MADRID
inter.noise 2019
June 16 - 19
NOISE CONTROL FOR A BETTER ENVIRONMENT

Automated environmental sound recognition for soundscape measurement and assessment

Yang, Ming[1]
HEAD acoustics GmbH
Herzogenrath, Germany

Yu, Lei[2]
School of Architecture, Harbin Institute of Technology
Shenzhen, China

Herweg, Andreas[3]
HEAD acoustics GmbH
Herzogenrath, Germany

ABSTRACT

Automated environmental sound recognition, the technique of machines/computers to recognise objects and events in environment as humans do, has an important role in diverse fields, ranging from security surveillance, warning/assistant systems, smart homes/buildings/cities, to autonomous robots. In soundscape, a sound environment is composed of various sound sources, such as natural sounds (moving water, bird song, etc.), mechanical sounds (transportation, construction, etc.), human activity sounds (speaking, footsteps, etc.), and cultural/historical sounds (church bells, music, etc.), the identification of which strongly influence humans' perception and assessment of the soundscape or the place/environment. Thus, for the assessment and management of soundscape, it is important and often necessary to measure the sound source information, in addition to acoustic indicators such as sound pressure level. However, measuring the sound source information requires a significant amount of human effort, which largely restricts such soundscape assessment/management approach in large scale real applications. This present paper therefore explores the method of automated environmental sound recognition, for recognising sound categories that strongly impact soundscape assessment (i.e. natural sounds, mechanical sounds, human sounds, and cultural/historical sounds), towards meeting the specific need of soundscape measurement, assessment and management. It examines the recognition performance of an automated recognition model, which uses a set of psychoacoustic/acoustic features and machine learning model of neural network, with multiple datasets covering various sound sources in these sound categories recorded in outdoor environment. The results show good recognition ability of the method.

Keywords: Soundscape assessment, Sound recognition, Machine learning
I-INCE Classification of Subject Number: 60

[1] mingkateyang@163.com
[2] Leilayu@hitsz.edu.cn
[3] Andreas.Herweg@head-acoustics.de

1. INTRODUCTION

A sound environment is composed of various sound sources, such as natural sounds (moving water, bird song, etc.), mechanical sounds (transportation, construction, etc.), human activity sounds (speaking, footsteps, etc.), and cultural/historical sounds (church bells, music, etc.). Soundscape research suggested that humans' perception and assessment of sound environment or soundscape are strongly influenced by the identification/recognition of the sound sources. Thus, for the assessment and management of soundscape, it is important and often necessary to measure the sound source information, in addition to acoustic indicators such as sound pressure level (SPL). However, unlike SPL, measuring the sound source information requires a significant amount of human effort, which largely restricts such soundscape measurement approach in large scale real applications.

With the rapid development of computer techniques, automated environmental sound recognition, the technique of machines/computers to recognise objects and events in environment as humans do, is being widely studied and has the increasingly important role in diverse fields, ranging from security surveillance, warning/assistant systems (for people with special needs), smart homes/buildings/cities, to autonomous robots. By applying the technique of environmental sound recognition, it would be possible to achieve the automated soundscape measurement of sound source information just like SPL. However, most of the studies of environmental sound recognition focused on the recognition of domestic sounds, urban sounds, etc., according to their specific applications, few had their focuses on soundscape or the categories of sound sources that have the most significant impacts on soundscape assessments, e.g. mechanical sounds, human sounds, natural sounds, and cultural/historical sounds.

This present paper therefore explores the automated recognition technique of environmental sound sources in recognising sound categories that are most concerned in soundscape from field recordings, towards meeting the specific need of soundscape measurement, assessment and management.

The rest of this paper first presents the critical importance of sound sources on soundscape assessment suggested by a number of previous soundscape studies, and finds the sound categories that soundscape mostly concerns. It then proposes an automated recognition model, which uses a set of psychoacoustic/acoustic features and machine learning model of neural network, and examines its recognition ability on the sound categories, with multiple datasets covering various sound sources recorded in outdoor environment.

2. EFFECT OF IDENTIFICATION OF SOUND SOURCES ON SOUNDSCAPE ASSESSMENT

2.1 Sound Sources and Soundscape Assessment/Evaluation

Since acoustic indicators such as SPL are correlated with people's subjective assessment (e.g. annoyance or sound quality) of noise or unpleasant sound environments, such as those dominated by traffic noise [1], which is the primary concern of the field of environment noise, the acoustic indicators have been used as a standard way for the measurement and assessment of such sound environments and further sound environments in general. However, when considering general sound environments including various environments/places that are composed of various sound sources, which is the concern of the field of soundscape, the acoustic indicators' abilities in explaining the subjective assessments are limited, and so are psychoacoustic indicators (e.g. loudness, roughness, sharpness and fluctuation) [2].

