Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09430-3

# Emotional quantification of soundscapes by learning between samples

Stavros Ntalampiras[1]

Received: 9 March 2020 / Revised: 7 July 2020 / Accepted: 27 July 2020 /
Published online: 15 August 2020

## Abstract

Predicting the emotional responses of humans to soundscapes is a relatively recent field of research coming with a wide range of promising applications. This work presents the design of two convolutional neural networks, namely ArNet and ValNet, each one responsible for quantifying arousal and valence evoked by soundscapes. We build on the knowledge acquired from the application of traditional machine learning techniques on the specific domain, and design a suitable deep learning framework. Moreover, we propose the usage of artificially created mixed soundscapes, the distributions of which are located between the ones of the available samples, a process that increases the variance of the dataset leading to significantly better performance. The reported results outperform the state of the art on a soundscape dataset following Schafer's standardized categorization considering both sound's identity and the respective listening context.

**Keywords** Acoustic ecology · Audio signal processing · Afffective computing

## 1 Introduction

The field aiming at assessing the emotional content of generalized sounds including speech, music and sound events is attracting the interest of an ever increasing number of researchers [12, 15–17, 21, 25]. However, there is still a gap regarding works addressing the specific case of soundscapes, i.e. the combination of sounds forming an immersive environment [20]. Soundscape emotion prediction (SEP) focuses on the understanding of the emotions perceived by a listener of a given soundscape. These may comprise the necessary stimuli for a receiver to manifest different emotional states and/or actions, for example, one may feel joyful in a natural environment. Such contexts demonstrate the close relationship existing between soundscapes and the emotions they evoke, i.e., soundscapes may cause emotional manifestations on the listener side, such as joy. That said, SEP can have a significant impact

✉ Stavros Ntalampiras
name.surname@unimi.it

[1] University of Milan, via Celoria 18, Milan, Italy

in a series of application domains, such as sound design [18, 22], urban planning [3, 24], and acoustic ecology [4, 11], to name but a few.

Affective computing has received a lot of attention [9] in the last decades with a special focus on the analysis of emotional speech, where a great gamut of generative and discriminative classifiers have been employed [21, 28], and music [7, 26] where most of the research is concentrated on regression methods. The literature analyzing the emotional responses to soundscape stimuli includes mainly surveys requesting listeners to characterize them. The work described in [1] details such a survey aiming to analyze soundscapes categorized as technological, natural or human. Davies et al. [3] provide a survey specifically designed to assess various emotional aspects of urban soundscapes. Another survey is described in [2] aiming at quantifying the relationship between pleasantness and environmental conditions. Moving on, the literature includes a limited amount of methods focused on the automatic emotional labeling of soundscapes. Among those, Fan et al. [5] employed a support vector regression scheme fed on a wide range of handcrafted features to assess the emotional characteristics of six classes of soundscapes. In their follow-up work [6] the authors used both handcrafted features and deep nets, boosting the achieved performance. Another framework was developed in [10] based on the bag-of-frames approach using handcrafted features and two support vector machines each one responsible for predicting the pleasantness and eventfulness of 77 soundscapes.

The main limitations of the related literature can be identified in the usage of extensive feature engineering which heavily depends on domain knowledge and poor data availability. This work proposes a deep learning framework for the automatic assessment of the emotional content of soundscapes. Addressing the existing limitations, the framework's novel aspects are *a*) relaxing the handcrafted features restriction, *b*) introduction of two convolutional neural networks (ArNet and ValNet) each one carrying out prediction of arousal and valence of soundscapes, and *c*) conceptualization and development of the between-sample learning scheme able to meaningfully augment the available feature space. The dataset includes soundscapes coming from six classes, i.e. *a*) natural, *b*) human, *c*) society, *d*) mechanical, *e*) quiet, and *f*) indicators following Schafer's organization [20]. After a thorough experimental campaign, we analyze the performance boosting offered by the between-sample learning scheme, while the reported results surpass the state of the art.

The rest of this paper is organized as follows: Section 2 analyzes the proposed between-samples learning paradigm including the entire pipeline. Section 3 presents the experimental set-up and results, while in Section 4, we draw our conclusions.

## 2 The between-samples learning paradigm

This section details the method used for predicting of the emotional evoked by a soundscape. The proposed method, demonstrated in Fig. 1, mixes sounds coming from multiple classes and complements the training set with the generated samples. Subsequently, the log-Mel spectrum is extracted which is fed to a convolutional neural network carrying out modeling and emotional quantification.

