CLASSIFICATION OF ACOUSTIC SCENES BASED ON THE MODULATION SPECTRUM

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
{rfraile, jmga, nslechon, vosma}@etsist.upm.es

ABSTRACT

A system for the automatic classification of acoustic scenes is proposed. This system calculates the spectral distribution of energy across auditory-relevant frequency bands and obtains some descriptors of the envelope modulation spectrum (EMS) by applying the discrete cosine transform to the logarithm of the EMS. This parametrisation scheme achieves good separation among scene classes, since it gets good classification results with a simple classifier consisting of a multilayer perceptron with only one hidden layer.

Index Terms— Acoustic scene classification, modulation spectrum, 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 CASR system proposals are based on parametrisation schemes which describe the signal in either spectral [3, 4, 5, 6] or cepstral domain [3, 7, 8, 9]. Consistently with CASA approaches for modelling the peripheral auditory system, 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, only [9] among 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 [3], convolutional [4] or time-delay neural networks [9].

In other applications of acoustic signal processing, such as speaker recognition, the temporal dimension is modelled by calculating frame-to-frame variations of parameters [10], 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 [11]. 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.

In this paper, we propose a system for the classification of acoustic scenes based on features obtained from the envelope modulation spectrum (EMS) [12] calculated using a gammatone filterbank [13]. These features are used as inputs for a simple multilayer perceptron (MLP) with only one hidden layer and as many *softmax* outputs as classes of acoustic scenes to be recognised [14].

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

#	Class name	Complementary information
1	Bus	Travelling by bus in the city
2	Café/Restaurant	Small café/restaurant
3	Car	Driving or travelling as a passenger
4	City centre	Outdoor
5	Forest path	
6	Grocery store	Medium size grocery store
7	Home	
8	Beach	Lakeside beach
9	Library	
10	Metro station	Indoor
11	Office	Several people, typical working day
12	Residential area	Outdoor
13	Train	Travelling
14	Tram	Travelling
15	Urban park	Outdoor

Table 1: Classes of acoustic scenes: 4 vehicle, 6 indoor, 5 outdoor.

2. MATERIALS

Audio recordings correspond to the TUT Acoustic Scenes 2017 dataset [15]. This dataset consists of recordings captured at distinct locations and split into 10-second segments. The duration of recordings ranged from 3 to 5 min. A Roland Edirol R-09 wave 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 44.1 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 combined to produce an alternative two-channel representation in which the first channel corresponded to the average of both original channels and the second transformed channel corresponded to the absolute value of the difference between the original channels. Afterwards, each transformed channel was split in frames with duration 0.5 seconds, and 50% overlap between consecutive frames.

Each frame was processed by a filter-bank consisting of 40

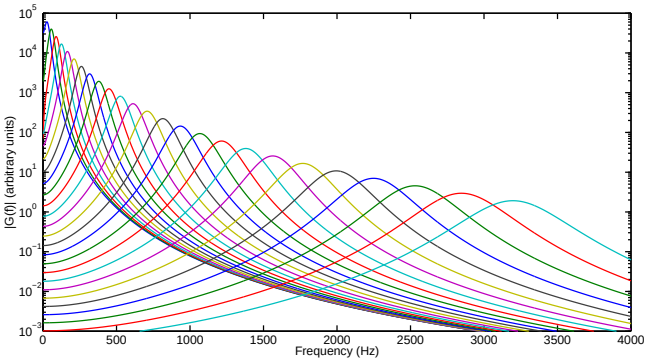

Figure 1: Frequency responses of the filters in the filter-bank with central frequencies up to 3.5 kHz (25 filters).

gammatone filters [13] 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 [16]. This non-linear system approximately performs compression of the higher signal peaks and half-wave rectification [17]. As this produces a too detailed set of signals, it is usual to apply low-pass filtering and decimation afterwards [18]. 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 [12] 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 40×13 elements. Figure 2 depicts the two EMS corresponding to 0.5 s of recording in a residential area. The first data column represents the average energy at the output of each gammatone filter, while the remaining 12 columns represent the energies of amplitude modulations at 2 Hz, 4 Hz, etc.