Nevertheless, a number of soundscape studies suggested that the identification or cognition of sound sources that compose the sound environments significantly affects the subjective assessment of a sound environment / soundscape [3]. The presence of mechanical and human sounds has negative impacts on reported soundscape assessments, whereas natural and cultural/historical sounds have positive impacts [4, 5]. For example in recreation parks, Kim and Shelby [6] found the mechanical sounds of airplane and truck engine decreased acceptability rating of environment, and the natural sounds of birds and water increased acceptability rating; Tse et al. [7] found hearing sounds from heavy vehicles or bikes reduced the acoustic comfort evaluation, and hearing breeze increased comfort evaluation. Importantly, these relationships remained after controlling overall sound level or Zwicker's loudness of soundscape [8].

2.2 Prediction of Soundscape Assessment from Sound Sources

Furthermore, soundscape studies suggested that the subjective soundscape assessment can be primarily determined by the dominant sound sources perceived, either anthropogenic sounds (mechanical and human sounds) or natural sounds, or the percentages of anthropogenic sounds and natural sounds [9]. Soundscapes dominated by anthropogenic sounds were found to be negative (e.g. unpleasant or lower acoustic comfort), and soundscape dominated by natural sounds to be positive (e.g. pleasant or comfort) [8].

A number of studies thus proposed prediction models on soundscape assessment based on multiple linear regression of the dominance or percentage of anthropogenic/natural sound sources in soundscape, reflected by either the perceived loudness or perceived length of time of the sources, as well as sound level / loudness of overall soundscape.

Through laboratory experiment, Pheasant et al. [10] proposed models on the perception of tranquillity (TR):

$$TR = 9.99 - 0.93L_{Amax} - 0.45PLM + 1.16PLB$$

where PLM is the perceived loudness of mechanical sounds and PLB is the perceived loudness of biological (natural) sounds.

Based on soundwalk, Aumond et al. [11] proposed a model of pleasantness (P) in urban context:

$$P = 9.70 - 0.47OL - 0.21T + 0.12V + 0.09B$$

where OL is perceived overall loudness, T is the perceived time of presence of traffic, V the perceived time of presence of voices, and B the perceived time of presence of birds.

Such models have rather similar structure and both have good prediction ability, the second explaining 58% of the specific assessment variance of the sound environment. They are among the generally best prediction models so far in soundscape research, which are more advanced than those based on sound level or loudness only and also more advanced than the more complex models based on additional soundscape factors such as demographical and behavioural factors [7, 12].

However, qualifying the sound sources requires a large amount of human effort, which restricts the application of the prediction models. Thus, some studies further used certain specific acoustic/psychoacoustic indicators in models to reflect the presence of the sound sources, e.g. Time and Frequency Second Derivative, describing the normalized deviations within frequency bands, to represent voices or birds [11], sharpness to represent the dominant presence of water fountains and voices, and $L_{Ceq} - L_{Aeq}$ (low frequency content) to represent vehicles [13]. However, these acoustic

indicators only work for such cases studied, and no accurate acoustic indicators have been found for general situations.

3. AUTOMATED RECOGNITION OF ENVIRONMENTAL SOUND SOURCES

The technique of environmental sound recognition would be a useful tool to replace the manual effort in qualifying sound sources of the above prediction models of soundscape assessment, to make it possible to achieve the automated soundscape measurement and assessment from field recordings.

As discussed in the last section, the categories of environmental sound sources, which include natural sounds, mechanical sounds, human activity related sounds, and cultural/historical sounds, have the most significantly impacts on soundscape assessments. This section thus focuses on the automatic recognition of these sound categories from field recordings, as a first step towards soundscape measurement, monitoring, assessment and management.

3.1 Sound Recognition Model

While there are a range of automated recognition methods studied for environmental sounds, the present study uses the method similar to that in the authors' previous research [14, 15], to examine the recognition ability of the technique of automatic environmental sound recognition for the soundscape purpose.

Following the fact that human listeners perceive sound through the peripheral auditory system and recognise/identify sound using the auditory sensations as potential cues by the higher-level neural system, the automated recognition model similarly consists of two such phases.