Initially, we briefly analyze the feature set and the regression algorithm, while the emphasis is placed on the way the learning is performed between the available samples.

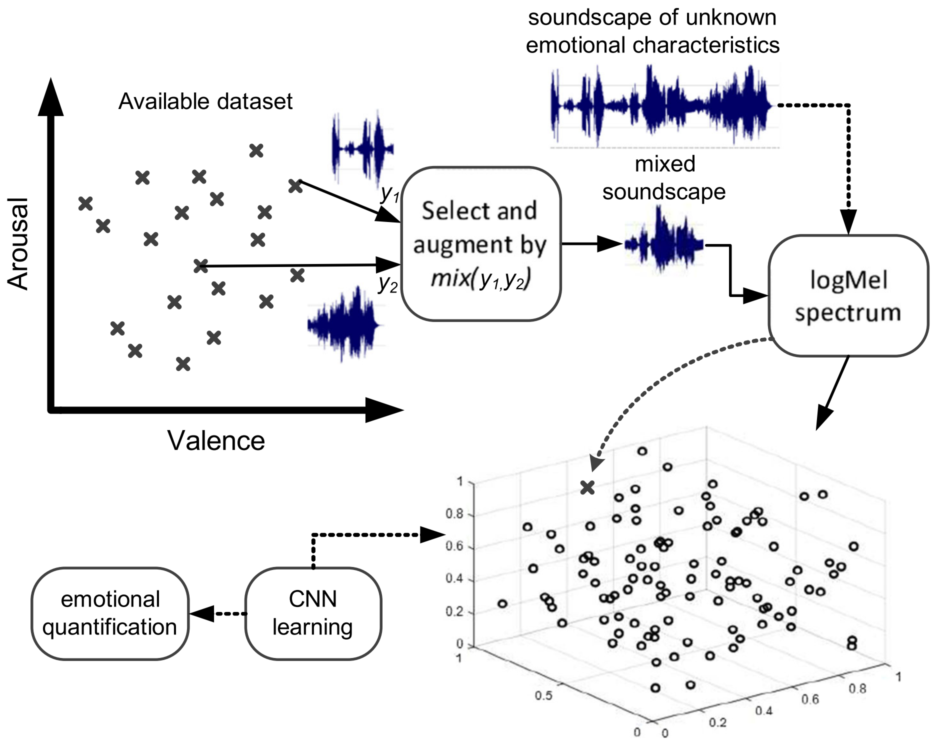

**Fig. 1** The proposed method starts by soundscape mixing, proceed with feature extraction and finally models the feature space using CNNs

## 2.1 Feature set

The present feature set is a simplification of the Mel-Frequency Cepstral Coefficients where the final dimensionality reduction step based on the discrete cosine transform is omitted [13, 14, 27]. To this end, we employed a triangular Mel scale filterbank for extracting 23 log-energies. Firstly, the audio signal is windowed and the short-time Fourier transform (STFT) is computed. The outcome of the STFT passes though the filterbank and the logarithm is computed to adequately space the data. It is worth noting that the usage of such a standardized feature extraction mechanism removes the need to conceptualize and implement handcrafted features specifically designed to address the given problem.

## 2.2 Convolutional neural network architecture

The structure of the proposed CNN was determined during early experimentations and is shown in Table 1. Starting from the standard, multilayer perceptron model, a CNN includes simple but relevant modifications. Commonly, a CNN is composed by a number of stacked

**Table 1** The structure of ArNet and ValNet (# of parameters: 2,674,721)

| Layer | Output shape | # of Parameters |
|---|---|---|
| Conv2D | (148,148,32) | 896 |
| MaxPooling2 | (74,74,32) | 0 |
| Conv2D | (72,72,64) | 18496 |
| MaxPooling2 | (36,36,64) | 0 |
| Flatten | 82944 | 0 |
| Dropout | 82944 | 0 |
| Dense | 32 | 265424 |
| Dense | 32 | 1056 |
| Dense | 1 | 33 |