The signal analysis scheme described so far transforms the audio recorded during 0.5 seconds into a feature vector of $40 \times 13 \times 2 = 1040$ 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 [11], so the corresponding 40 values for each EMS were kept unchanged. Similarly to the approach in [19], the remaining 12 columns of each EMS were processed as if they were a grey-scale image. Specifically, the two-dimensional discrete cosine transform (DCCT) [20] of the logarithm of the EMS was calculated, and the block corresponding to the first 7×7 DCT coefficients was chosen as a lower-dimensional representation of each 40×12 EMS. Therefore, after this dimensionality reduction, each

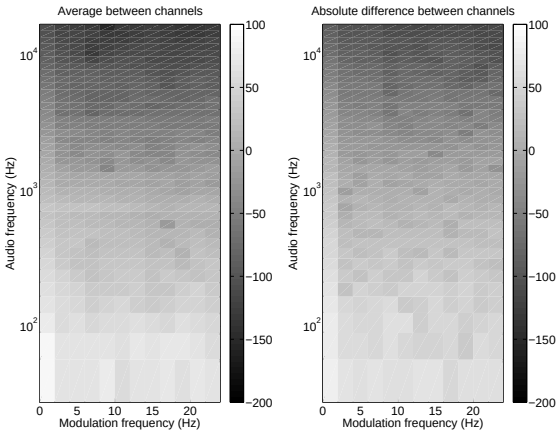

Figure 2: EMS (in dB) of the two transformed channels corresponding to one frame of audio recorded in a residential area.

audio frame of duration 0.5 s was represented by a feature vector with $(40 + 49) \cdot 2 = 178$ components.

4. CLASSIFICATION

The afore-mentioned feature vectors were used as inputs for a multilayer perceptron (MLP) with only one hidden layer comprising 20 hidden neurons. The selected activation function for these hidden neurons was the hyperbolic tangent due to its symmetry. The output layer was formed by 15 neurons, one corresponding to each class in Tab. 1. These output neurons had *softmax* activation functions [14]. 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 multiplying the probabilities of its frames. Similarly, the probabilities associated to the full recordings were calculated by multiplication of the probabilities of their corresponding segments. For all frames, segments and recordings, the class assigned by the MLP was estimated to be the class yielding the highest *a posteriori* probability.

5. EXPERIMENTS & RESULTS

Two classification experiments were run using the system described so far. In the first experiment, all recordings were randomly divided into training (80%) and test (20%) sets. Care was taken to ensure that all classes were present in both sets in the same proportion. This experiment was repeated for five times, with new sets randomly chosen for each repetition.

The second experiment consisted in the baseline evaluation procedure proposed for the acoustic scene classification challenge in DCASE 2017[1] [21].

[1] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification

5.1. First experiment

The results of the first experiment are summarised in the graphs of Fig. 3. These represent the correct classification rate (CCR) of frames, segments and recordings belonging to each scene class. The average CCR (*Overall*) is also plotted for reference in each graph.

One can observe that the CCR tends to improve as the length of the processed audio signals is increased. For all classes except for *Urban park*, the CCR for segments is greater than for frames, with values over 75%. In contrast, the *Urban park* yields the worst results, which are fairly independent from signal length (58%). The overall CCR reaches 90% for full recordings.

5.2. Second experiment

The confusion matrix corresponding to the second experiment is in Tab. 2. The overall CCR for audio segments is 79.8%. It is noteworthy that some of the highest error rates happen between classes that may be difficult to distinguish even for a human listener: *Urban park* vs *Residential area*, *Train* vs *Tram*, *Grocey store* vs *Café/Restaurant*, *Home* vs *Library*.

6. CONCLUSIONS

This paper presents a system for the automatic classification of acoustic scenes based on the EMS. The proposed system exploits the availability of two channels in the stereophonic recordings by computing the average and the absolute difference of both channels and processing them independently. Features from both signals 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 [18]. 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 [11]. 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 [20].

The reported results indicate a good performance of the system (CCR$\approx$ 75% at segment level), except for the *Train* and *Urban park* classes (CCR$\approx 57-59\%$). The graphs in Fig. 3 show that classification results improve as longer audio signals are processed. This implies that there are audio frames much more relevant for the identification of acoustic scenes than the rest. In other words, some acoustic scenes seem to be especially characterised by certain acoustic events. If recordings are made long enough for such events to happen, then the classification accuracy can be increased.

