inter.noise
21-24 AUGUST
SCOTTISH EVENT CAMPUS
GLASGOW 2022

Artificial intelligence-based collaborative acoustic scene and event classification to support urban soundscape analysis and classification

Yuanbo Hou[1], Dick Botteldooren[2]
Ghent University
Research group WAVES, Department of information Technology, Technologiepark 126, 9052 Gent, Belgium

ABSTRACT

A human listener embedded in a sonic environment will rely on meaning given to sound events as well as on general acoustic features to analyse and appraise its soundscape. However, currently used measurable indicators for soundscape mainly focus on the latter and meaning is only included indirectly. Yet, today's artificial intelligence (AI) techniques allow to recognise a variety of sounds and thus assign meaning to them. Hence, we propose to combine a model for acoustic event classification trained on the large-scale environmental sound database AudioSet, with a scene classification algorithm that couples direct identification of acoustic features with these recognised sound for scene recognition. The combined model is trained on TUT2018, a database containing ten everyday scenes. Applying the resulting AI-model to the soundscapes of the world database without further training shows that the classification that is obtained correlates to perceived calmness and liveliness evaluated by a test panel. It also allows to unravel why an acoustic environment sounds like a lively square or a calm park by analysing the type of sounds and their occurrence pattern over time. Moreover, disturbance of the acoustic environment that is expected based on visual clues, by e.g. traffic can easily be recognised.

1. INTRODUCTION

ISO standard ISO 12913-1:2014 defines an urban soundscape *as the acoustic environment as perceived and understood by people or society within a context*. This definition stresses the importance of the human in the emergence of a soundscape and thus goes well beyond the physics of sound waves, amount of energy, or even spectral information. *Understanding* implies that meaning is associated to the sound that is perceived. An meaning can be understood as the associations triggered in the human mind by the perception of the sounds in a given context [1]. Assigning a verbal label to the sounds that are heard, is a typical human way of assigning and communicating meaning. Thus, measurement equipment capable of automatically classify the sonic environment with soundscape in mind should be able to classify the sounds that could be heard in the sonic environment, where classifying in this context refers to assigning verbal labels (auditory event classification, AEC). In earlier work, prior to the advent of modern artificial intelligence, we showed that there is a relationship between automatic event classification and soundscape quality [2].

The type of sounds that can be identified contribute to the overall soundscape perception and hence the above-mentioned ISO standard also stipulates that assessment of soundscapes via interviews

[1] yuanbo.hou@ugent.be

[2] dick.botteldooren@ugent.be

should also include a question on the types of sounds that can be heard. But, soundscape is more holistic than a simple sum of the sounds that people hear. It is governed by a complex interplay of attention and saliency [3], expectations [4], and audiovisual interactions [5].

Artificial Intelligence methods today are far from being able to take into account all these factors that affect human perception and understanding. However, strong advances have been made in the areas of Acoustic Event Classification (AEC) [6] and Acoustic Scene Classification (ASC). The latter referring to the detection of a typical sound environment for a specific (urban) environment, e.g. the soundscape of a park. It stands beyond doubt that ASC and AEC are strongly related and hence multiple attempts have been made to combine part of the task [7,8]. Most prior work in this area assumes that the relevant basic features for classification are the same resulting in a common set of feature extracting layers in the convolutional neural networks (CNN) or transformer models. Here, we introduce a second link at the level of the labeled events which allows the scene classification not only to rely on the acoustic features but also on the pseudo-labels assigned to specific events.

Acoustic scene classification adds an additional layer of meaning to the acoustic environment by identifying combinations of sounds together with a background ambience that make an environment sound typical. Yet, not only the probability of hearing specific sounds determine how an acoustic environment is perceived, also the sequence and interweaving of these sounds over time can have a noticeable influence. In this, predictability and complexity may be suitable indicators. Early music theory and urban soundscape analysis identified the complexity in the sequence of level and pitch as possible indicators [9,10]. This type of complexity could be referred to as sensory complexity as it does not involve recognition and understanding. Similarly, once a sequence of recognized sounds is identified, the complexity this could be assessed [11]. This form of complexity may be referred to as semantic complexity to distinguish it from the former. The relevance of semantic complexity and predictability [12] for appraisal of soundscapes can be hypothesized on the basis of theories on the pleasure in predictability [13]: an environment that is sufficiently predictable to allow predictive coding to be efficient, yet not too predictable may give the highest pleasure. Very complex and hardly predictable partners could only be appreciated by field experts [14].