layers forming a deep topology. Here, we consider two convolutional layers each one followed by a max-pooling operation. The convolutional layers organize the hidden units so that local structures are revealed in the 2-d plane and subsequently exploited. This is accomplished by connecting each hidden unit to only a small portion, so-called receptive field, of the input space (e.g. 4 × 4 pixel blocks). In essence, the weights of such units form filters (also called convolutional kernels) applied to the entire input plane and thus, extracting a feature map. At this point we make the assumption that such locally extracted features are useful in other parts of the input plane, thus the same weights are applied on its entirety. This assumption is highly important since not only it minimizes the number of trainable parameters but it also renders the network indifferent to translational shifts of the input data [19]. The max-pooling layers carry out further dimensionality reduction by merging adjacent units and retaining their maximum value, a process which boosts translational indifference. Rectified Linear Units (ReLUs) are employed with the activation function being $f(x) = max(0, x)$. ReLUs dominate the current literature as they tend to offer *a*) faster gradient propagation than conventional units (logistic sigmoid, hyperbolic tangent, etc.), *b*) biological plausibility, and *c*) an activation form characterized by high sparsity [8].

The network is completed by a flattening, a dropout and three densely connected layers responsible for the regression process. The dropout layer helps to avoid overfitting due to the large number of parameters in need of estimation (2,674,721). The specific layer randomly removes 50% of the present hidden units. In general, such an operation removes irrelevant relationships and secures that the learned filters are able to provide reliable modeling and in the present study, SEP.

### 2.3 Generating samples between the original ones

The dataset chosen in the current study follows the widely accepted Schafer's soundscape taxonomy [5, 20] based on the referential meaning of environmental sounds. In Schafer's work, the grouping criterion is the identity of the sound source and the listening context without taking into account audio features. Interestingly, the Emo-Soundscapes corpus described in [5] includes 600 clips equally distributed among the six classes proposed by Schafer, i.e. *a*) natural, *b*) human, *c*) society, *d*) mechanical, *e*) quiet, and *f*) sounds as indicators.

The second part of the Emo-Soundscapes corpus includes mixed soundscapes coming from these six classes. The initial idea was to study the emotional impact of sound design; however, this work shows that such mixed soundscapes are useful to predict the emotions perceived both by single- and multi-class soundscapes. Each mix is designed so that it includes content coming from either two or three audio clips. Interestingly, there is no restriction during class selection meaning that a certain mixing can include audio clips belonging to the same class. The duration of each clip is 6 seconds which suffices for annotators to efficiently characterize its emotional content in terms of arousal and valence.

In mixed sounds, humans can understand the existence of more than one classes, perceive the one dominating the mixture, etc. quite effortless. Thus, a point placed within the limits of the entire feature distribution must have a meaningful semantic correspondence, while this is not necessarily true for points outside the distribution. Feature distributions characterizing mixed sounds are expected to be located between the distributions of the sounds composing the mixture. At the same time, the mixed variance is proportional to the original feature distributions similarly to the classification problematic described in [23].

Figure 2 demonstrates how the mixing affects the feature distributions. More specifically, two cases are shown: a) mixing of two classes (mechanical and human) and b) mixing of three classes (mechanical, human, and nature). In both cases, we observe that the vast

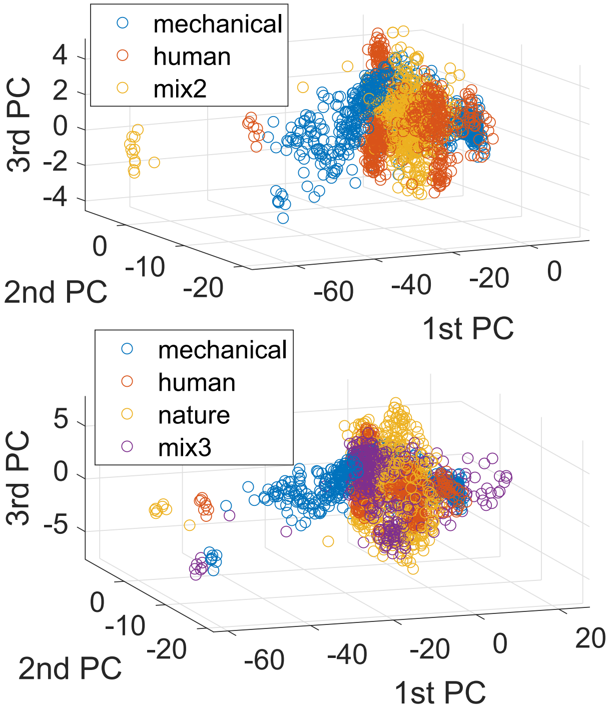

**Fig. 2** The feature space demonstrating the cases of mixing samples coming from 2 (top) and 3 classes (bottom)

majority of the principal components of the mixed feature vectors lies within the principal components extracted out of the single classes. Figure 3 demonstrates ArNet's intermediate activations showing how single (top-row), mixture of two (middle row), and mixture of three (bottom row) soundscapes are decomposed unto the different filters learned by the network.

