DOI: 10.61782/fa.2023.0713

forum acusticum 2023

EXPLORING ANNOYANCE IN A SOUNDSCAPE CONTEXT BY JOINT PREDICTION OF SOUND SOURCE AND ANNOYANCE

Yuanbo Hou[1*] **Andrew Mitchell** [2] **Qiaoqiao Ren**[3]
Francesco Aletta[2] **Jian Kang**[2] **Dick Botteldooren**[1*]
[1] WAVES Research Group, Department of Information Technology, Ghent University, Belgium
[2] Institute for Environmental Design and Engineering, University College London, UK
[3] Department of Electronics and Information Systems, Ghent University - IMEC, Belgium

ABSTRACT

Soundscape, the sonic environment as perceived and understood by people, is a conglomerate of different sounds. It has been established that its appraisal by instantaneous annoyance is not solely determined by its calculated loudness, but also by recognised sounds. Hence, most previous research on annoyance has focused on single-source environments. Audio analytics aims at detecting and classifying sound sources, but does not explore human perception of these. This paper proposes a dual-input model to simultaneously perform sound source classification (SSC) and human annoyance rating prediction (ARP). The model takes mel features and root-mean-square value (rms) features as input, and uses convolutional blocks to extract high-level acoustic features. These are used to predict sound source classes and to estimate the human annoyance rating for the whole fragment. Experiments on the DeLTA dataset show that: 1) models using mel features and rms features outperform models using only one of them; 2) The proposed model achieves a SSC accuracy of 90.06%, and an ARP (scale 1 to 10) root mean square error of 1.05.

Keywords: *sound source classification, annoyance prediction, convolutional neural networks*

**Corresponding author*: *Yuanbo.Hou@UGent.be, Dick.Botteldooren@UGent.be.*

1. INTRODUCTION

When environmental sound disturbs instantaneous intentions and activities, it becomes noise and is appraised as being annoying. This definition of annoyance is today used to explore the long-term effects of environmental noise through surveys (ISO/TS 15666:2021). In this context, annoyance is often related to a specific source, e.g. being annoyed by the sound of X. It is worth noting that annoyance in this context refers to long-term appraisal and therefore is the result of non-focused listening during other activities. Noticing the sound can be seen as an important prerequisite for becoming annoyed [1]. Lab research on annoyance - in this definition - has to carefully avoid that participants' change to a focused listening style, e.g. by creating an ecologically valid setting [2] [3].

Soundscape research takes a more holistic approach to perception and understanding of the sonic environment (ISO 12913-1:2014). The combination of sounds that together form the sonic environment becomes important. Appraisal of the soundscape can be done by focused listening, preferably in context (ISO/TS 12913-2:2018). As it is well known that the sounds that are heard affect the appraisal of a soundscape, the standard foresees asking users of the space about the sounds that they hear. In addition, perceived affective quality is assessed along 8 dimensions, roughly matching the circumflex model of effect. One of these dimensions is annoying, and it is precisely this definition of annoyance that is used in this work.

Sound source classification (SSC) has been used for audio event recognition [4], acoustic scene classification [5], and monitoring [6]. In this work, we will augment it with annoyance rating prediction (ARP), which aims to predict the overall appraisal of the soundscape along the

annoyance axis. As both sound recognition and perceived affective quality are suggested as soundscape descriptors by the standard, joint prediction of SSC and ARP can help in the design of a friendly soundscape [7] and the construction of smart cities [8].

In real life, the song of birds in the park usually makes people feel relaxed, while the horn roar of speeding cars on the road usually makes people feel annoyed. To accurately identify these diverse audio events, deep learning-based convolutional neural networks (CNN) [9], convolutional recurrent neural networks (CRNN) [10] and Transformer [11] [12] with multi-head attention are proposed to identify different kinds of real-world audio events, which are from different sound sources. These SSC-related studies mostly focus on the category identification of sound sources. The feeling brought by these sound sources in the soundscape to humans is ignored. In this paper, we explore the possibility of simultaneously performing SSC and ARP tasks based on real-life polyphonic audio clips.

The paper is organized as follows. Section 2 introduces the proposed method. Section 3 describes the dataset, experimental setup, and analyzes results. Section 4 draws conclusions.

