A Temporal Convolutional Neural Network for Multi-label Sound Recognition and Annoyance Detection of Complex Soundscapes

Andrew Mitchell,[1,2,a)] Christopher Soelitsyo,[3] Mercede Erfanian,[1] Tin Oberman,[1] Jing-Hao Xue,[4,2] Jian Kang,[1] and Francesco Aletta[1]

[1] *Institute for Environmental Design & Engineering, University College London, UK*

[2] *The Alan Turing Institute, London, UK*

[3] *Charras Lab, Institute for the Physics of Living Systems, University College London, UK*

[4] *Department of Statistical Science, University College London, UK*

(Dated: 20 June 2021)

Increasing urban noise pollution and simultaneous improvements in smart city sensor technology and deployment have created a necessity for increasingly sophisticated approaches to automated noise recognition. Sensor networks which are focused primarily on sound level monitoring have proved to be insufficient to adequately identify harmful sound events or to reflect the human impact of noise in cities. Therefore, ...

 [http://dx.doi.org(DOI number)]

[XYZ] Pages: 1–2

I. INTRODUCTION

Increasing urban noise pollution and simultaneous improvements in smart city sensor technology and deployment have created a necessity for increasingly sophisticated approaches to automated noise recognition. Sensor networks which are focused primarily on sound level monitoring have proved to be insufficient to adequately identify harmful sound events or to reflect the human impact of noise in cities. Therefore, the development of automated environmental sound recognition (ESR) systems has become a necessary component of next-generation approaches to noise pollution mitigation.

A. Importance of sound source and annoyance detection

B. AI for sound source recognition

1. Previous approaches

C. DCASE Challenge

D. SONYC

1. Datasets

(Cartwright *et al.*)

2. Component Parts

E. Empirical Models of Annoyance

1. Zwicker Psychoacoustic Annoyance

The field of psychoacoustics has had a particular focus on annoyance modelling, however this field presents some typical limitations. Firstly, from its inception it has made use of simple, simulated sounds for conducting laboratory tests. These are useful in that they enable much more control over the acoustic characteristics of the sound, allowing for isolated testing of the independent variables, in a conventional experimental approach. This is a limiting approach also taken in the field of auditory neuroscience ct *Need to add citations for this* . Second, the field is primarily developed towards and focussed on annoyance modelling of single sound sources, typically commercial products such as vacuum cleaners and high-end cars ct .

2. Soundscape Models

F. AI Models of Annoyance

II. METHODS

A. Temporal convolutional neural network

III. EXPERIMENTS

A. Datasets

1. DeLTA / SSID Binaural Dataset

Recording splitting In order to increase the available dataset and to make all of the recordings a consistent length, the original recordings were split into 15 seconds chunks. For all recordings, as many complete 15s chunks

a) andrew.mitchell.18@ucl.ac.uk

	dBFS	max_dBFS
count	2891	289
mean	-36.33	-18.92
std	7.45	7.68
min	-63.86	-46.34
25%	-40.04	-23.23
50%	-35.49	-18.40
75%	-31.42	-13.69
max	-15.35	-0.66

as possible were extracted and the remaining portion was excluded; for instance, for a 34s original recording, two sequential 15s chunks are extracted from the beginning, and the remaining 4s are not used. The original dataset of 1,453 recordings then results in 2,921 15s mp3s.

Gain Boost Due to the limitations of the means of delivery of the stimuli and to ensure the sounds did not exceed a safe level, we excluded the top 30 most loudest recordings as outliers. This was done by calculating the peak volume of the recording and excluding the top 1% (> -8.64 dBFS) loudest recordings. The peak value was used to ensure no recordings would clip. We then added a gain boost of 8 dB to all recordings, enabling us to include 250 very soft acoustic environments featuring little or no specific sound sources. This results in a total dataset of 2,891 recordings, with the relative volumes given in Table III A 1. The audio processing was done in Python, using pydub (Robert *et al.*, 2018).

2. DCASE 2018 / SONYC

B. Model architecture

C. Training and Evaluation

IV. DISCUSSION

"Cumulative annoyance due to compounding acute annoyance events."

V. CONCLUSION

And in conclusion...

ACKNOWLEDGMENTS

This research was supported by ...

Cartwright, M., Mendez, A. E. M., Dove, G., Cramer, J., Lostanlen, V., Wu, H.-H., Salamon, J., Nov, O., and Bello, J. P. "SONYC Urban Sound Tagging (SONYC-UST): a multi-label dataset from an urban acoustic sensor network" http://markcartwright.com/files/cartwright2019sonycust.pdf, doi: 10.5281/ZENODO.3338310.

Robert, J., Webbie, M. *et al.* (**2018**). "Pydub" http://pydub.com/.