CLASSIFICATION OF ACOUSTIC SCENES BASED ON MODULATION SPECTRA AND POSITION-PITCH MAPS

Technical Report

Rubén Fraile, Elena Blanco-Martín, 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
{rfraile, eblanco, jmga, nslechon, vosma}@etsist.upm.es

ABSTRACT

A system for the automatic classification of acoustic scenes is proposed that uses the stereophonic signal captured by a binaural microphone. This system uses one channel for calculating the spectral distribution of energy across auditory-relevant frequency bands. It further obtains some descriptors of the envelope modulation spectrum (EMS) by applying the discrete cosine transform to the logarithm of the EMS. The availability of the two-channel binaural recordings is used for representing the spatial distribution of acoustic sources by means of position-pitch maps. These maps are further parametrized using the two-dimensional Fourier transform. These three types of features (energy spectrum, EMS and position-pitch maps) are used as inputs for a standard multilayer perceptron with two hidden layers.

Index Terms— Acoustic scene classification, modulation spectrum, position-pitch map, multilayer perceptron

1. INTRODUCTION

The automatic classification of acoustic scenes, or computerised acoustic scene recognition (CASR) [1] aims at recognising the context in which a given acoustic signal is produced. While its objectives are different from those of computerised auditory scene analysis (CASA), both CASR and CASA share some common challenges and can thus be considered close to one another [2].

A significant portion of computerised acoustic scene recognition (CASR) system proposals are based on parametrisation schemes which describe the signal in either spectral or cepstral domains [3]. Consistently with typical approached for modelling the peripheral auditory system in computerised auditory scene analysis (CASA) [4], all the cited proposals include spectral analyses with greater bandwidths for higher frequencies. While the temporal dimension of perceived signals seems to be key for perception [4], only some of the previous works included modelling of the temporal evolution of the parameters in the set of proposed features. Alternative options for considering the temporal dimension in the classification scheme imply designing classifiers with time-varying outputs such as recurrent, convolutional or time-delay neural networks [3].

In other applications of acoustic signal processing, such as speaker recognition, the temporal dimension is modelled by calculating frame-to-frame variations of parameters [5], the so called Δ (short for 1^{st} derivative) and $\Delta\Delta$ (2^{nd} derivative) parameters. However, these are of limited value in the case of sound event detection, since $\Delta\Delta$ parameters added no significant improvement to the results in [6]. The problem of CASR is closely related to the problem of sound event detection [1]; therefore, this limited informative value of fast variations in parameter values is to be expected also in CASR.

From another point of view, the availability of binaural acoustic signals can lead to relevant improvements in CASR systems, specially if the spatial information present in these two-channel signals is exploited, instead of simply processing each channel independently and subsequently aggregating parameters into a single feature vector [7].

In this paper, we propose a system for the classification of acoustic scenes based on features obtained from the envelope modulation spectrum (EMS) [8] calculated using a gammatone filter-bank [9]. This EMS is calculated from one of the available audio channels, while the spatial information conveyed by the binaural signal is modelled by the position-pitch plane obtained after the cross-correlation function of the two channels [10]. These features are used as inputs for a simple multilayer perceptron (MLP) with only two hidden layers and as many *softmax* outputs as classes of acoustic scenes to be recognised [11].

2. MATERIALS

Audio recordings correspond to the TUT Urban Acoustic Scenes 2018 dataset [12]. This dataset consists of recordings captured at distinct locations and split into 10-second segments. The duration of recordings ranged from 5 to 6 min. A Zoom F8 multitrack recorder and a Soundman OKM II Klassik/studio A3 binaural microphone were used for recording, hence producing a stereophonic signal. The microphone response can be considered flat between 20 Hz and 20 kHz. Recordings were captured with sampling rate equal to 48 kHz and 24 quantization bits. Each recording location corresponded to one of the classes listed in Tab. 1.

3. SIGNAL ANALYSIS

The two audio channels comprising each recoding were first preprocessed to remove their mean values. Their combined mean square value was subsequently normalised. Normalisation was performed by the same factor in both channels so as to preserve their level

*This work has been partially funded by the Spanish Ministry for Economy and Competitiveness through project grant MAT2015-64139-C4-3-R.

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

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

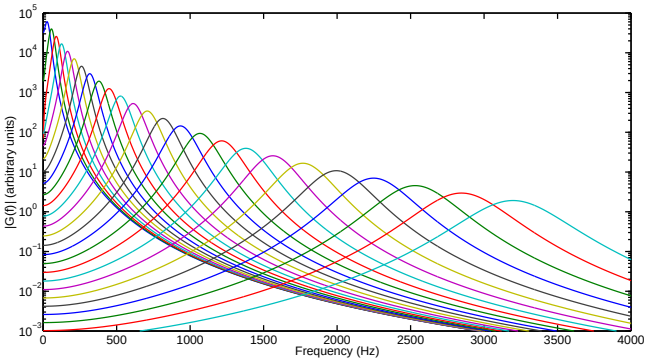

Figure 1: Frequency responses of the filters in the filter-bank with central frequencies up to 3.5 kHz (25 filters).

differences, that is, the root mean square value of all samples included in both channels was computed for normalisation. Afterwards, each channel was split in frames with duration 0.5 seconds, and 50% overlap between consecutive frames.

