Author Accepted Manuscript
Presented at the Federated Conference on Computer Science and Information Systems
(FedCSIS) in Sep 2018

Feature Extraction of Binaural Recordings for Acoustic Scene Classification

Sławomir K. Zieliński
Faculty of Computer Science, Białystok University of Technology, Białystok, Poland
Email: s.zielinski@pb.edu.pl

Hyunkook Lee
Applied Psychoacoustics Laboratory (APL), University of Huddersfield, Huddersfield, HD1 3DH, United Kingdom
Email: h.lee@hud.ac.uk

Abstract
Binaural technology becomes increasingly popular in the multimedia systems. This paper identifies a set of features of binaural recordings suitable for the automatic classification of the four basic spatial audio scenes representing the most typical patterns of audio content distribution around a listener. Moreover, it compares the five artificial-intelligence-based methods applied to the classification of binaural recordings. The results show that both the spatial and the spectro-temporal features are essential to accurate classification of binaurally rendered acoustic scenes. The spectro-temporal features appear to have a stronger influence on the classification results than the spatial metrics. According to the obtained results, the method based on the support vector machine, exploiting the features identified in the study, yields the classification accuracy approaching 84%.

I. Introduction

Due to a growing popularity of binaural technology [1], large repositories of audio material with binaural sound will soon be created. This will inevitably give rise to challenges concerning the management of spatial audio content. The method proposed in this paper could potentially be used for automatic indexing, search and retrieval of binaural recordings according to their spatial properties, helping to manage future audio repositories.

Most of the studies in the area of acoustic scene classification (ASC) aim to identify an environment where a given scene was recorded [2]-[4]. Little work has been done towards the classification of the recordings according to their spatial characteristics. The key idea underlying this work is, therefore, to extract the features from binaural recordings and to develop a prototype classifier allowing for classification of the spatial properties of acoustic scenes.

Taking advantage from feeding binaural signals to the input of ASC algorithms does not constitute a new approach. Chu et al. developed an environment-aware robotic system equipped with binaural microphones [5]. Trowitzsch et al. demonstrated benefits from using a binaural signal processor for detection of environmental sounds [6]. More recently, such researchers as Han and Park, as well as Weiping at al., exploited binaural signals in their ASC algorithms submitted to the DCASE2017 Challenge [7], [8]. However, to the best of the authors' knowledge, no-one has yet attempted to classify spatial properties of auditory scenes evoked by binaural recordings.

This study extends and builds on the recent work by Zieliński [9]. In contrast to the aforementioned study, which was focused on the classification of five-channel surround sound recordings, the experiment described in this paper was devoted to the classification of binaural audio content.

II. Taxonomy of Basic Spatial Audio Scenes

Information provided at the output of the proposed classifier identifies one of the four basic spatial scenes, labeled as *FB*, *FF*, *BF*, and *BB*. These scenes constitute the typical distribution patterns of foreground and background audio content around the listener in the horizontal plane (see Table I). Foreground sound objects represent easily identifiable, important and clearly perceived audio sources, whereas background objects normally represent reverberant, unimportant, unclear, ambient, "foggy" and distant sound sources. A taxonomy of the acoustic scenes adopted in this study was inspired by Rumsey's simplified spatial audio scene-based paradigm [10].

TABLE I.
THE BASIC SPATIAL AUDIO SCENES

Acoustic Scene	Description
Foreground-Background (*FB*)	A listener perceives foreground audio content in the front and background content behind the head.
Foreground-Foreground (*FF*)	A listener is surrounded by foreground audio content.
Background-Foreground (*BF*)	A listener perceives background audio content in the front and foreground content behind the head.
Background-Background (*BB*)	A listener is surrounded by background audio content.

III. CORPUS OF BINAURAL RECORDINGS

In total 600 binaural recordings were gathered for the purpose of this experiment. Most of the selected excerpts were extracted from the recordings available in the Internet, while 28 recordings, which constitutes 4.7% of all the items, were obtained through a binaural processing of the commercially available 5.0 surround sound recordings. The gathered sound clips represented such recording genres as classical music, pop music, jazz, electronic music, nature, documentary, drama, ambient recordings, and film soundtracks. During the selection procedure, care was taken that each excerpt exemplified a single spatial scene (*FB*, *FF*, *BF* or *BB*). The recordings were annotated manually by the first author. The average duration of the acquired audio samples was equal to 20 seconds. The recordings were stored in uncompressed two-channel audio files with a sampling rate of 44.1kHz and a 16-bit resolution. The available recordings in the audio corpus were split into the two subsets intended for the training (75% of items) and validation purposes (25% of excerpts), respectively.

