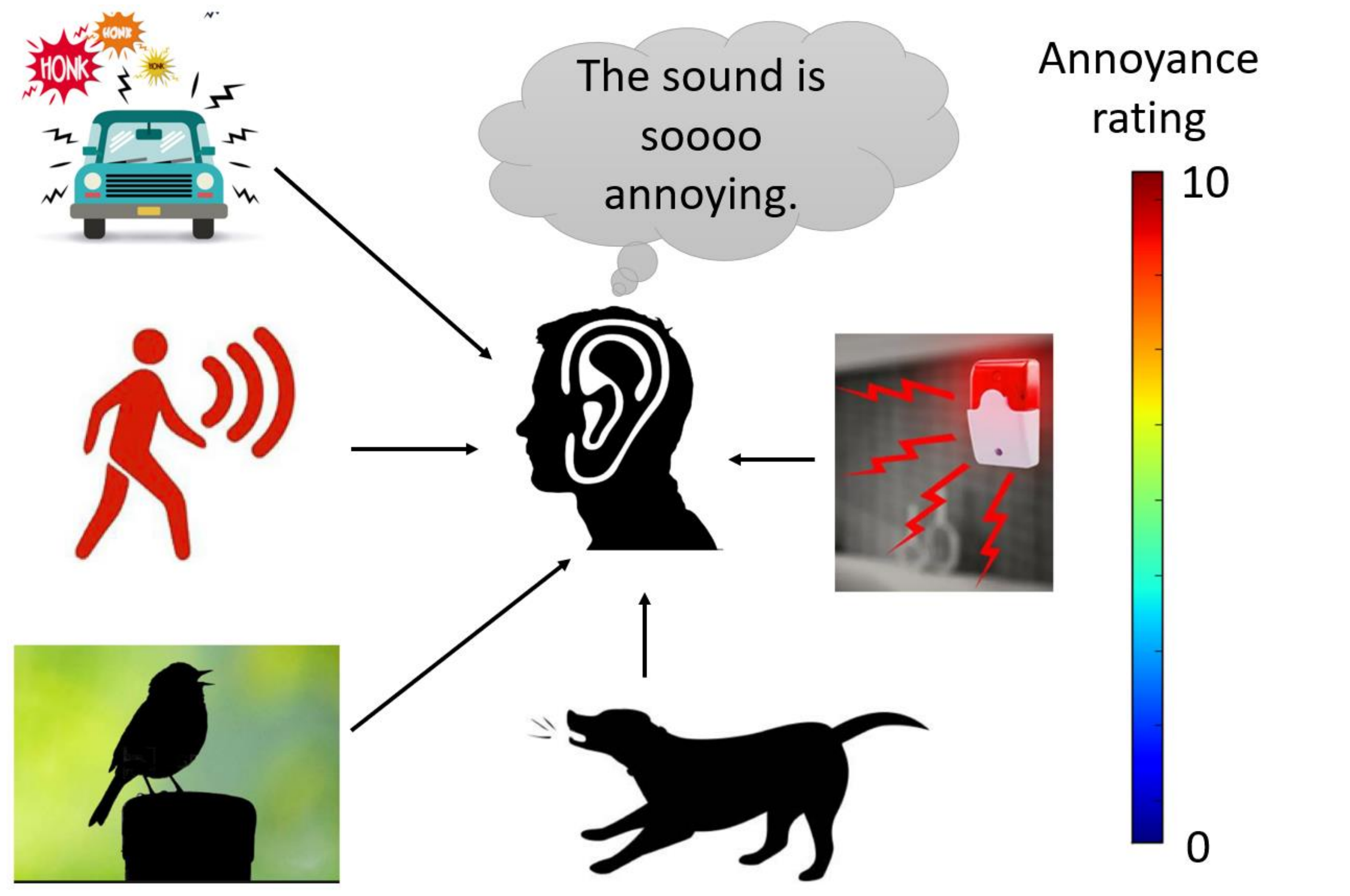
Multi-level graph learning for audio event classification and human-perceived annoyance rating prediction

Yuanbo Hou[1], Qiaoqiao Ren[1], Siyang Song[2], Yuxin Song[3], Wenwu Wang[4], Dick Botteldooren[1]

[1]WAVES, Ghent University, Belgium [2]University of Leicester, UK [3]Baidu Inc., China [4]CVSSP, University of Surrey, UK

1. Introduction

Audio events (AEs) in daily life carry rich information about the objective world. The composition of these sounds affects the mood of people in a soundscape.