Results from the second experiment (Tab. 2) are better than the baseline system provided in DCASE 2017. In addition, it should be noted that some of the highest error rates happen between classes that can be difficult to identify for human listeners.

Last, the adequacy of the proposed signal processing system to the problem of acoustic scene classification is suggested by the fact that good classification results can be achieved even with a simple classifier consisting of a MLP with only one hidden layer.

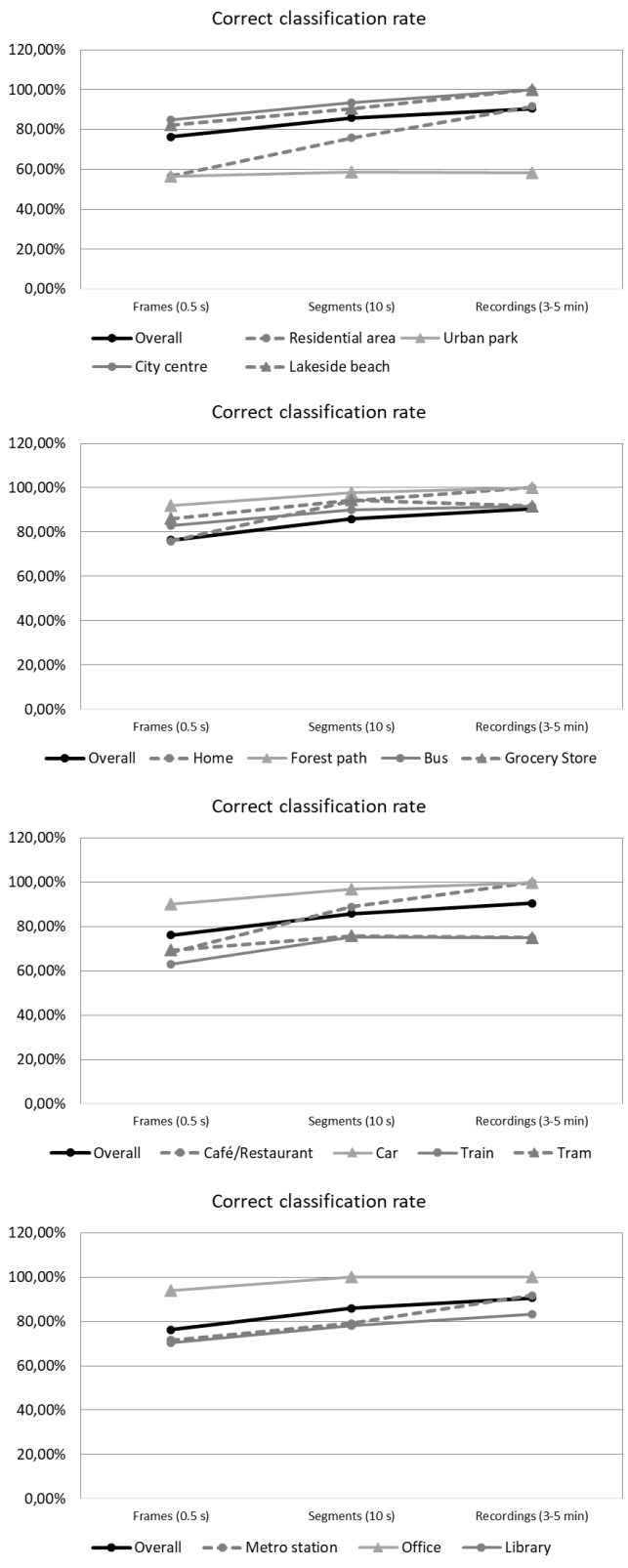

Figure 3: Correct classification rate for the test sets of the first experiment.