In this contribution we introduce a new model for collaborative acoustic scene and event classification and apply it to a collection of urban soundscape recordings. In Section 4 we will evaluate how this AI classification is related to human evaluation of these audiovisual environments and how semantic complexity calculated on the basis of the AEC typically varies between acoustic scenes.

2. COLLABORATIVE ACOUSTIC SCENE AND EVENT CLASSIFICATION

A few previous studies explored joint classification of scenes and events based on two frameworks: classification based on the same embedding space [7] (denoted as Framework1), and classification based on shared low-level and separated high-level embedding spaces [8,15,16] (denoted as Framework2). Framework1 attempts to learn the same acoustic representations applicable to both scenes and events based on the multi-task learning paradigm [17]. The resulting models are thus more efficient as multiple tasks can be performed at the same time. However, models based on Framework1 have difficulties adapting to the intricate scenes and events in real life using the same learned embeddings. Thus, Framework2 explores the shared joint scene-event representations and separated task-dependent representations for ASC and AEC. The models based on Framework2 are able to utilize diverse information of both joint and individual representations of scenes and events, resulting in better performance [18]. However, real-life acoustic scenes and audio events naturally have implicit relationships with each other, and these relationships between scenes and events are not fully explored

and used in Framework2. To this end, we recently proposed a new Relation-Guided ASC (RGASC) model to further exploit and coordinate the scene-event relation for the mutual benefit of scene and event recognition [19].

Inspired by the idea of RGASC [19], to jointly classify the auditory scene and label sound events, the collaborative scene-event classification (CSEC) framework is introduced. It uniquely extends current practice models by introducing a learnable coupling matrix between a scene classification branch that solely relies on basic acoustic features and an event identification branch that solely relies on acoustic features, to assist the acoustic scene classification.

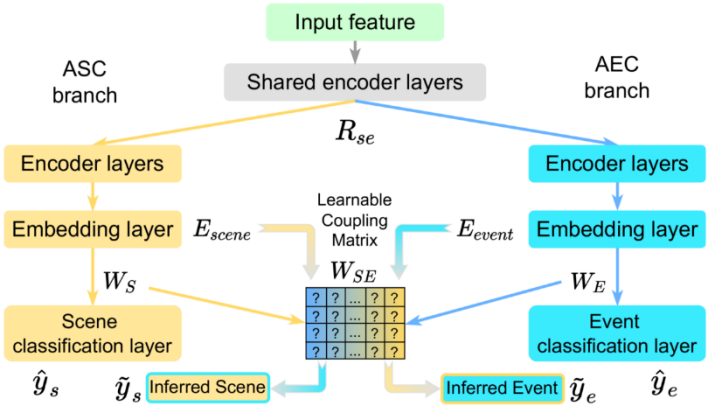

Figure 1 The proposed collaborative scene-event classification (CSEC) framework.

As shown in Fig. 1, the input time-frequency representations are mapped by the encoder layers and transformed into the joint classification space of scenes and events. The shared encoder encodes input representations into joint scene-event representations $\boldsymbol{R_{se}}$. Next, from the joint representations, the separated encoder further extracts scenes representations $\boldsymbol{R_S}$ and events representations $\boldsymbol{R_E}$, respectively. Subsequent embedding layers transform $\boldsymbol{R_S}$ and $\boldsymbol{R_E}$ into the embeddings of scene $\boldsymbol{E_s}$ and the embeddings of event $\boldsymbol{E_e}$ in the latent semantic spaces. Then, the scene classification layer maps the embeddings of scenes $\boldsymbol{E_s}$ onto target scene components by $\boldsymbol{W_S}$ and outputs the prediction of scene class $\hat{y}_s$. Similarly, the event classification layer maps the embeddings of scenes $\boldsymbol{E_e}$ by $\boldsymbol{W_E}$ and outputs the prediction of scene class $\hat{y}_e$. The weight matrices $\boldsymbol{W_S}$ and $\boldsymbol{W_E}$ in final classification layers can be viewed as the core knowledge about targets learned by the model.

