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).

2.2. Local context-aware graphs (LcGs)

2.2.1. fAG: fAEs-AR graph

2.2.2. fcG: fAEs-cAEs graph

2.2.3. cAG: cAEs-AR graph

Attention-based node fusion.

2.2.4. fAE24-AR graph

Attention-based node fusion.

2.2.5. cAE7-AR graph

Attention-based node fusion.

3. Results and Analysis

3.1. Model parameters and size.

Table 2: Comparison of HGR1 and MLGL in detail.

<table>
<thead>
<tr>
<th>Model</th>
<th>Params (M)</th>
<th>Model Size (MB)</th>
<th>Inference time (s)</th>
<th>ACC (%)</th>
<th>AUC</th>
<th>MAE</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGR1</td>
<td>92.3</td>
<td>35.0</td>
<td>0.531</td>
<td>91.71</td>
<td>0.901</td>
<td>0.802</td>
<td>0.458</td>
</tr>
<tr>
<td>MLGL</td>
<td>4.1</td>
<td>16.0</td>
<td>0.438</td>
<td>91.96</td>
<td>0.921</td>
<td>0.706</td>
<td>0.515</td>
</tr>
</tbody>
</table>

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

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

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.