2. METHOD

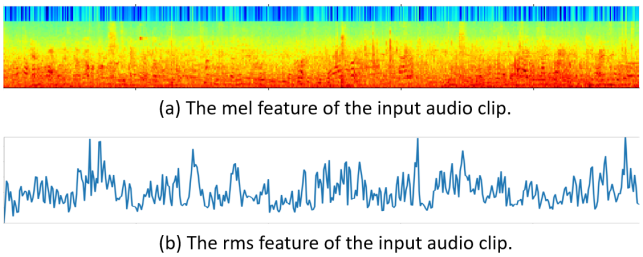
This section introduces the proposed model: the dual-input convolutional neural network (DCNN).

For the input of DCNN, since the log mel features and root-mean-square value (rms) features are used in this paper, there are two branches of inputs to the proposed DCNN model. The dual-input model uses four convolutional blocks to extract high-level representations of the two acoustic features separately. The representations of mel and rms features generated by the convolution block are fed to the fusion module to generate the embeddings of the human annoyance ratings. The embeddings of the human annoyance ratings will be input into the final annoyance rating prediction layer to complete the ARP task. The embeddings of sound sources are fed into the final sound source classification layer to complete the SSC task.

3. EXPERIMENTS AND RESULTS

3.1 Dataset, Experiments Setup, and Metric

In this paper, the publicly available DeLTA [13] dataset, which includes both ground-truth sound source labels and human annoyance rating scores, is used. Each audio clip in DeLTA has a clip-level 24-dimensional multi-hot vector as the sound source label, and an annoyance rating (continuously from 1 to 10). DeLTA comprises 2890 15-second binaural audio clips, where the training, validation, and test sets contain 2081, 231, and 578 audio clips, respectively. The log-magnitude Mel-filter 64-bank spectrogram [14] and frame-level root-mean-square value [15] are used as the acoustic features in this paper.



(a) The mel feature of the input audio clip.

(b) The rms feature of the input audio clip.

Figure 1: The mel feature and rms feature of the same input audio clip.

For SSC, accuracy (*Acc*) is used to measure the classification results. For ARP, root mean square error (*RMSE*) is used to evaluate the prediction results.

3.2 Results and Analysis

Two kinds of acoustic features are used in this paper, Tab. 1 shows the ablation experiments of the two acoustic features on the proposed DCNN model to specifically present the performance of the DCNN model based on different features.

As shown in Tab. 1, the DCNN model performs the worst on the ARP and SSC tasks when using only the rms features related to sound loudness. Moreover, its SSC results are the worst in Tab. 1. As shown in Fig. 1 (b), the dimension of the frame-level rms used in this paper is (T, 1), where T is the number of frames. In other words, the loudness-related rms can be considered as a

Table 1: Ablation study on the acoustic features.

#	Acoustic feature		ARP	SSC
	mel	rms	*RMSE*	*Acc. (%)*
1	✔	✗	1.18 ± 0.12	89.11 ± 1.06
2	✗	✔	1.27 ± 0.10	79.09 ± 2.31
3	✔	✔	**1.05** ± 0.13	**90.06** ± 1.38

one-dimensional feature. Compared with the mel features with a dimension of (T, 64) in Fig. 1 (a), the information contained in the loudness-related rms is slightly scarce. These factors make it difficult to distinguish the 24 different types of sound sources from real-life sources in the DeLTA dataset based only on the rms features alone. The DCNN using mel features outperforms the results of its corresponding rms features overall. While DCNN combining mel and rms features achieves the best results, which clarifies that using these two acoustic features benefits the model's performance on SSC and ARP tasks.

4. CONCLUSION

Previous soundscape research based on human listening tests and questionnaires is often time-consuming and expensive. This paper explores the feasibility of using the deep learning-based model to recognize sound sources and infer human perception within soundscapes without human questionnaires. The successful identification of various sound sources in the soundscape and the prediction of human annoyance ratings by the model indicate that the automatic analysis of soundscapes based on artificial intelligence deep neural networks is promising.