The proposed CSEC framework will first attempt to construct a coupling matrix $\boldsymbol{W_{SE}}$ based on the core knowledge $\boldsymbol{W_S}$ and $\boldsymbol{W_E}$ to capture the bidirectional relation between scenes and events, and next map the scene embeddings $\boldsymbol{E_s}$ to event space based on the learned two-way relation matrix to infer the corresponding event. Then, the loss between the inferred event and the actual output of the event branch is calculated, and the loss is backpropagated to update relevant weights. Similarly, the event embeddings $\boldsymbol{E_e}$ are first mapped into the scene space to infer the corresponding scene output, next the loss between the inferred scene output and the actual output of the scene branch is measured to correct the learnable weights to obtain more accurate estimates. In this process, the learnable $\boldsymbol{W_{SE}}$ will model the implicit two-way scene-event relation, and the scene branch and event branch will collaborate to estimate each other's output and classify their own targets with the assistance of the modeled scene-event relation, to learn better representations by considering different levels of information in input samples from different aspects simultaneously.

On audio-related tasks, the Transformer-based Audio Spectrogram Transformer (AST) [20] has recently shown promising results. So, the proposed CSEC framework is instantiated on the AST to test

the performance of CSEC-AST. Compared to convolution-based models [21,22], convolution-free purely attention-based AST can be applied to audio spectrograms to capture long-range global context. The spectrogram is split into a sequence of patches [20]. To learn the individual high-level representations for the ASC and AEC, the scene-event joint representations $\boldsymbol{R_{se}}$ are fed into the encoder layers of the ASC branch and the AEC branch in the CSEC-AST, respectively. The learned representations are then fed into the embedding layer to learn better mappings in the latent semantic space that will be later used for the final classification. The learning process of the coupling matrix in CSEC-AST is jointly driven by the losses between the derived results $\widehat{y_e}$ and $\widehat{y_s}$ and the actual predictions $\hat{y}_e$ and $\hat{y}_s$.

For the dataset and training details, the development dataset of TUT Urban Acoustic Scenes 2018 (TUT2018) [23] with 8640 10-seconds segments from real life, totaling 24 hours of clips, is used in this paper. TUT2018 was recorded in six large European cities, in different locations for each scene class. The training/testing split of the TUT2018 dataset follows the default split of the DCASE 2018 Task 1 Subtask A3. There are no event labels in the scene dataset TUT2018. Thus, the pretrained model AST is used to tag each audio clip with a pseudo label to indicate the probabilities of the corresponding audio events. During training, log Mel filterbank (fbank) [24] is used as the acoustic feature. The clip-level spectrograms are standardized by subtracting the mean and dividing the standard deviation along the frequency axis. For CSEC-AST, the audio clip is converted into a sequence of 128-dimensional fbank computed with the 25ms Hamming window and a hop size of 10ms, then the spectrogram is split into a sequence of patches following the settings of [20]. The CSEC- AST is trained for a maximum of 50 epochs. A batch size of 64, and an Adam optimizer [25] with an initial learning rate of 1e-6 [20] are used to minimize the losses in CSEC-AST. To prevent over-fitting, dropout [26] and normalization are used in this paper.

3. THE SOUNDSCAPES-OF-THE-WORLD DATABASE AND ITS CLASSIFICATION

The pretrained CSEC model is used to investigate a collection of soundscapes (https://urban-soundscapes.org/) that was collected at typical locations in cities around the world using the methodology described in [27]. The database contains 130 unique settings where 360 degree video and ambisonics sound are recorded. From this database 60 sound environments were played back in virtual reality and evaluated by 20 persons [28]. Of particular interest for this work is the answer to the first question asked after each one-minute fragment: In general how would you characterize the environment you just experienced? with five answer categories ranging from very calming/tranquil to very lively/active. This question evaluates the environment as a whole, not the sonic environment in particular. The third question: How much did the sound draw your attention? is also considered here.

Based on the visual setting each environment was manually classified as a park, a square or a street.

The CSEC model is used to automatically classify the acoustic scenes. In this it should be noted that the training of the model allowed it to identify a wide range of acoustic scenes some of which did not occur in the soundscape-of-the-world database. Relevant categories are park, public square, and street/traffic acoustic scenes.

Every second the acoustic event classifier estimates the probability for one of 527 labels to apply to the current sound. The sequence of probabilities indicates how strongly the meaning of the sound varies over time. The complexity of this sequence could be indicative for the appraisal of the sound

[3] https://dcase.community/challenge2018/task-acoustic-scene-classification

environment [11,14]. Here, the semantic complexity is quantified using the recurrence plot and the indicators derived from this representation as presented in [29]. For this analysis only the 10 on-average most probable sounds are considered for every recording. Sounds that have a probability of occurrence of less than 5% at a given second are considered to be absent.