IV. FEATURE EXTRACTION

In total 1012 features were extracted for the purpose of this study. They could be divided into two broad categories: spatial and spectro-temporal. An overview of the extracted features was given in Table II. The rms-based metrics and binaural cues were classified in this study as spatial features, whereas the spectral features, the Mel-frequency cepstral coefficients (MFCCs) and the discrete cosine transformed amplitude modulation spectrogram coefficients (DCT AMS) were categorized as the spectro-temporal metrics. The procedure used to extract the features was outlined below.

Let x and y denote the left and right ear signals of the binaural recordings, respectively. Some of the metrics were extracted directly from the above signals whereas the other features were calculated based on m and s signals, where $m = x + y$ and $s = x - y$. Prior to calculating the metrics, the signals were split into 20 ms time-frames with a 10 ms overlap. In order to save the computation time the duration of the analyzed time-blocks of the recordings was reduced to 7 seconds.

For each time-frame, a ratio between the rms values of the x and y signals was computed. This way the obtained descriptors constituted a crude approximation of the interaural level differences (*ILD*). Similarly, for every time-frame, a ratio between m and s signals was also calculated. It was assumed by the authors that this ratio could also be considered to be a simple descriptor of spatial characteristics.

All the metrics, including those described in the remainder of the paper, were calculated for every time-frame of the signals. Then, they were summarized using the absolute mean values and standard deviations. In order to account for temporal fluctuations of the rms ratio across the time-frames, the standard *delta* metrics [11] were also computed in a similar way as explained above.

There are three fundamental cues responsible for the spatial perception of sound: interaural level difference (*ILD*), interaural time difference (*ITD*), and interaural coherence (*IC*) [1], [12]. These cues were computed separately for each output of a 40-channel gammatone filter bank using their corresponding rate-maps. The rate-maps constitute a representation of auditory nerve firing rates [13] and are used in ASC algorithms [6]. The standard *delta* metrics [11] were also computed based on the *ILD*, *ITD*, and *IC* cues. The binaural cues were estimated using the publically available software package developed as an auditory front-end of the TWO!EARS system [14].

The following spectral features were included in the study: *centroid*, *spread*, *brightness*, *high-frequency content*, *crest*, *decrease*, *entropy*, *flatness*, *irregularity*, *kurtosis*, *skewness*, *roll-off*, *flux*, and *variation*. They all

constitute the standard metrics commonly used in music information retrieval algorithms [15]. The above spectral features were extracted separately from the *x* and *y* signals. Then, the differences between the obtained spectral descriptors (difference features) were computed for each time-frame. In addition, the same procedure was also applied to the *m* and *s* signals.

Mel-frequency cepstral coefficients (MFCCs) are commonly used in the ASC algorithms as spectral descriptors [4]. In our study, the first 20 coefficients were extracted for the *m* and *s* signals, respectively, and summarized using means and standard deviations. The similar calculations were also performed for the *delta*-MFCC coefficients. Moreover, the same procedure was also applied to the difference values between the MFCC coefficients obtained for the *m* and *s* signals, respectively.

The last group of features included in this study was derived from the amplitude modulation spectrograms (AMSs) [16]. First, the AMSs were calculated for the *m* and *s* signals, respectively. Then, the modulation spectrograms were transformed using the discrete cosine transform (DCT). As a result, for each time frame 600 DCT coefficients were produced. In order to compress the data, only the first 40 coefficients were preserved (the value adjusted during the pilot experiments). Finally, the DCT coefficients were summarized across time-frames using the mean values and standard deviations.

TABLE II.
OVERVIEW OF THE EXTRACTED FEATURES (1012 METRICS IN TOTAL)

	Spatial Features		**Spectro-Temporal Features**		
Feature Acronym	RMS	Binaural Cues	Spectral Features	MFCC	DCT AMS
No. of Features	8	492	112	240	160

V. EXPERIMENTS AND RESULTS

The following five algorithms were selected and compared in terms of their ability to classify the spatial scenes: (1) *k*-nearest neighbors algorithm (*k-nn*), (2) multinomial regression with a least absolute shrinkage and selection operator (*lasso*) [17], (3) random forest, (4) neural network, and (5) support vector machine (*svm*).

The training data consisted of 451 observations and 1012 variables (features). A standard 10-fold cross-validation was performed during the supervised training procedure.

Fig. 1. shows the average classification accuracy results obtained using a single classification algorithm, namely the random forest. The classifier employing a subset of only 8 features based on the rms estimators produced the worst results, with the mean accuracy below 60%. This outcome shows that such simplistic metrics are inadequate, on their own (that is used in isolation from the other features), to reliably discriminate between the audio scenes. Far better results could be obtained by using a set of 492 features based on the binaural cues, with an accuracy reaching approximately 70%. Spectral features (112 metrics), when used on their own, yielded a similar level of accuracy. Slightly better accuracy could be obtained employing solely the MFCC features (240 metrics). DCT-AMS features (160 metrics) used in isolation from the other descriptors produced slightly disappointing results with the accuracy level of approximately 65%. The best classification outcome was obtained by incorporating all the features simultaneously (1012 metrics), yielding a mean classification accuracy of approximately 78%.