## 3 Experimental set-up and results

This section includes details regarding the dataset, the parameterization of the proposed approach, performance analysis as well as how it compares with the state of the art.

### 3.1 Dataset

Up until recently, there was a gap as regards to a dataset including emotionally-annotated soundscapes. The work presented in [5] covered this gap by designing and making available to the scientific community, the Emo-soundscapes dataset. It facilitates soundscape emotion recognition tasks by the study of single as well as mixed soundscapes. As mentioned in Section 2.3 the dataset follows Schafer's organization (human, nature, indicators, ) and includes 1213 6-second Creative Commons licensed audio clips. The annotation of the perceived emotion was carried out by means of a crowdsourcing listening experiment. They

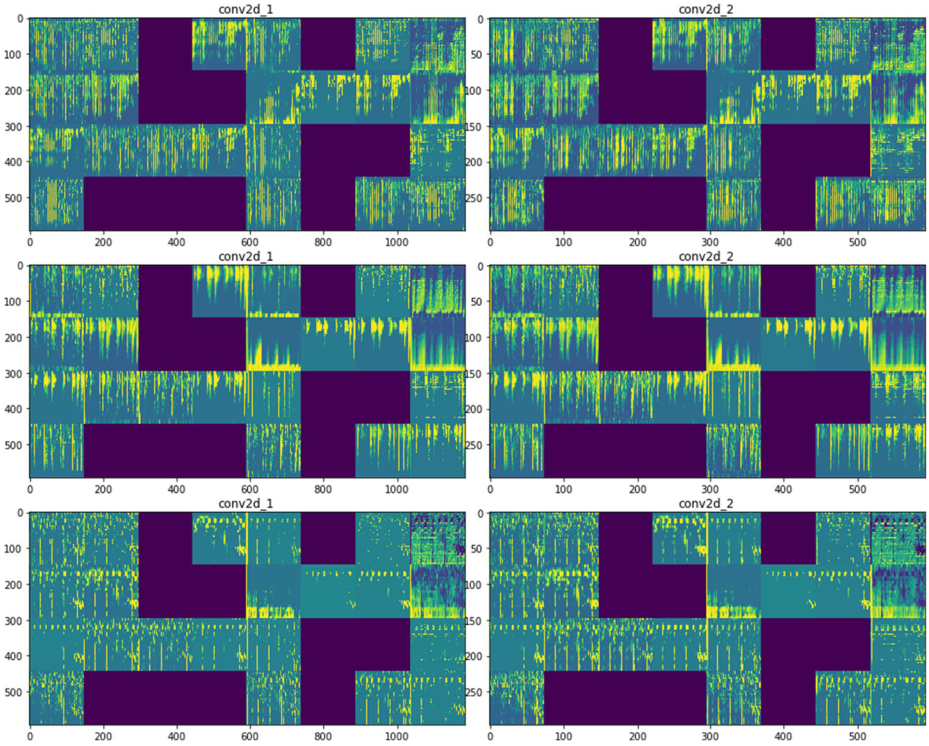

**Fig. 3** ArNet's intermediate activations showing how single (top-row), mixture of two (middle row), and mixture of three (bottom row) soundscapes are decomposed unto the different filters learned by the network

recorded both valence and arousal perceived by 1182 annotators from 74 different countries. Detailed information regarding the dataset and its annotation is available in [5].

### 3.2 Parameterization

The log-mel spectrogram was extracted out of windows is 30 ms with 10 ms overlap. The sampled data are hamming windowed to smooth potential discontinuities, while the FFT size is 512. The CNN operates on a receptive field as in Table 1, the activation function is ReLU, while two networks were trained to model valence and arousal respectively, i.e. ValNet and ArNet. The training process terminated after 100 epochs at a learning rate of 0.01. Each network is trained on minimizing the mean squared error.