4. RESULTS

Table 1 compares the automatic acoustic scene classification with the visual classification of each of the recordings. The numbers indicate the co-occurrence of each combination of classifications. Most environments that look like parks also sound like parks although in 10 cases they sound like streets with a substantial amount of traffic. Hence, for a vast majority of the parks in the soundscapes of the world database, the soundscape matches expectations based on visual classification. For visually classified public squares, the situation is quite different. For most places, the sound environment is classified as street with traffic. In the training set, the public squares auditory scenes have been carefully selected to avoid traffic-noise dominated situations. For streets, the situation is again clear: when it looks like a street with traffic, it sounds like a street with traffic.

Table 1: number of soundscapes in each combination of visual classification (manual) and automatic acoustic scene classification.

visual classification	automatic acoustic scene classification				
	park	public square	street / traffic	tram	metrostation
park	34	3	10	3	1
square	9	11	35	1	0
street	0	1	17	0	1

For 60 audiovisual recordings, an evaluation by 25 volunteers was performed previously [28]. In a first question the environment experienced in virtual reality was rated on a five point scale ranging from very calming/tranquil (1) to very lively/active (5). Figure 2 shows the mean and standard deviation of the response to this question for the automatically acoustic scene classifications that resulted in 19 park, 3 public square, and 25 street/traffic environments. As expected, acoustic scenes automatically classified as parks are rated more calming and tranquil by people while acoustic scenes classified as public squares or street/traffic are rated more lively and active. A similar result is found for the answer on a question to what extend the sound has drawn attention of the evaluators of the audiovisual environment experienced in virtual reality. The differences between classes are nevertheless somewhat smaller.

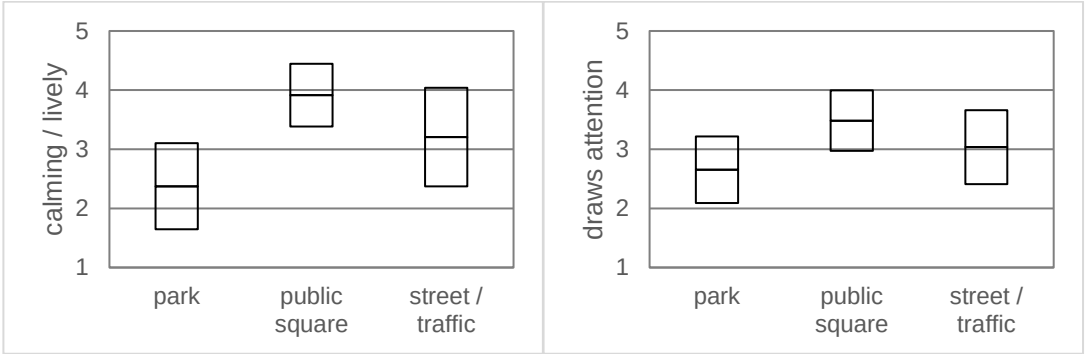

Figure 2: response of human listener (mean and standard deviation) for soundscapes categorized as parks, public squares and streets with traffic by the automatic auditory scene classification; left:

evaluation on a calming-lively 5-point axis; right: how strongly does the sound environment attracts attention.

The analysis of semantic complexity is based on two indicators extracted from the recurrence plot: the trapping time TT and the recurrence time of the second kind T2. Examples of soundscapes with very high T2 are R0118(Templo de Debod, Madrid), R0096 (Millennium Park – Crown Fountain, Chicago), R0084 (Fifth Avenue, New York), and R0007 (Chalet du Mont Royal, Montreal). Simply listening to these sounds online will already illustrate what this complexity indicator mean. Soundscapes where TT is particularly high can be found at R0126 (Užupis Art Incubator, Vilnius), R0119 (Lukiškes Square, Vilnius), R0008 (McGill University campus, Montreal), and R133 (National Museum of Lithuania, Vilnius). The monotonicity of these soundscapes becomes particularly clear, but this is not necessarily correlated to a low noise level.

When analyzing complexity for different acoustic scenes (Figure 3), it becomes clear that T2 seems to be lower and TT higher in acoustic scenes that are classified as parks. Acoustic scenes classified as public squares tend to have a more semantically complex sound environment.