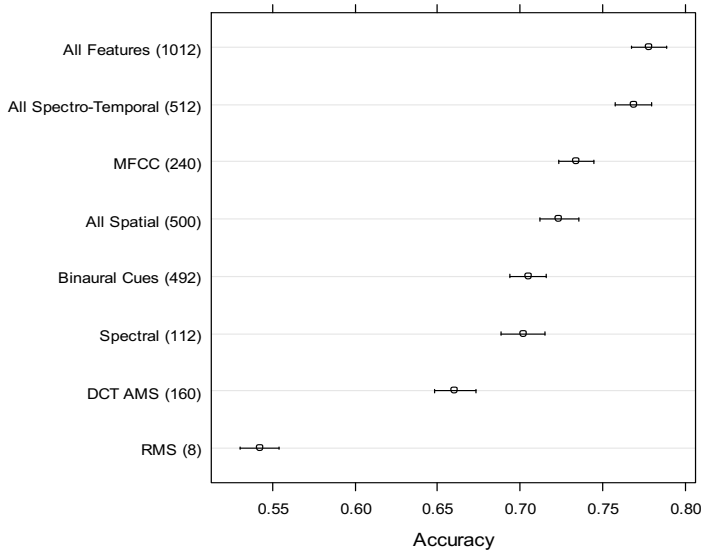

Fig. 1 Classification accuracy obtained using lasso regression for selected groups of features. The results show means and associated 95% confidence intervals. Numbers in brackets denote a quantity of features in each group.

Note that a conglomerate of all the 500 spatial features produced markedly worse results compared to those obtained using the combined group of all the 512 spectro-temporal features. This surprising outcome showed that the spectro-temporal features might be better at discriminating between the spatial scenes than the spatial metrics. This observation was confirmed during the validation test described below.

In order to reduce the risk of overfitting, a backward stepwise selection technique [17] was applied to the test data. An overview of the obtained results, including the accuracy levels, the number of retained features and the values of the model parameters, were presented in Table III. The obtained results show that the best models obtained for the lasso regression method, random forest, and support vector machines produced very similar results, with the accuracy level being equal to approximately 79.8%. The main difference between these models was the number of the selected features. For the lasso regression method, 116 features were selected, whereas for the random forest only 33 metrics were retained. The best model obtained for the support vector machine was based on 490 selected features. The worst outcomes were produced by the neural network and *k*-nn algorithms. The best models selected for each classifier during the feature selection procedure were subsequently used in a validation test.

During the validation test, based on the test dataset, the best classification accuracy results were obtained using the support vector machine (83.89%), followed by the random forest (77.18%), and the lasso regression method (76.51%). The neural network and the method based on the *k*-nearest neighbors produced the worse accuracy results, at the level of 75.17%. The confusion matrix obtained for the support vector machine (the winning method) was presented in Fig. 2. It can be seen that the algorithm could make a particularly good distinction between the *BB* scene and the remaining three scenes (sensitivity of 90.7%).

TABLE III.
OVERVIEW OF THE BEST MODELS OBTAINED THROUGH THE PROCEDURE OF FEATURE SELECTION

Classifier	**Accuracy (%)**	**No. of Features**	**Parameters**
k-nn	73.17	445	$k = 7$
lasso regression	79.84	116	*Alpha* = 0.55 *Lambda* = 6.460145×10^{-3}
random forest	79.81	33	*No. of trees* = 500 *mtry* = 17
neural network	77.64	394	*No. of hidden layers* = 1 *No. of hidden units* = 3 *Weight decay* = 0.1
svm	79.83	490	*Kernel – radial basis function (RBF)* *Sigma* = 1.822721×10^{-3} *Cost* = 1

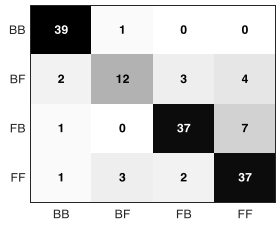

Fig. 2 Confusion matrix for the best classification algorithm (SVM, accuracy 83.89%, 490 features)

VI. DISCUSSION AND CONCLUSIONS

The aim of this study was to identify the features useful for discrimination of the four basic spatial audio scenes of binaural recordings, labeled as *FB*, *BF*, *FF*, and *BB* (see Table I). The obtained results showed that spatial audio scenes could be classified using a mixture of spatial and spectro-temporal metrics with an accuracy exceeding 80%. This outcome indicates that the standard spectro-temporal descriptors combined with the fundamental binaural cues (*ITD*, *ILD*, and *IC*) are adequate for the aforementioned task. Moreover, it provides evidence that the task of spatial audio scene classification may be successfully undertaken without employing a blind source separation algorithm or any other sophisticated techniques aiming to isolate and/or localize audio sources in complex binaural audio scenes. Such an approach could simplify the design of spatial audio scene classifiers.

