration of recordings ranged from 5 to 6 min. The first deivce was a Zoom F8 multitrack recorder connected to a Soundman OKM II Klassik/studio A3 binaural microphone, hence producing a stereophonic signal. The microphone response can be considered flat between 20 Hz and 20 kHz. The other three devices were consumer devices not designed for professional audio performance: two mobiles phones (Samsung Galaxy S7, and iPhone SE) and a video camera (GoPro Hero5 Session). Some recordings obtained with the first device were used as sources for generating audio signals corresponding to 11 additional simulated devices by calculating the convolution with corresponding estimated impulse responses. Each recording location corresponded to one of the classes listed in Tab. 1.

#	Class name	Location type
1	Airport	Indoor
2	Indoor shopping mall	Indoor
3	Underground station	Indoor
4	Pedestrian street	Outdoor
5	Public square	Outdoor
6	Street with medium level of traffic	Outdoor
7	Travelling by tram	Transport
8	Travelling by bus	Transport
9	Travelling by underground	Transport
10	Urban park	Outdoor

Table 1: Classes of acoustic scenes: 3 transport, 4 indoor, 3 outdoor.

For task 1A all audio recordings were converted to a sampling rate equal to 44.1 kHz and 24 quantization bits, regardless the capabilities of the recording device. For device A, only one channel was used. In the case of task 1B, only recordings from device A were used, but including both channels, and sampled at 48kHz.

3. SIGNAL ANALYSIS

All audio crecodings were first preprocessed to subtract their mean values. Their mean square values were subsequently normalised. When both binaural channels were processed (task 1B), normalisation was performed by the same factor in both channels so as to preserve their level differences, that is, the root mean square value of all samples included in both channels was computed for normalisation. Afterwards, each audio signal was split in frames with duration 1.5 seconds, and 35% overlap between consecutive frames.

Each frame was processed by a filter-bank consisting of 26 gammatone filters [2] with central frequencies ranging from 27.5 Hz to 3587 Hz. The central frequencies of the filter-bank were chosen so that the pass-bands of contiguous filters were adjacent but not overlapping, i.e. the upper cut-off frequency of one filter was the same as the lower cut-off frequency of the next. Figure 1 illus-

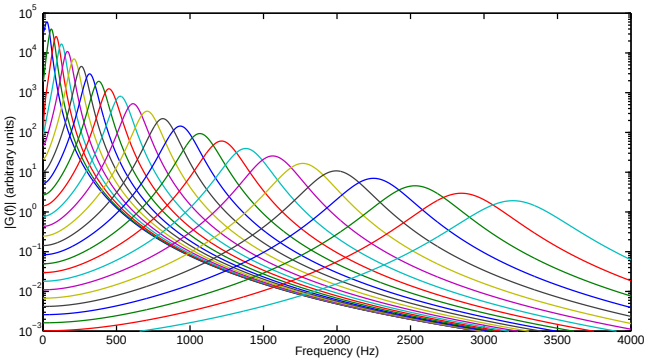

Figure 1: Frequency responses of the filters in the filter-bank with central frequencies up to 3.5 kHz (25 filters).

trates the frequency responses for the first filters.

In CASA systems, the filter-bank modelling the cochlear frequency behaviour is followed by a non-linear model of neuromechanical transduction [6]. This non-linear system approximately performs compression of the higher signal peaks and half-wave rectification [7]. As this produces a too detailed set of signals, it is usual to apply low-pass filtering and decimation afterwards [8]. The implementation of this model is computationally expensive due to its non-linearities. For this reason, we substitute it by full-wave rectification followed by a 5^{th} order Butterworth low-pass filter with cut-off frequency equal to 80 Hz and decimation to yield a sampling frequency equal to 200 Hz.

Each resulting frame is further processed by computing its discrete Fourier transform (DFT). The EMS [1] is obtained by stacking the square modulus of the DFT corresponding to the 26 gammatone filters. In order to reduce the dimensionality of the EMS, its components corresponding to the fastest variations of the signal were discarded. Specifically, a threshold of 30 Hz was set for the modulation frequency. Therefore, each signal frame was represented by a matrix, i.e. EMS, of 26×11 elements. The first data column represents the average energy at the output of each gammatone filter, i.e. the long-term average spectrum (LTAS) of the audio frame. The remaining 10 columns represent the energies of amplitude modulations between 0 and 3 Hz, between 3 and 6 Hz, etc.

The signal analysis scheme described so far transforms one channel of the audio recorded during 1.5 seconds into a feature vector of $26 \times 11 = 286$ components. The dimensionality of this feature space was reduced as follows. As stated before, the first column in the EMS corresponds to the average energy at each frequency band. This is relevant for discriminating among certain types of acoustic events [9], so the corresponding 26 values for each EMS were kept unchanged. Only a logarithm operation was applied in order to reduce the skewness of their distribution. Similarly to the approach in [10], the remaining 15 columns of each EMS were processed as if they were a grey-scale image. Specifically, the two-dimensional discrete cosine transform (DCT) [11] of the logarithm of the EMS was calculated, and the block corresponding to the first 10×10 DCT coefficients was chosen as a lower-dimensional representation of each 26×15 EMS. Therefore, after this dimensionality reduction, each audio frame of duration 1.5 s was represented by a feature vector with $26 + 100 = 126$ components.