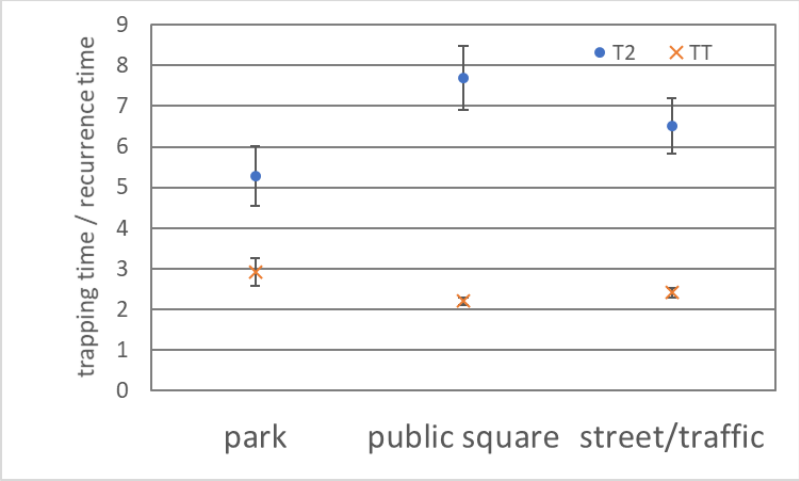

Figure 3:Indicators of complexity (trapping time TT and recurrence time of the second kind T2) of the sequence of sounds identified in each scene grouped by automatically classified acoustic scene; error bars indicated 95% confidence interval on the mean.

4. CONCLUSIONS

It was shown that collaborative acoustic scene and event classification using artificial intelligence, even with models trained on other datasets, allows to label sounds in a collection of sound recordings from around the world, and at the same time can label the acoustic scene as a whole. Evidence of relevance of this labeling has been given using secondary analysis of the outcome. Firstly, the acoustic scene labels were compared to visual classification of the environment. The acoustic scene in places visually identified as parks seems to match this classification, while places visually identified as public squares are quite often acoustically labeled as streets with traffic. The latter comes as no surprise as many of the public squares are indeed very close to traffic routes. Secondly, the AI-based acoustic scene classification was compared to human labeling on a calm versus lively axis. Public square acoustic scenes are generally rated more lively whiles scenes classified as parks are rated on average less lively.

To gain a better understanding on the reason for labeling sound environments lively or calming, a new concept has been introduced: semantic complexity. This concept measures the complexity of the temporal evolution of the different sounds that can be heard in a place. A constant hum even at relatively high amplitude would be rated as non-complex, while an environment filled with voices, music, laughter; but also an environment filled with the sound of cars, busses, trucks and people become more complex. It can safely be assumed that the semantic complexity dimension aligns more with the arousal than with the valence axis in a circumplex model. Nevertheless, on average, the fluctuation of sounds in a park acoustic scene shows less complex behavior than the fluctuation of sound in a public square acoustic scene. Finally, it should be noted that the appraisal of complex sonic environments may depend on the abilities of the visitor to understand this environment, and thus on its age.

5. ACKNOWLEDGEMENTS

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen" program.

6. REFERENCES

[1] D. Botteldooren et al., *From Sonic Environment to Soundscape*, in *Soundscape and the Built Environment* (2016).

[2] M. Boes, K. Filipan, B. de Coensel, and D. Botteldooren, *Machine Listening for Park Soundscape Quality Assessment*, Acta Acustica United with Acustica **104**, (2018).

[3] N. Huang and M. Elhilali, *Push-Pull Competition between Bottom-up and Top-down Auditory Attention to Natural Soundscapes*, Elife **9**, (2020).

[4] K. Filipan, M. Boes, B. de Coensel, C. Lavandier, P. Delaitre, H. Domitrović, and D. Botteldooren, *The Personal Viewpoint on the Meaning of Tranquility Affects the Appraisal of the Urban Park Soundscape*, Applied Sciences **7**, 91 (2017).

[5] G. M. Echevarria Sanchez, T. van Renterghem, K. Sun, B. de Coensel, and D. Botteldooren, *Using Virtual Reality for Assessing the Role of Noise in the Audio-Visual Design of an Urban Public Space*, Landscape and Urban Planning **167**, (2017).

[6] 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 **28**, (2020).