In the first phase, a range of psychoacoustic/acoustic features as well as their temporal variations of sound are extracted via digital signal processing, according to the human auditory sensations [16, 17]. The psychoacoustic/acoustic features extracted include:

- Level (unweighted) (L)
- Level A (L_A)
- Loudness (N), calculated according to the DIN 45631/A1 standard for time-variant sound. Loudness is an auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud.
- Sharpness (S), calculated according to DIN 45692 standard. Sharpness is correlated to the spectral energy distribution, and increases with increasing centre frequency.
- Tonality, calculated according to DIN 45681 standard. Tonality indicates whether a sound consists of mainly tonal components or broadband noise.
- Tonality frequency, calculated according to DIN 45681 standard. It indicates the frequency at which the highest tonality appears.
- Roughness (R), calculated according to the hearing model of Sottek [18] (1Bark solution is used). Roughness is related to the beating phenomenon, or relatively quick changes of sound.
- Impulsiveness, calculated according to the hearing model of Sottek.

The instantaneous psychoacoustic/acoustic features are calculated for the whole duration of sound. For the statistics of the temporal variations of each of the features, average (Ave), standard deviation (StD), maximum (Max), percentile 5, percentile 25, percentile 75 and percentile 95 are used. Thus, 56 features (e.g. N Ave, S StD) are extracted in total.

In the second phase, the machine learning model of artificial neural network (ANN) that is frequently used in audio and visual pattern recognition is applied for the recognition/classification of sound categories based on the features extracted. Feed-forward ANN with back-propagation training (supervised learning) is used, which consists of a set of nodes that are organized in layers, input, hidden and output layers. Networks with one hidden layer is used for the demonstration purpose. The distance or error between the network's output and the desired output (target) is measured by the cross-entropy (CE) error. The training process iteratively adjusts the relevant weights of the nodes/ connections between the nodes to minimise the error. Scaled conjugate gradient algorithm is used for the iteration process of training.

The implementation of the sound recognition model, including the calculation of the psychoacoustic/acoustic features, statistics, and neural network training and testing, is made through the commercial software ArtemiS SUITE from HEAD acoustics and MATLAB with Neural Network Toolbox 11.0.

3.2 Datasets

To examine the recognition method with a large set of sound samples, a number of open datasets of environmental sounds available on the public internet are used in the present study. Such datasets were created to be suitable for benchmarking methods of environmental sound recognition. The samples in all the datasets were manually collected from field recordings gathered in the online audio database Freesound.org. Each dataset contains sound files (in the wave format) and the corresponding semantical class (annotation information) label for each file/sample. While each sample may contain multiple sound sources, it is only labelled with a single class. The duration of the samples varies from a few seconds to tens of seconds. These datasets include:

- Acoustic Event Dataset [19], which contains 5223 sound samples organised in 28 semantical classes.
- DBR (Dog, Bird, and Rain) dataset [20], which consists of three classes of sounds, i.e. dog, bird, and rain, each containing 50 samples.
- ESC-50 (Environmental Sound Classification 50 semantical classes) [21], which consists of a collection of 2000 5-second-long sound samples organised in 50 classes (with 40 samples per class).
- UrbanSound8K [22], which contains 8732 sound excerpts (<=4s) organised in 10 classes, i.e. air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music.

For each of these datasets, a number of classes of sounds that are related to the common sound sources in urban soundscape according the concern of the present study are selected among the original classes. For example, 18 classes among the 50 classes are selected of ESC-50 dataset. The sound classes selected and included in the present study include bird, sea waves, rain, wind, crickets, children playing, dog, footsteps, speech, airplane, car, train, engine, drilling, church bells, music, etc. The sound classes of each dataset that are selected are shown in Table 1.

The selected sound classes are then arranged into the four sound categories considered, i.e. natural sounds, mechanical sounds, human activity related sounds, and cultural/historical sounds. The classification of the sound classes or sources (the category that they are belonged to) is made roughly according to Schafer [23] and Brown et al. [24]. In the present study, dog barking sound is among the human activity related sounds, since in general everyday urban environment dog sound (unlike bird sound) is closely related to human activity, and people's perception/assessment (such as preference) of it is similar to human sounds compared to natural sounds or mechanical sounds. Table 1.

also shows the categories of the sound classes included for each dataset. It is noted that while the present study classifies the sound classes/sources in the current way, other classification is possible, which would not significantly change the main results of the paper, given the aim of the paper is to investigate the possibility of automated recognition of sound categories.

For each dataset with the selected classes of sound samples, the samples are further divided into three sets: training set to train the neural networks, validation set to monitor network's performance on samples outside the training set (i.e. generalisation) and automatically stop the training process when the CE error of the validation set increases for avoiding overtraining, and testing set to test the network by providing an independent measure of the network performance. Within each dataset, 70% of the selected samples are randomly picked into the training set, 15% into validation set, and 15% into testing set.