True class	Assigned class #														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Bus	94.2	0	2.2	0	0	0	0.6	0	0	0	0	0	0.6	2.2	0
Café/Restaurant	0	70.2	0	1.3	0	11.5	7.7	0	1.9	5.1	0	0	1.6	0.6	0
Car	1.0	0	95.5	0	0	0	0	0	0	0	0	0	1.9	1.3	0.3
City centre	0	0.6	0	88.5	1.0	0.6	0	1.28	0	1.0	0	4.8	0	0	2.2
Forest path	0	0.6	0	1.9	87.8	0	1.0	0.6	0	0	0	6.1	0	0	0.3
Grocery store	0	8.0	0	0	0	82.7	0.6	0	0	7.1	0	1.6	0	0	0
Home	0.6	1.3	0	0	0	1.9	77.0	0	9.1	0	9.8	0	0	0.3	0
Beach	0	0	0.3	0	2.6	0.3	1.0	79.8	0	0	0	13.8	0	0	2.2
Library	0	0.3	0	0	0	1.3	9.6	0	80.5	0.6	0	0.6	5.5	1.6	0
Metro station	0	2.9	0	0.6	0.6	4.2	0	0	6.7	80.8	0.6	0	0.6	0	0
Office	0	0	0	0	0	0	5.8	0	0.3	1.6	92.3	0	0	0	0
Residential area	0	0.6	0	6.7	2.9	0	0	0.6	0	0.3	0	69.9	0.6	0	18.3
Train	3.5	8.65	2.2	5.1	0	0	0.3	0.3	0.6	1.6	0	2.9	57.4	17.0	0
Tram	1.9	0.3	0	0	0	6.7	0	0	1.0	0	0	0	9.3	80.8	0
Urban park	0	2.6	0	5.8	0	0	1.3	4.5	0	0	0	26.6	0	0	59.3

Table 2: Confusion matrix (in %) for the second experiment. Class numbers in column headers correspond to the order in Tab. 1. Note that rows corresponding to *Forest path*, *Metro station* and *Train* do not sum up 100% because no class was assigned to audio segments with all their frames containing recording errors.

Class	CCR (%)
Bus	46.3
Café/Restaurant	47.2
Car	76.9
City centre	88.9
Forest path	65.7
Grocery store	48.1
Home	95.4
Beach	61.1
Library	35.2
Metro station	63.0
Office	24.1
Residential area	63.9
Train	75.0
Tram	53.7
Urban park	29.6

Table 3: Correct classification rate (CCR) corresponding to the DCASE 2017 evaluation results.

APPENDIX: EVALUATION RESULTS

Performance of the proposed system for the DCASE 2017 evaluation dataset is reported in [22]. The overall accuracy was 58.3%. Per-class results are summarised in Tab.3.

7. REFERENCES

[1] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *IEEE Internat. Con. 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] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. of DCASE2016*, 2016, pp. 11–15.

[4] T. Lidy and A. Schindler, "CQT-based convolutional neural networks for audio scene classification," in *Proc. of DCASE2016*, 2016, pp. 60–64.

[5] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, S. Squartini, and B. Schuller, "Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification," in *Proc. of DCASE2016*, 2016, pp. 65–69.

[6] G. Sena Mafra, N. Q. K. Duong, A. Ozerov, and P. Pérez, "Acoustic scene classification: An evaluation of an extremely compact feature representation," in *Proc. of DCASE2016*, 2016, pp. 85–89.

[7] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and I. Lane, "Experiments on the DCASE challenge 2016: Acoustic scene classification and sound event detection in real life recording," in *Proc. of DCASE2016*, 2016, pp. 20–24.

[8] Q. Kong, I. Sobieraj, W. Wang, and M. Plumbley, "Deep neural network baseline for DCASE challenge 2016," in *Proc. of DCASE2016*, 2016, pp. 50–54.

[9] N. Moritz, J. Schröder, S. Goetze, J. Anemüller, and B. Kollmeier, "Acoustic scene classification using time-delay neural networks and amplitude modulation filter bank features," in *Proc. of DCASE2016*, 2016, pp. 70–74.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] A. Mesaros, T. Heittola, and T. Virtanen, "TUT acoustic scenes 2017, development dataset," 2017. [Online]. Available: https://doi.org/10.5281/zenodo.400515

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

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

[18] D. Wang and G. J. Brown, "Fundamentals of computational auditory scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. J. Brown, Eds. New York, USA: Wiley Interscience, 2006, pp. 1–44.

[19] 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.

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

[21] 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.

[22] "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]