[7] H. L. Bear, I. Nolasco, and E. Benetos, *Towards Joint Sound Scene and Polyphonic Sound Event Recognition*, in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vols. 2019-September (2019).

[8] N. TONAMI, K. IMOTO, R. YAMANISHI, and Y. YAMASHITA, *Joint Analysis of Sound Events and Acoustic Scenes Using Multitask Learning*, IEICE Transactions on Information and Systems **E104D**, (2021).

[9] B. de Coensel, D. Botteldooren, and T. de Muer, *1/f Noise in Rural and Urban Soundspaces*, Acta Acustica (Stuttgart) **89**, (2003).

[10] D. Botteldooren, B. de Coensel, and T. de Muer, *The Temporal Structure of Urban Soundscapes*, Journal of Sound and Vibration **292**, (2006).

[11] D. Botteldooren, *Urban Soundscape Complexity*, J Acoust Soc Am **150**, (2021).

[12] G. Boffetta, M. Cencini, M. Falcioni, and A. Vulpiani, *Predictability: A Way to Characterize Complexity*, Physics Reports.

[13] J. A. Litman, *Curiosity and the Pleasures of Learning: Wanting and Liking New Information*, Cognition and Emotion.

[14] D. Botteldooren, *Urban Sound Design for All*, in *Proceedings of 2020 International Congress on Noise Control Engineering, INTER-NOISE 2020* (2020).

[15] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, and Y. Yamashita, *Sound Event Detection by Multitask Learning of Sound Events and Scenes with Soft Scene Labels*, in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Vols. 2020-May (2020).

[16] T. Komatsu, K. Imoto, and M. Togami, *Scene-Dependent Acoustic Event Detection with Scene Conditioning and Fake-Scene-Conditioned Loss*, in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Vols. 2020-May (2020).

[17] Y. Zhang and Q. Yang, *A Survey on Multi-Task Learning*, IEEE Transactions on Knowledge and Data Engineering (2021).

[18] N. Tonami, K. Imoto, M. Niitsuma, R. Yamanishi, and Y. Yamashita, *Joint Analysis of Acoustic Events and Scenes Based on Multitask Learning*, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Vols. 2019-October (2019).

[19] Y. Hou, W. van Hauwermeiren, and D. Botteldooren, *Relation-Guided Acoustic Scene Classification Aided with Event Embeddings*, in *Proceedings of 2022 International Joint Conference on Neural Networks (IJCNN). IEEE.* (2022).

[20] Y. Gong, Y. A. Chung, and J. Glass, *Ast: Audio Spectrogram Transformer*, in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 1 (2021).

[21] Y. Hou, Q. Kong, and S. Li, *A Comparison of Attention Mechanisms of Convolutional Neural Network in Weakly Labeled Audio Tagging*, in *Lecture Notes in Electrical Engineering*, Vol. 568 (2019).

[22] M. Lim, D. Lee, H. Park, Y. Kang, J. Oh, J. S. Park, G. J. Jang, and J. H. Kim, *Convolutional Neural Network Based Audio Event Classification*, KSII Transactions on Internet and Information Systems **12**, (2018).

[23] A. Mesaros, T. Heittola, and T. Virtanen, *TUT Database for Acoustic Scene Classification and Sound Event Detection*, in *European Signal Processing Conference*, Vols. 2016-November (2016).

[24] A. BALA, *Voice Command Recognition System Based on Mfcc and Dtw*, International Journal of Engineering Science and Technology **2 (12)**, (2010).

[25] D. P. Kingma and J. L. Ba, *Adam: A Method for Stochastic Optimization*, in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015).

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, Journal of Machine Learning Research **15**, (2014).

[27] B. de Coensel, K. Sun, and D. Botteldooren, *Urban Soundscapes of the World: Selection and Reproduction of Urban Acoustic Environments with Soundscape in Mind*, in *INTER-NOISE 2017 - 46th International Congress and Exposition on Noise Control Engineering: Taming Noise and Moving Quiet*, Vols. 2017-January (2017).

[28] K. Sun, B. de Coensel, K. Filipan, F. Aletta, T. van Renterghem, T. de Pessemier, W. Joseph, and D. Botteldooren, *Classification of Soundscapes of Urban Public Open Spaces*, Landscape and Urban Planning **189**, (2019).

[29] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, *Geometry from a Time Series*, Physical Review Letters **45**, (1980).

Proceedings

Internoise 2022

ISBN 978-1-906913-42-7