Each frame in the left channel was processed by a filter-bank consisting of 40 gammatone filters [9] with central frequencies ranging from 27.5 Hz to 17.09 kHz. 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 illustrates 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 [13]. This non-linear system approximately performs compression of the higher signal peaks and half-wave rectification [14]. As this produces a too detailed set of signals, it is usual to apply low-pass filtering and decimation afterwards [4]. 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 [8] is obtained by stacking the square modulus of the DFT corresponding to the 40 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 24 Hz was set for the modulation frequency. Therefore, each signal frame was represented by a matrix, i.e. EMS, of 40×9 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 8 columns represent the energies of amplitude modulations between 0 and 3 Hz, between 3 and 6 Hz, etc. Figure 2 depicts the LTAS (top) and the modulation energies (bottom) corresponding to a 0.5 s frame in an underground station.

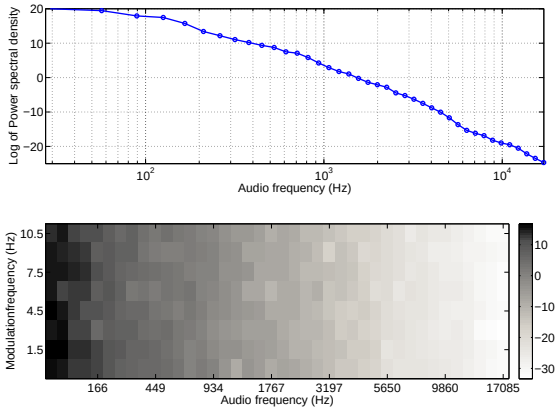

Figure 2: Logarithm of the LTAS (top) and map of modulation energies (bottom) of a sample 0.5 s audio frame.

The signal analysis scheme described so far transforms one channel of the audio recorded during 0.5 seconds into a feature vector of $40 \times 9 = 360$ components. The dimensionality of this feature space was reduced as follows. As stated before, the first column in the EMS (see Fig. 2) corresponds to the average energy at each frequency band. This is relevant for discriminating among certain types of acoustic events [6], so the corresponding 40 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 [15], the remaining 8 columns of each EMS were processed as if they were a grey-scale image. Specifically, the two-dimensional discrete cosine transform (DCT) [16] of the logarithm of the EMS was calculated, and the block corresponding to the first 8×8 DCT coefficients was chosen as a lower-dimensional representation of each 40×8 EMS. Therefore, after this dimensionality reduction, each audio frame of duration 0.5 s was represented by a feature vector with $(40 + 64) \cdot 2 = 104$ components.

The spatial information provided by the 2-channel recordings was represented by generating the position-pitch map $\rho(\varphi, f)$ defined as [10]:

$$\rho(\varphi, f) = \frac{1}{2K+1} \sum_{k=-k}^{K} R_{\text{lr}}\left(k\frac{f_{\text{s}}}{f} + \frac{df_{\text{s}}}{c}\cos\varphi\right) \qquad (1)$$

where φ (azimut - rad) and f (frequency - Hz) are the independent variables of the map, $R_{\text{lr}}(\tau)$ is the estimated cross-correlation between left and right channels at time lag τ, f_{s} is the sampling frequency (48 kHz), d is the interaural distance (estimated to be 14 cm for this experiment), c is the phase speed of sound (estimated to be 343 m/s for this experiment), and K is the largest possible integer given the maximum time lag τ for which $R_{\text{lr}}(\tau)$ has been estimated ($\tau \leq 100$ ms in our case).

The position-pitch map was calculated for each 0.5 s audio frame for $-\pi < \varphi \leq \pi$ with a resolution of $\frac{\pi}{60}$ rad, and for

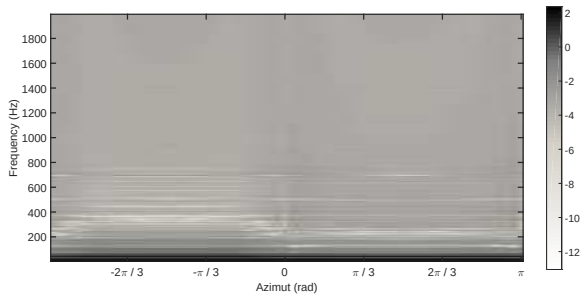

Figure 3: Position-pitch map of corresponding to a sample 0.5 s audio frame.

$20 < f \leq 2\,000$ with a resolution of 10 Hz. This produced a 120×199 map with shifts in the φ dependent on the orientation of the head-mounted microphone system. For illustration purposes, Fig. 3 depicts the position-pitch map corresponding to a 0.5 s frame in an underground station. In order to reduce the number of dimensions, a bidimensional discrete Fourier transform (2D DFT) was calculated, and only the 20×20 elements corresponding to the lowest spatial frequencies were taken as input features for the acoustic scene classifier. Furthermore, in order to make the parameters orientation-independent, only the modulus of the 2D DFT was considered.