Table 1. Sound classes selected in each sound category

	Acoustic Event Dataset	DBR dataset	ESC-50	UrbanSound8K
Nature	bird	bird	chirping birds	
	ocean surf		sea waves	
		rain	rain	
	rustle		wind	
			crickets	
			thunderstorm	
Human	child			children playing
	crowd			
	dog barking	dog	dog	dog bark
	footstep		footsteps	
	laughter		laughing	
	speech			
	whistle			
Mechanics	airplane		airplane	
	helicopter		helicopter	
	car		car horn	car horn
			siren	siren
			train	air conditioner
	engine		engine	engine idling
	hammer		chainsaw	jackhammer
			hand saw	drilling
Culture	acoustic guitar			
	church bell		church bells	
	violin			

3.3 Recognition Performance Results

For each dataset, a series of neural networks with different network structure, i.e. number of nodes in the hidden layer (from 12 to 100), are trained/developed, to automatically recognise/classify the sound samples into the 4 categories, i.e. nature, human, mechanics, and culture. The results of the best network among them (evaluated by the average recognition accuracy of training, validation and testing sets) for each

dataset are shown in Table 2, which thus generally has the optimal network structure (number of hidden nodes) and represents the highest accuracy that can be achieved based on the recognition method. It shows the percentage of correctly classified samples, respectively for all samples (considering training, validation and testing samples together), the training samples, the validation samples, the testing samples, and average accuracy of training, validation and testing sets. It can be seen that the recognition accuracies are above 82% for the testing sample sets of all the datasets. Also, the accuracy is reliable across the different datasets.

In particular, Tables 2 and 3 show the confusion matrix of correctly-classified and misclassified samples of each sound category, considering all the training, validation and testing samples together, for Audio Event Dataset and ESC-50 dataset respectively as examples. In general, the results suggest the good recognition ability of the method.

Table 2. Recognition accuracy (%) of the optimal networks for each of the datasets

	Hidden nodes	Accuracy of all sample	Training accuracy	Validation accuracy	Testing accuracy	Average accuracy
Audio Event Dataset	68	85.73	87.04	83.04	82.28	84.12
DBR dataset	14	100.00	100.00	100.00	100.00	100.00
ESC-50	88	88.47	89.68	87.04	84.26	86.99
UrbanSound8K	38	85.64	85.60	86.97	84.51	85.69

Table 3. Confusion matrix of the optimal network for Audio Event Dataset

	Nature	Human	Mechanics	Culture
Nature	386	90	45	2
Human	37	910	49	10
Mechanics	39	66	686	7
Culture	0	15	16	276

Table 4. Confusion matrix of the optimal network for ESC-50 dataset

	Nature	Human	Mechanics	Culture
Nature	217	3	20	0
Human	5	106	9	0
Mechanics	29	3	282	6
Culture	0	0	8	32

4. CONCLUSIONS AND FUTURE WORKS

This present paper studies the automatic recognition of categories of environmental sound sources that significantly impact soundscape assessments, i.e. natural sounds, mechanical sounds, human activity related sounds, and cultural/historical sounds, from field recordings, as a first step towards soundscape measurement, monitoring, assessment and management.

An automated sound recognition method is proposed, which uses a set of psychoacoustic/acoustic features and machine learning model of neural network. The recognition ability of the method on the sound categories are examined with multiple open datasets, covering various urban environmental sound sources recorded in different

places/situations, by different equipment and with different qualities, which brings the challenge for the recognition. The results show that the recognition accuracies are above 82% in testing for all the datasets. It suggests good recognition ability of the method. Also, it suggests the possibility of using the technique of environmental sound recognition for automated soundscape measurement in terms of the category information of the dominant sound source.

Since the current sound recognition models are based on the neural network with relatively simple network structure, e.g. one hidden layer is used, and established psychoacoustic/acoustic features, there is still room for the recognition accuracy to improve by using such as more complex network structure or deep learning, or new features to be developed. Also, additional general models can be developed using such as all the samples of the datasets together. The recognition ability of any of the current models trained on a single dataset can be further validated with the other different datasets, to examine their generalisations.

Furthermore, while the current study provides a first step towards automated soundscape measurement, assessment and management, further works would be needed on such as comprehensive classification of sound sources that correlates with people's perceptions and assessments of soundscape, and estimation of the loudness and duration of the recognised sound sources/categories to predict soundscape assessment as discussed above, to achieve an automated soundscape assessment.

