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

Technical Report

Rubén Fraile, Juan Carlos Reina, Juana M. Gutiérrez-Arriola, Elena Blanco-Martín

Research Center on Software Technologies and Multimedia Systems for Sustainability (CITSEM)
Universidad Politécnica de Madrid, Madrid, Spain
r.fraile@upm.es, jc.reina@alumnos.upm.es, juana.gutierrez.arriola@upm.es, elena.blanco@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 Gaussian Mixture Model with 64 components.

***Index Terms*—** Acoustic scene classification, modulation spectrum, position-pitch map, Gaussian Mixture Models

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 position-pitch plane obtained after the cross-correlation function of the left and right channels [3]. The EMS is calculated from both audio channels. These features are used as inputs for a standard Gaussian mixture model (GMM) with 64 Gaussian components[4].

2. MATERIALS

Audio recordings correspond to the TUT Urban Acoustic Scenes 2018 dataset [5]. 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.

#	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.

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 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 1.5 seconds, and 50% overlap between consecutive frames.

Each frame in the both left and right channels was processed by a filter-bank consisting of 40 gammatone filters [2] 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 [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 5th 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

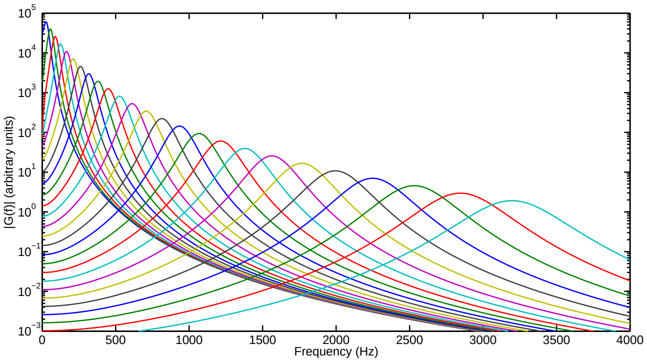

Figure 1: Frequency responses of the filters in the filter-bank with central frequencies up to 3.5 kHz (25 filters).

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.

The signal analysis scheme described so far transforms one channel of the audio recorded during 1.5 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 corresponds to the average energy at each frequency band. This is relevant for discriminating among certain types of acoustic events [9], 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 [10], 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) [11] 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 1.5 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 $\rho\left(\varphi, f\right)$ defined as [3]:

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

where φ (azimut - rad) and f (frequency - Hz) are the independent variables of the map, $R_{\mathrm{lr}}\left(\tau\right)$ 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 21 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 $R_{\mathrm{lr}}\left(\tau\right)$ has been estimated ($\tau < 1.5$ s in our case).

The position-pitch map was calculated for each 15 s audio frame for $-\pi < \varphi \leq \pi$ with a resolution of $\frac{\pi}{180}$ rad, and for $20 < f \leq 2\,000$ with a resolution of 1 Hz. This produced a 181×1981 map with shifts in the φ dependent on the orientation of the head-mounted microphone system. In order to reduce the number of dimensions, a bidimensional discrete Fourier transform (2D DFT) was calculated, and only the 4×4 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 ten GMMs with 64 components [4] each. Each GMM modelled the likelihood of one of the scene classes given the feature vector composed by the EMS and the position-pitch features, and corresponding to one 1.5 s. The overall *a posteriori* probability of each class for a 10 s audio segment was estimated by adding up the logarithms of the likelihoods of its frames, assuming that all *a priori* probabilities are equal. For all frames, segments and recordings, the class assigned by the system was estimated to be the class yielding the highest of these *a posteriori* log-probabilities.

5. EXPERIMENTS & RESULTS

The classification experiment corresponding to the baseline evaluation procedure proposed for the acoustic scene classification challenge in DCASE 2019[5] was run. The overall correct classification rate (CCR) for audio segments is 53.86%, while the per-class performance is as indicated in table The confusion matrix corresponding to this experiment is in Tab. 2.

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 [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]. 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.

7. REFERENCES

[1] 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.

Class	**CCR (%)**
Airport	42.76
Indoor shopping mall	58.50
Underground station	51.26
Pedestrian street	49.88
Public square	32.56
Street with medium level of traffic	85.57
Travelling by tram	56.42
Travelling by bus	57.59
Travelling by underground	47.11
Urban park	56.99

Table 2: Per-class correct classification rates

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

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

[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," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[7] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Amer.*, 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," *IEEE Signal Processing Lett.*, vol. 18, no. 2, pp. 130–133, 2011.

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

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