### 3.3 Results

We followed the experimental protocol described in [5] where the dataset is randomly divided $n$ times into training and testing data with a ratio 4:1 and $n = 10$. At each iteration there is no overlap between the training and testing sets while the reported mean square errors are the averages over $n$.

The achieved results are summarized in Fig. 4. As we can see, the CNN trained only on the original single-class sounds surpasses the state of the art MSEs w.r.t valence prediction, while the arousal one lies at similar levels. However, the proposed method employing sound mixtures outperforms the other methods significantly at both valence and arousal prediction. The final MSEs are 0.0168 and 0.0107 for valence and arousal respectively. The bottom part of Fig. 4 demonstrates the way these results vary as per Schafer's categorization. A similar behavior is observed for the majority of the classes. Best valence prediction is achieved for the quiet class while the indicators one is the hardest to predict. Best arousal prediction is achieved for the nature class and the worst for the society class. In general, sounds coming from the society, mechanical, and indicators classes provide the highest MSEs, i.e. worst performance, which may be due to the respective intra-class variability as it can be assessed by a human listener. Following the analysis provided in Section 2.3, we see how the variance offered by mixed samples boosts network's prediction capabilities. Overall, the method based on learning between samples provided excellent results and surpasses the state of art in emotional quantification of soundscapes.

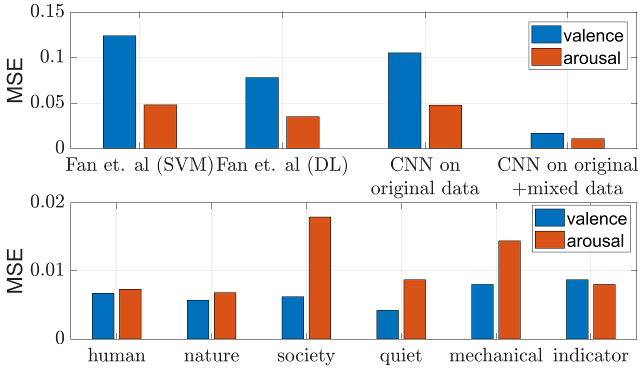

**Fig. 4** The results achieved by the proposed approach and how it compares with the state of the art (top) as well as how they vary per Schafer's categorization (bottom)

## 4 Conclusions

This work presented a deep learning framework achieving SEP able to surpass the state of the art based on handcrafted features and traditional machine learning algorithms. Interestingly, the accompanying module carrying out between samples learning manages to significantly boost the prediction performance.

In the future, we wish to evaluate the usefulness and practicality of an emotional space formed not only by soundscapes but incorporating generalized sound events, music, and speech. Such a jointly created space may offer improved prediction in multiple applications domains. To this end, we intent to exploit transfer learning technologies [12] forming a synergistic framework able to incorporate and transfer knowledge coming from multiple domains favoring diverse applications, such as music information retrieval, bioacoustic signal processing, etc.

**Acknowledgements** We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

**Funding Information** Open access funding provided by Università degli Studi di Milano within the CRUI-CARE Agreement.

## References

1. Berglund B, Nilsson M, Axelsson S (2007) Soundscape psychophysics in place. 6, 3704–3711. Proc. Inter-Noise 2007 2007(p.):IN07114
2. Brocolini L, Waks L, Lavandier C, Marquis-Favre C, Quoy M, Lavandier M (2010) Comparison between multiple linear regressions and artificial neural networks to predict urban sound quality
3. Davies W, Adams M, Bruce N, Cain R, Jennings P, Carlyle A, Cusack P, Hume K, Plack C (2009) A positive soundscape evaluation system
4. Drossos K, Floros A, Giannakoulopoulos A, Kanellopoulos N (2015) Investigating the impact of sound angular position on the listener affective state. IEEE Trans Affect Comput 6(1):27–42. https://doi.org/10.1109/TAFFC.2015.2392768
5. Fan J, Thorogood M, Pasquier P (2017) Emo-soundscapes: a dataset for soundscape emotion recognition. In: 2017 Seventh international conference on affective computing and intelligent interaction (ACII), pp 196–201. https://doi.org/10.1109/ACII.2017.8273600
6. Fan J, Tung F, Li W, Pasquier P (2018) Soundscape emotion recognition via deep learning. Sound and Music Computing Conference
7. Fukayama S, Goto M (2016) Music emotion recognition with adaptive aggregation of gaussian process regressors. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 71–75
8. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: Gordon G, Dunson D, Dudík M (eds) Proceedings of the fourteenth international conference on artificial intelligence and statistics, proceedings of machine learning research, vol 15. PMLR, Fort Lauderdale, FL, USA, pp 315–323. http://proceedings.mlr.press/v15/glorot11a.html