WHO's report on environmental noise estimates that 22 M people suffer from chronic annoyance related to noise caused by AEs from various sources. Annoyance may lead to health issues and adverse effects on metabolic and cognitive systems.

To create annoyance-related monitoring, this paper proposes a graph-based model to identify AEs in a soundscape, and explore relations between diverse AEs and human-perceived annoyance rating (AR)

2. Lightweight attention-fused multi-level graph learning

2.1. Dataset

Publicly available dataset DeLTA includes both AE labels and human AR scores. Each audio clip in DeLTA has a clip-level 24-dimensional multi-hot vector as the ***fAE*** label, and an ***AR*** (continuously from 1 to 10).

Based on the ontology of the event labels in AudioSet, we ***further group*** the 24 ***fAE*** classes of DeLTA into 7 ***cAE*** classes:

1) ***Vehicle***: [aircraft, bus, car, general traffic, motorcycle, rail, screeching brakes]; 2) ***Music***: [bells, music];

3) ***Animals***: [bird tweet, dog bark];

4) ***Human sounds***: [children, laughter, speech, shouting, footsteps];

5) ***Alarm***: [siren, horn];

6) ***Natural*** sounds: [rustling leaves, water];

7) ***Other***: [construction, non-identifiable, ventilation, other].

2.2. Local context-aware graphs (LcGs)

2.2.1. fAG: fAEs-AR graph

2.2.2. fcG: fAEs-cAEs graph

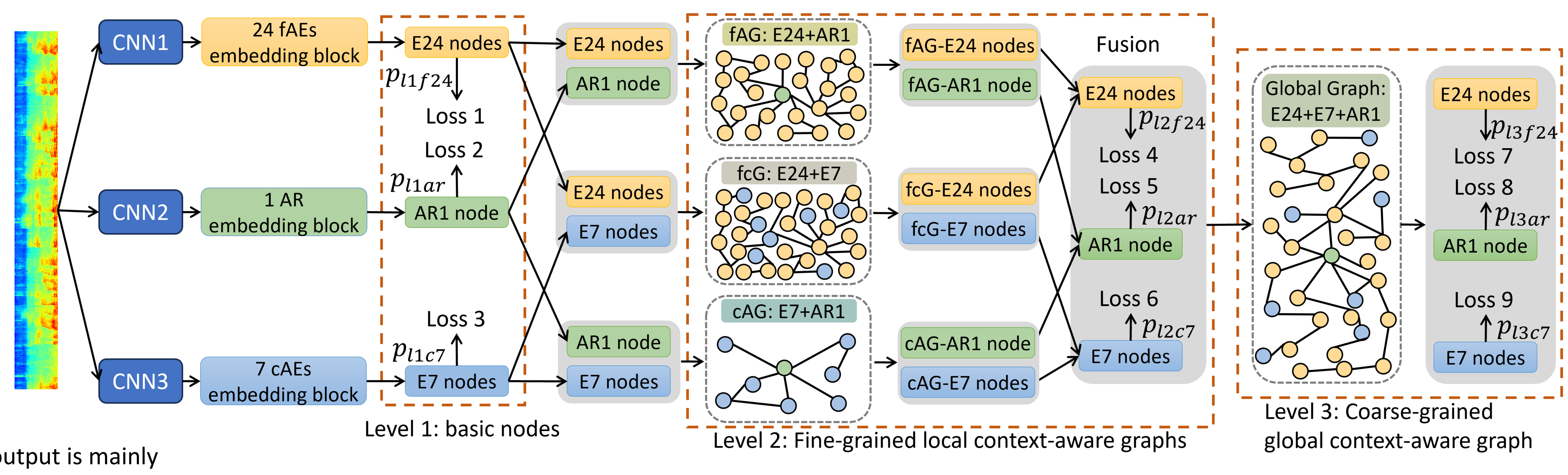
2.2.2. cAG: cAEs-AR graph

Attention-based node fusion.