In detail, this paper extends the environmental sound source classification (SSC) task with real-life soundscape annoyance rating prediction (ARP). As soundscape affective rating relates to the recognition of sounds, both tasks are intermingled. Hence, to simultaneously perform SSC and ARP tasks, this paper proposes a dual-branch convolutional neural network (DCNN) using mel features and rms features. Experimental results on the DeLTA dataset show that the proposed DCNN using mel features and rms features outperform models using only one of them, and the proposed model achieves a SSC accuracy of 90.06%, and an ARP (scale 1 to 10) root mean square error of 1.05. It should however be noted that the DeLTA dataset is based on attentive listening to short sound recordings which may have triggered participants to relate their annoyance rating more to the sounds they recognized as they would have done in an ecologically valid context. This is expected to make the task of the model easier.

5. ACKNOWLEDGEMENTS

The WAVES Research Group received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. The UCL authors were supported by UCL Health of the Public, via the Small Grants Scheme 2020–2021, project: "Deep Learning Techniques for Noise Annoyance Detection" (DeLTA).

6. REFERENCES

[1] B. De Coensel, D. Botteldooren, T. De Muer, B. Berglund, M. E. Nilsson, and P. Lercher, "A model for the perception of environmental sound based on notice-events," *The Journal of the Acoustical Society of America*, vol. 126, no. 2, pp. 656–665, 2009.

[2] B. De Coensela, D. Botteldoorena, B. Berglund, M. E. Nilsson, T. De Muer, and P. Lercher, "Experimental investigation of noise annoyance caused by high-speed trains," *Acta Acustica united with Acustica*, vol. 93, no. 4, pp. 589–601, 2007.

[3] K. Sun, B. De Coensel, G. M. E. Sanchez, T. Van Renterghem, and D. Botteldooren, "Effect of interaction between attention focusing capability and visual factors on road traffic noise annoyance," *Applied Acoustics*, vol. 134, pp. 16–24, 2018.

[4] J. Ren, X. Jiang, J. Yuan, and N. Magnenat-Thalmann, "Sound-event classification using robust texture features for robot hearing," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 447–458, 2016.

[5] Y. Hou, B. Kang, W. Van Hauwermeiren, and D. Botteldooren, "Relation-guided acoustic scene classification aided with event embeddings," in *Proc. of International Joint Conference on Neural Networks*, pp. 1–8, 2022.

[6] R. T. Buxton, M. F. McKenna, M. Clapp, E. Meyer, E. Stabenau, L. M. Angeloni, K. Crooks, *et al.*, "Efficacy of extracting indices from large-scale acoustic recordings to monitor biodiversity," *Conservation Biology*, vol. 32, no. 5, pp. 1174–1184, 2018.

[7] E. Margaritis, J. Kang, F. Aletta, and Ö. Axelsson, "On the relationship between land use and sound sources in the urban environment," *Journal of Urban Design*, vol. 25, no. 5, pp. 629–645, 2020.

[8] E.-L. Tan, F. A. Karnapi, L. J. Ng, K. Ooi, and W.-S. Gan, "Extracting urban sound information for residential areas in smart cities using an end-to-end iot system," *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 14308–14321, 2021.

[9] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[10] Y. Hou, Q. Kong, J. Wang, and S. Li, "Polyphonic audio tagging with sequentially labelled data using crnn with learnable gated linear units," in *Proceedings of the DCASE Workshop*, pp. 78–82, November 2018.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, *et al.*, "Attention is all you need," in *Proc. of International Conference on Neural Information Processing Systems*, pp. 5998–6008, 2017.

[12] Y. Hou, Z. Liu, B. Kang, Y. Wang, and D. Botteldooren, "CT-SAT: Contextual transformer for sequential audio tagging," in *INTERSPEECH*, pp. 4147–4151, 2022.

[13] A. Mitchell, M. Erfanian, C. Soelistyo, T. Oberman, J. Kang, R. Aldridge, J.-H. Xue, and F. Aletta, "Effects of soundscape complexity on urban noise annoyance ratings: A large-scale online listening experiment," *International Journal of Environmental Research and Public Health*, vol. 19, no. 22, p. 14872, 2022.

[14] A. Bala, A. Kumar, and N. Birla, "Voice command recognition system based on MFCC and DTW," *International Journal of Engineering Science and Technology*, vol. 2, no. 12, pp. 7335–7342, 2010.

[15] M. Mulimani and S. G. Koolagudi, "Acoustic event classification using spectrogram features," in *TENCON 2018-2018 IEEE Region 10 Conference*, pp. 1460–1464, 2018.