4. CLASSIFICATION

The afore-mentioned feature vectors were used as inputs for a multilayer perceptron (MLP) two hidden layers. The first hidden layer comprised 24 neurons. The first 4 neurons were connected to the 40 inputs corresponding to the LTAS of each frame, a second group of 8 neurons were connected to the 8×8 DCT coefficients representing the EMS, and the remaining 12 neurons had the 20×20 2D DFT coefficients or the position-pitch map as inputs. The second hidden layer was composed by 12 neurons fully connected to the first hidden layer. The output layer was formed by 10 neurons, one corresponding to each class in Tab. 1. These output neurons had *softmax* activation functions [11]. Thus, their outputs corresponded to the estimated *a posteriori* probabilities of the input feature vector, or the 0.5 s frame, corresponding to each scene class.

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.

5. EXPERIMENTS & RESULTS

The classification experiment corresponding to the baseline evaluation procedure proposed for the acoustic scene classification challenge in DCASE 2018[12] was run. The confusion matrix corresponding to this experiment is in Tab. 2. The overall correct classification rate (CCR) for audio segments is 62.3%. It is noteworthy that if classes are grouped in three types: indoor (airport, shopping mall and underground station), outdoor (public squate, pedestrina street, street with traffic and urban park), and in-transport (tram, bus and underground), the majority of confusions happen between classes of the same type. In fact, the system classifies audio segments from indoor environments as corresponding to one of the indoor classes in 79.7% of cases. Similarly, for outdoor classes this rate reaches 88.3%, and for in-transport classes the rate is 92.9%.

6. CONCLUSIONS

This paper presents a system for the automatic classification of acoustic scenes based on the EMS and position-picth maps. The proposed system exploits the availability of two channels in the stereophonic recordings by building a representation of the spatial distribution of sound sources from the cross correlation between the binaural signals. Features from both types of analysis are subsequently combined to build a feature vector for each audio frame.

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 [4]. 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 [6]. 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 [16]. In turn, the dimensionality of position-pitch maps was reduced by calculating the 2D DFT, and the parametrization scheme was made orientation-invariant by taking only the modulus of such 2D DFT.

Reported results (Tab. 2) indicate that the proposed system performs better (CCR $\approx$ 62.3%)than the baseline system provided in DCASE 2018. In addition, system errors mainly happen between classes of the same type, either indoor, outdoor or in-transport. Thus, the system correctly indentifies the type of class in 75.3% of cases, without any explicit training in this regard.

7. REFERENCES

[1] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *IEEE Internat. Conf. on Acoust., Speech, and Signal Process. (ICASSP)*, vol. 2, 2002, pp. II/1941–II/1944.

[2] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[3] "Acoustic scene classification. Challenge results," Tampere University of Technology," DCASE, 2017. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification-results[Visited: 13/10/2017]

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

[5] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Adv. Signal Process.*, vol. 2004, no. 4, pp. 1–22, 2004.

[6] 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.

True class	Assigned class #									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Airport	52.1	17.0	4.9	19.6	2.6	0	0.4	0	3.4	0
Shopping mall	7.2	87.1	1.1	4.7	0	0	0	0	0	0
Underground station	7.7	14.3	46.7	9.3	0.8	4.6	3.1	3.5	8.9	1.2
Pedestrian street	8.5	6.9	3.2	45.8	29.2	4.9	0	0	1.6	0
Public square	0.5	0	6.9	6.9	50.9	15.3	0	3.2	0	16.2
Street with traffic	0	0.4	2.4	5.3	10.2	78.1	0	0	0.8	2.9
Tram	3.83	0	0.8	0	0	0	67.4	12.3	14.9	0.8
Bus	0	0	0	0	0	0	19.0	78.9	1.2	0.8
Underground	0	0	2.3	2.3	5.8	3.8	47.9	4.6	33.0	0.4
Park	0	0	5.0	0.4	0.8	4.6	6.2	0.4	0.4	82.2

Table 2: Confusion matrix (in %) for the baseline evaluation experiment. Class numbers in column headers correspond to the order in Tab. 1.

[7] S. Adavanne and T. Pertilä, P. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *IEEE Internat. Con. on Acoust., Speech, and Signal Process. (ICASSP)*, 2017, pp. 771–775.

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

[9] 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.

[10] M. Kepesi, F. Pernkopf, and M. Wohlmayr, "Joint position-pitch tracking for 2-channel audio," in *Proc. of CBMI'07.*, 2007, pp. 303–306.

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

[12] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," 2018, submitted to DCASE2018 Workshop. [Online]. Available: https://arxiv.org/abs/1807.09840

[13] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[14] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Amer.*, vol. 79, no. 3, pp. 702–711, 1986.

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

[16] W. H. Chen and W. Pratt, "Scene adaptive coder," *IEEE Trans. Commun.*, vol. 32, no. 3, pp. 225–232, 1984.

[17] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. of DCASE2017*, 2017, submitted.