9. Kim BH, Jo S (2020) Deep physiological affect network for the recognition of human emotions. IEEE Trans Affect Comput 11(2):230–243
10. Lundén P, Axelsson Ö, Hurtig M (2016) On urban soundscape mapping : a computer can predict the outcome of soundscape assessments
11. Moscoso P, Peck M, Eldridge A (2018) Emotional associations with soundscape reflect human-environment relationships. J Ecoacoustics 2, YLFJ6Q. https://doi.org/10.22261/jea.ylfj6q
12. Ntalampiras S (2017a) A transfer learning framework for predicting the emotional content of generalized sound events. J Acoust Soc Am 141(3):1694–1701. https://doi.org/10.1121/1.4977749
13. Ntalampiras S (2017b) Hybrid framework for categorising sounds of mysticete whales. IET Signal Process 11(4):349–355. https://doi.org/10.1049/iet-spr.2015.0065
14. Ntalampiras S (2019) Automatic acoustic classification of insect species based on directed acyclic graphs. J Acoust Soc Am 145(6):EL541–EL546. https://doi.org/10.1121/1.5111975
15. Ntalampiras S (2020) Toward language-agnostic speech emotion recognition. J Audio Eng Soc 68(1/2):7–13. https://doi.org/10.17743/jaes.2019.0045
16. Ntalampiras S, Arsic D, Stormer A, Ganchev T, Potamitis I, Fakotakis N (2009) Prometheus database: A multimodal corpus for research on modeling and interpreting human behavior. In: 2009 16th International conference on digital signal processing, pp 1–8
17. Ntalampiras S, Avanzini F, Ludovico LA (2019) Fusing acoustic and electroencephalographic modalities for user-independent emotion prediction. In: 2019 IEEE International conference on cognitive computing (ICCC), pp 36–41. https://doi.org/10.1109/ICCC.2019.00018
18. Ntalampiras S, Potamitis I (2019) A statistical inference framework for understanding music-related brain activity. IEEE J Sel Topics Signal Process 13(2):275–284. https://doi.org/10.1109/JSTSP.2019.2905431
19. Piczak KJ (2015) Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP), pp 1–6. https://doi.org/10.1109/MLSP.2015.7324337
20. Schafer R (1993) The soundscape: our sonic environment and the tuning of the world. Inner Traditions/Bear. https://books.google.it/books?id=-FsoDwAAQBAJ
21. Schuller BW (2018) Speech emotion recognition. Commun ACM 61(5):90–99. https://doi.org/10.1145/3129340
22. Thorogood M, Pasquier P (2013) Computationally generated soundscapes with audio metaphor. https://doi.org/10.13140/2.1.4191.0084
23. Tokozume Y, Ushiku Y, Harada T (2017) Learning from between-class examples for deep sound recognition. arXiv:1711.10282
24. Ward T (2014) The soundscapes landscapes project: sound and video art in an urban installation. In: 2014 International conference on interactive mobile communication technologies and learning (IMCL2014), pp 265–268. https://doi.org/10.1109/IMCTL.2014.7011145
25. Weninger F, Eyben F, Schuller BW, Mortillaro M, Scherer KR (2013) On the acoustics of emotion in audio: what speech, music, and sound have in common. Front Psychol 4. pp 1–12 https://doi.org/10.3389/fpsyg.2013.00292
26. Yang Y-H, Chen HH (2012) Machine recognition of music emotion: a review. ACM Trans Intell Syst Technol 3(3):40:1–40:30. https://doi.org/10.1145/2168752.2168754
27. Zhang C, Wang P, Guo H, Fan G, Chen K, Kämäräinen J-K (2017) Turning wingbeat sounds into spectrum images for acoustic insect classification. Electron Lett 53(25):1674–1676. https://doi.org/10.1049/el.2017.3334
28. Zhang Z, Hong W, Li J (2020) Electric load forecasting by hybrid self-recurrent support vector regression model with variational mode decomposition and improved cuckoo search algorithm. IEEE Access 8:14642–14658

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.