The cross correlation between the signals captured at microphones placed in different positions is known to incorporate information about the directions of arrival [12], and hence the spatial distribution of sound sources. The shape of the cross correlation is coded using the first coefficients of its power cepstrum [13]. Specifically, coefficients up to the 200^{th} were taken, but discarding the first one since it only represents the average spectral energy. Additionally, for task 1B the resolution for modulation frequencies was 2Hz, thus producing 15 modulation bands: from 2Hz to 30Hz in steps of 2Hz. Therefore, for task 1B the feature vectors representing each 1.5s audio frame consisted of 126 coefficients describing the EMS of the left channel plus 199 coefficients corresponding to the cepstrum of the cross-correlation between left and right channels.

4. CLASSIFICATION

The afore-mentioned feature vectors were used as inputs for a multilayer perceptron (MLP) two hidden layers. For task 1A the first hidden layer comprised 40 neurons. The first 8 neurons were connected to the 26 inputs corresponding to the LTAS of each frame; the reamining 32 neurons were connected to the 10 × 10 DCT coefficients representing the EMS. The second hidden layer was composed by 24 neurons fully connected to the first hidden layer. The hyperbolic tangent was chosen as the activation function for hidden neurons. The output layer was formed by 10 neurons, one corresponding to each class in Tab. 1. These output neurons had *softmax* activation functions[3]. Thus, their outputs represented the estimated a posteriori probabilities of each scene class corresponding to the input feature vector associated to each 1.5 s frame. The MLP corresponding to task 1A had 11,264 adjustable parameters, each having a length of 8 bytes.

The overall a posteriori probability of each class for a 10 s audio segment was estimated by adding up the logarithms of the probabilities of its frames. For all frames, segments and recordings, the class assigned by the MLP was estimated to be the class yielding the highest a posteriori log-probability.

For task 1B, a MLP with a similar structure was used, but with a different number of neurons per layer. In this case the first hidden layer consisted of 81 neurons: 50 linked to the inputs corresponding to the cepstrum of the cross-correlation, 6 linked to the LTAS, and 25 taking inputs from the 10×10 coefficients of the DCT of the EMS. The second hidden layer included 6 neurons, fully connected to adjacent layers. The output layer comprised 3 neurons, one corresponding to each scene type (indoor, outdoor,transport). The MLP corresponding to task 1B had 13,197 adjustable parameters, each having a length of 8 bytes.

5. EXPERIMENTS & RESULTS

The classification experiment corresponding to the baseline evaluation procedure proposed for the Acoustic Scene Classification with Multiple Devices task (1A) DCASE 2020[4] was run. The overall correct classification rate (CCR) for audio segments is 57.07%, while the per-class performance is as indicated in table The confusion matrix corresponding to this experiment is in Tab. 2.

Results corresponding to the Low-Complexity Acoustic Scene Classification task (1B) are summarised in Tab. 2.

6. CONCLUSIONS

This paper presents a system for the automatic classification of acoustic scenes based on the EMS and the cross-correlation between

Class	CCR (%)
Airport	41.1
Indoor shopping mall	54.9
Underground station	37.4
Pedestrian street	35.7
Public square	29.0
Street with medium level of traffic	82.2
Travelling by tram	95.3
Travelling by bus	74.4
Travelling by underground	48.1
Urban park	72.7

Table 2: Per-class correct classification rates for task 1A.

Class	CCR (%)
Indoor	84.05
Outdoor	89.30
Transport	89.45
OVERALL	**87.77**

Table 3: Per-class correct classification rates for task 1B.

binaural channels when available. The signal analysis scheme was designed taking into account several issues. The first stages of the system are a simplification of the peripheral auditory system [8]. The specific responses of the gammatone filters were chosen so that the filter-bank fully covered the pass-band of the microphone. The average energy at the output of each filter was kept as a feature, hence accounting for the relevance of the energy spectrum for acoustic event detection [9]. Slow modulations of these energies were described by reducing the dimensionality of the EMS using the DCT, a common-use tool for data compression in image processing [11]. Information on the spatial distribution of sound sources present in binarual recordings has been represented using the power cepstrum of the cross-correlation between audio channels. In all cases, the signal bandwidth has been limited to less than 4kHz in order to increase robustness against diversity in device bandwidth.

7. REFERENCES

[1] J. M. Liss, S. LeGendre, and A. J. Lotto, "Discriminating dysarthria type from envelope modulation spectra," vol. 53, no. 5, pp. 1246–1255, 2010.

[2] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Speech-Group Meeting of the Institute of Acoustics on Auditory Modelling*, RSRE, Malvern, 1987.

[3] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2015.

[4] "Dcase2020 challenge," DCASE Community," DCASE, 2020. [Online]. Available: http://dcase.community/challenge2020/[Visited:15/06/2020]

[5] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. of DCASE2018*, 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

[6] G. J. Brown and M. Cooke, "Computational auditory scene analysis," vol. 8, no. 4, pp. 297–336, 1994.

[7] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," vol. 79, no. 3, pp. 702–711, 1986.

[8] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE press, 2006.

[9] J. M. Gutiérrez-Arriola, R. Fraile, A. Camacho, T. Durand, J. L. Jarrín, and S. R. Mendoza, "Synthetic sound event detection based on MFCC," in *Proc. of DCASE2016*, 2016, pp. 30–34.

[10] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," vol. 18, no. 2, pp. 130–133, 2011.

[11] W. H. Chen and W. Pratt, "Scene adaptive coder," vol. 32, no. 3, pp. 225–232, 1984.

[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," vol. 24, no. 4, pp. 320–327, 1976.

[13] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," vol. 65, no. 10, pp. 1428–1443, 1977.

[14] I. Nabney, *NETLAB: algorithms for pattern recognition*. Springer Science & Business Media, 2002.