5. ACKNOWLEDGEMENTS

This research is funded by HEAD Genuit Foundation research fellowship (project no.: P-17/02-W). The authors would like to thank Prof. Klaus Genuit at HEAD acoustics GmbH, Prof. Brigitte Schulte-Fortkamp and Prof. André Fiebig at Technical University of Berlin for the valuable discussions. The authors also would like to thank Vishwesh as an intern student at HEAD acoustics GmbH for preparation of the dataset files.

6. REFERENCES

1. A. J. Torija and I. H. Flindell, "Listening laboratory study of low height roadside noise barrier performance compared against in-situ field data", *Building and Environment*, vol. 81, pp. 216-225 (2014)

2. D. A. Hall, et al., "An exploratory evaluation of perceptual, psychoacoustic and acoustical properties of urban soundscapes", *Applied Acoustics*, vol. 74, no. 2, pp. 248-254 (2013)

3. J. Kang and B. Schulte-Fortkamp, *"Soundscape and the Built Environment"*, edited by J. Kang and B. Schulte-Fortkamp, CRC Press, Boca Raton (2016)

4. B. Schulte-Fortkamp and A. Fiebig, "Soundscape analysis in a residential area: An evaluation of noise and people's mind", *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 875-880 (2006)

5. J. A. Benfield, et al., "Aesthetic and affective effects of vocal and traffic noise on natural landscape assessment", *Journal of Environmental Psychology*, vol. 30, no. 1, pp. 103-111 (2010)

6. S. O. Kim and B. Shelby, "Effects of soundscapes on perceived crowding and encounter norms", *Environmental Management*, vol. 48, no. 1, pp. 89-97 (2011)

7. M. S. Tse, et al., "Perception of urban park soundscape", *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2762-2771 (2012)

8. O. Axelsson, M. E. Nilsson, and B. Berglund, "A principal components model of soundscape perception", *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 2836-2846 (2010)

9. G. Perez-Martinez, A. J. Torija, and D. P. Ruiz, "Soundscape assessment of a monumental place: A methodology based on the perception of dominant sounds", *Landscape and Urban Planning*, vol. 169, pp. 12-21 (2018)
10. R. J. Pheasant, et al., "The importance of auditory-visual interaction in the construction of 'tranquil space'", *Journal of Environmental Psychology*, vol. 30, no. 4, pp. 501-509 (2010)
11. P. Aumond, et al., "Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context", *Acta Acustica United with Acustica*, vol. 103, no. 3, pp. 430-443 (2017)
12. L. Yu and J. Kang, "Modeling subjective evaluation of soundscape quality in urban open spaces: An artificial neural network approach", *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1163-1174 (2009)
13. A. Maristany, M. Recuero Lopez, and C. Asensio Rivera, "Soundscape quality analysis by fuzzy logic: A field study in Cordoba, Argentina", *Applied Acoustics*, vol. 111, pp. 106-115 (2016)
14. M. Yang and J. Kang, "Automatic identification of environmental sounds in soundscape", in *Proceedings of the 42nd International Congress and Exposition on Noise Control Engineering (Inter-noise 2013)*, Innsbruck, Austria (2013)
15. M. Yang and J. Kang, "Psychoacoustical evaluation of natural and urban sounds in soundscapes", *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 840-851 (2013)
16. K. Genuit and A. Fiebig, "Psychoacoustics and its benefit for the soundscape approach", *Acta Acustica United with Acustica*, vol. 92, no. 6, pp. 952-958 (2006)
17. E. Zwicker and H. Fastl, *"Psychoacoustics – Facts and Models"*, Springer, Berlin (1999)
18. R. Sottek, "Gehörgerechte Rauhigkeitsberechnung", in *DAGA*, Dresden (1994)
19. N. Takahashi, et al, "Deep convolutional neural networks and data augmentation for acoustic event recognition", in *Interspeech*, San Fransisco (2016)
20. V.-V. Eklund, DBR dataset [Data set], Zenodo, http://doi.org/10.5281/zenodo.1069747 (2017)
21. K. J. Piczak, "ESC: Dataset for environmental sound classification", in *Proceedings of the 23rd Annual ACM Conference on Multimedia*, Brisbane, Australia (2015)
22. J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research", in *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, USA (2014)
23. R. M. Schafer, *"The Tuning of the World"*, Knopf, New York (1977)
24. A. L. Brown, J. Kang, and T. Gjestland, "Towards standardization in soundscape preference assessment", *Applied Acoustics*, vol. 72, no. 6, pp. 387-392 (2011)