Both fAG-E24 and fcG-E24 describe the same target E24 enriched by related context information in different perspectives. Hence, MLGL fuses the common information between these nodes by

$$Attention(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax(\mathbf{Q}\mathbf{K}^{\mathbf{T}}/\sqrt{d_k})\mathbf{V}$$

Here, ***Q*** acts as an index to adjust ***V***. The attention output is mainly based on ***V***, so a more informative ***V*** will lead to better attention results.



For the E24 nodes, fAG-E24 learns E24 with 1 AR node, but fcG-E24 learns E24 with 7 cAEs nodes, so fcGE24 will contain more context information.

- For fusing E24 nodes, ***Q*** is fAG-E24, ***K*** is fcG-E24, that is, ***using the AR-aware information to adjust the pure-AE information***.
- For fusing the E7 nodes, ***Q*** is cAG-E7, and ***K*** is fcG-E7.
- For fusing the AR1 node, ***Q*** is cAG-AR1, and ***K*** is fAG-AR1.

2.2. Global context-aware graph (GcG)

3. Results and Analysis

3.1. Model parameters and size.

Table 2: Comparison of HGRL and MLGL in detail.

Model	Params (M)	Model Size (MB)	Inference time (s)	AEC		ARP	
				Acc. (%)	AUC	MAE	R2
HGRL	92.3	353.0	0.531	91.71	0.901	0.802	0.458
MLGL	**4.1**	**16.0**	**0.448**	**91.96**	**0.921**	**0.706**	**0.515**

Params of MLGL are reduced by (92.3-4.1)/92.3×100%≈96%, and the model size is reduced by about 95%.

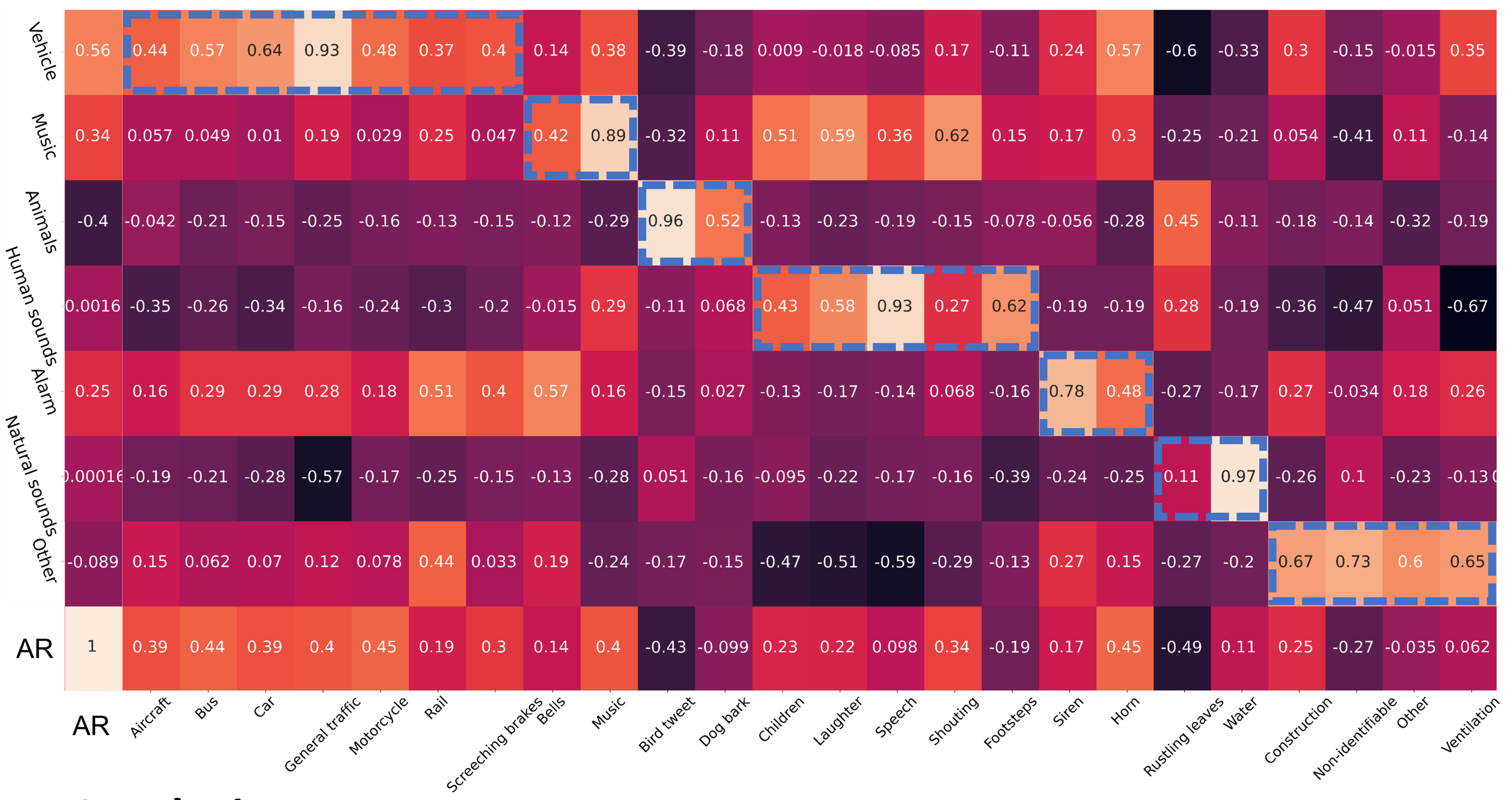
Note: ** indicates statistical significance at the 0.001 level.