It was surprising to observe that the spectro-temporal features appeared to have a stronger influence on the classification results than the spatial metrics. This effect, which requires further investigation, could have been caused by an unintended correlation between the spectral and spatial characteristics of the audio recordings used in this study.

Out of the five machine-learning algorithms compared in this study, the support vector machine exhibited the best classification performance, reaching an accuracy of 83.89% upon the validation test. While this result can be considered as satisfactory at this stage of research, there is still scope for improvements. In order to enhance the proposed method, a model accounting for a well-known binaural precedence effect [18] could be incorporated in future studies.

REFERENCES

[1] J. Blauert, *The Technology of Binaural Listening*. Springer, New York, 2013, ch. 1., DOI https://doi.org/10.1007/978-3-642-37762-4

[2] D. Barchiesi, D. Giannoulis, D. Stowell, and M.D. Plumbley, “Acoustic scene classification: Classifying environments from the sounds they produce,” *IEEE Signal Process. Mag.*, vol. 32, pp. 16–34, 2015. DOI https://doi.org/10.1109/msp.2014.2326181

[3] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M.D. Plumbley, “Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 379–393, 2018. DOI https://doi.org/10.1109/taslp.2017.2778423

[4] 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, Oct. 2015. DOI https://doi.org/10.1109/tmm.2015.2428998

[5] S. Chu, S. Narayanan, C.C.J. Kuo, and M. J. Matarić, "Where am I? Scene recognition for mobile robots using audio features," in *Proc. of IEEE International Conference on Multimedia and Expo, IEEE, Toronto*, Canada, July, 2006. DOI https://doi.org/10.1109/icme.2006.262661

[6] I. Trowitzsch, J. Mohr, Y. Kashef, and K. Obermayer, "Robust Detection of Environmental Sounds in Binaural Auditory Scenes," *IEEE Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1344−1356, 2017. DOI https://doi.org/10.1109/taslp.2017.2690573

[7] Y. Han and J. Park, "Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification," *Workshop on Detection and Classification of Acoustic Scenes and Events*, Munich, Germany, November, 2017.

[8] Z. Weiping, Y. Jiantao, X. Xiaotao, L.Xiangtao and P. Shaohu, "Acoustic Scene Classification Using Deep Convolutional Neural Network and Multiple Spectrograms Fusion," *Workshop on Detection and Classification of Acoustic Scenes and Events*, Munich, Germany, November, 2017.

[9] S.K. Zieliński, "Feature extraction of surround sound recordings for acoustic scene classification," In: Rutkowski L., Scherer R., Korytkowski M., Pedrycz W., Tadeusiewicz R., Zurada J. (eds) Artificial Intelligence and Soft Computing. ICAISC 2018. Lecture Notes in Computer Science, vol. 10842. Springer. DOI https://doi.org/10.1007/978-3-319-91262-2_43

[10] F. Rumsey, "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," *J. Audio Eng. Soc.*, vol. 50, pp. 651−666, 2002.

[11] L. Rabiner, B.-H. Juang, B. Yegnanarayana, *Fundamentals of Speech Recognition*, Pearson Education, 2008.

[12] J. Blauert, *Spatial Hearing. The Psychophysics of Human Sound Localization*. The MIT Press, London, 1996, ch. 3.

[13] G.J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[14] A. Raake *et al.*, "Two!ears—Integral interactive model of auditory perception and experience," *Proc. DAGA*, 2014.

[15] G. Peeters, B.Giordano, P. Susini, N. Misdariis, and S. McAdams, The Timbre Toolbox: Extracting audio descriptors from musical signals. *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 2902–2916, 2011. DOI https://doi.org/10.1121/1.3642604

[16] T. May, and T. Dau, "Computational speech segregation based on an auditory-inspired modulation analysis," *J. Acoust. Soc. Am.*, vol. 136, no. 6, pp. 3350–3359, 2014. DOI https://doi.org/10.1121/1.4901711

[17] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, London, 2017, ch. 6.

[18] A.D. Brown, G.C. Stecker, and D.J. Tollin, "The Precedence Effect in Sound Localization," *J. Assoc. Res. Otolaryngol.*, vol. 16, no. 1, pp. 1–28, 2015. DOI https://doi.org/10.1007/s10162-014-0496-2