Table 5: Spearman's rho correlation coefficients for AEs with AR.

#	AE	CC w/ AR	#	AE	CC w/ AR	#	AE	CC w/ AR
1	Bus	0.504**	7	Speech	0.193**	13	General traffic	0.391**
2	Car	0.199**	8	Children	0.173**	14	Motorcycle	0.374**
3	Rail	0.329**	9	Shouting	0.443**	15	Screech brakes	0.488**
4	Bells	0.458**	10	Bird tweet	-0.522**	16	Rustling leaves	-0.585**
5	Horn	0.653**	11	Dog bark	0.029**	17	Ventilation	0.085
6	Water	-0.097**	12	Aircraft	0.415**	18	Other	0.011

In Table 5, several AEs, namely ***Aircraft***, ***Bus***, ***Screeching brakes***, ***Bells***, ***Shouting***, and ***Horn*** exhibit strong positive correlations with AR, which indicates that an increase in the occurrences of these AEs increases the level of annoyance. In contrast, AEs like ***Bird tweet*** and ***Rustling leaves*** show strong negative correlations with AR, implying a decrease in annoyance when these are present. These statistical results are consistent with earlier human-perception-based soundscape research.

Fig. 2 visualizes the Pearson correlation coefficients between node embeddings. The correlations between the nodes in ***fcG*** part successfully match the associations of ***fAEs*** to ***cAEs*** in the DeLTA dataset. The correlations in the ***fAG*** part indicate that ***motorcycle***, ***bus***, and ***horn*** are the most likely sounds to cause people annoyance, while ***rustling leaves*** and ***bird tweet*** sounds are the least likely to be annoying. The cAG part shows that ***vehicles*** sounds are more likely to annoy people, while ***animals*** sounds are not.



4. Conclusions

This paper presented the MLGL to identify audio events (AEs) generated by diverse environmental sound sources and predict human-perceived annoyance rating (AR) in real-life soundscapes.

Experiments show that:

- MLGL with 4.1 M parameters works well;
- MLGL captures relations between coarse- and fine-grained AEs and AR well;
- Statistical analysis shows that some AEs significantly correlate with AR, which is consistent with previous soundscape research based on human perception.

SCAN ME

Dataset
Code
Models