INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Joint Prediction of Audio Event and Annoyance Rating in an Urban Soundscape by Hierarchical Graph Representation Learning

Yuanbo Hou[1], *Siyang Song*[2], *Cheng Luo*[3], *Andrew Mitchell*[4], *Qiaoqiao Ren*[5], *Weicheng Xie*[3], *Jian Kang*[4], *Wenwu Wang*[6], *Dick Botteldooren*[1]

[1]WAVES Research Group, Ghent University, Belgium. [2]University of Leicester, UK.
[3]Shenzhen University, China. [4]University College London, UK.
[5]AIRO-IDlab, Ghent University-Imec, Belgium. [6]University of Surrey, UK.
{Yuanbo.Hou, Dick.Botteldooren}@UGent.be

Abstract

Sound events in daily life carry rich information about the objective world. The composition of these sounds affects the mood of people in a soundscape. Most previous approaches only focus on classifying and detecting audio events and scenes, but may ignore their perceptual quality that may impact humans' listening mood for the environment, e.g. annoyance. To this end, this paper proposes a novel hierarchical graph representation learning (HGRL) approach which links objective audio events (AE) with subjective annoyance ratings (AR) of the soundscape perceived by humans. The hierarchical graph consists of fine-grained event (fAE) embeddings with single-class event semantics, coarse-grained event (cAE) embeddings with multi-class event semantics, and AR embeddings. Experiments show the proposed HGRL successfully integrates AE with AR for AEC and ARP tasks, while coordinating the relations between cAE and fAE and further aligning the two different grains of AE information with the AR.

Index Terms: hierarchical graph representation learning, audio event classification, human annoyance rating prediction

1. Introduction

Audio event classification (AEC) aims to recognise predefined semantic events from audio clips to indicate whether the audio clip contains target events. AEC is useful for acoustic scene recognition [1], robot hearing [2], and monitoring [3]. Meanwhile, human annoyance rating prediction (ARP) aims to predict the annoyance rating (AR) given by humans to express their appraisal of a soundscape containing multiple undesired audio events. Joint APR and AEC can be used in soundscape design [4], human-robot interaction [5], and smart cities [6]. While AEC focuses on describing which audio events (AE) are present in the soundscape, ARP aims to identify how the combination of these sound events may induce the particular listening mood, i.e. annoyance, for humans in the soundscape. While noise annoyance at the community level has typically been explored in reference to coarse-grained sound classes, e.g. traffic and aircraft noise [7], soundscape pleasantness (considered the opposite of annoyance) more often considers specific audio events such as bird sounds [8] (occasionally even differentiating between different species of birds sounds [9]) or water fountain sounds [10]. Psychoacoustic annoyance is also defined at a coarse-grained level, originally formulated to be applied to consumer products such as vacuums, refrigerators, and car engines [11]. As such, both coarse-grained audio events (cAE) and fine-grained audio events (fAE) are considered in this paper.

To capture the high-level acoustic representations of AE, convolutional neural networks (CNN)-based models with local receptive fields have been widely proposed for AEC tasks, which show outstanding performance [12]. The convolutional recurrent neural networks (CRNN) [13, 14] combining recurrent layers, which excel in temporal modelling, further enhance the model's ability to recognize diverse AE. The Audio Spectrogram Transformer [15] has also been proposed for AEC, which outperforms the CNN-based pretrained audio neural networks (PANNs) [12] on the large-scale audio event dataset AudioSet [16], thanks to the ability of the Transformer [17] with multi-headed attention (MHA) for modelling long-term dependency information in audio clips. However, global attention in MHA may smooth out the boundaries between audio events and background noises [18]. To alleviate this problem, event-related data conditioning [18] is proposed for AEC. Audio events in real-life audio clips are usually not isolated, but exist as temporal sequences. To exploit this property, CRNN-based models [19] and contextual Transformer [20] have been proposed for sequential audio tagging.

The above studies aim to provide an objective description of the content of audio clips by exploring what audio events are present. In real life, various audio events in soundscapes can be overlapped and coupled to form polyphonic audio clips, bringing people different perceptual experiences and listening moods [21, 22]. For example, people can feel relaxed and pleasant when hearing the sound of running water and bird songs in a park scene, while they can be annoyed when hearing the noise of speeding cars and harsh horns on the street scene. Previous studies on soundscape and its impact on people, have acknowledged the importance of the source of sound but did not elaborate on matching indicators. Rather, they focused on noise levels, psycho-acoustic indicators such as calculated loudness, and various other indicators for the overall sound [23, 24]. This paper aims to combine AEC with human perception-related ARP. Furthermore, this paper explores the feasibility of predicting AR, which is a dominant performance metric for perceptual evaluation of soundscapes, based on the detected AE.

Inspired by the scene-dependent event relational graph representation method [25], this paper proposes a hierarchical graph representation learning (HGRL) method, which performs AEC and ARP tasks simultaneously, and is trained on the public DeLTA dataset [22] with 24-class AE labels and the overall human AR. To enrich the semantic information of the hierarchical graph learned by HGRL, we summarise the 24-class fAE into 7-class cAE referred to the label topology in AudioSet [16]. Next, a three-level hierarchical graph representation, which consists of the low-level fAE graph, mid-level cAE graph, and top-level human AR graph, is defined for each audio clip. The hierarchical graph with different semantic nodes and edges is input to a gated graph convolutional network (Gated GCN) [26] to classify 24-class fAE and 7-class cAE, and to predict AR. The fAE labels and AR scores are inherent in the DeLTA dataset, and the

10.21437/Interspeech.2023-1021

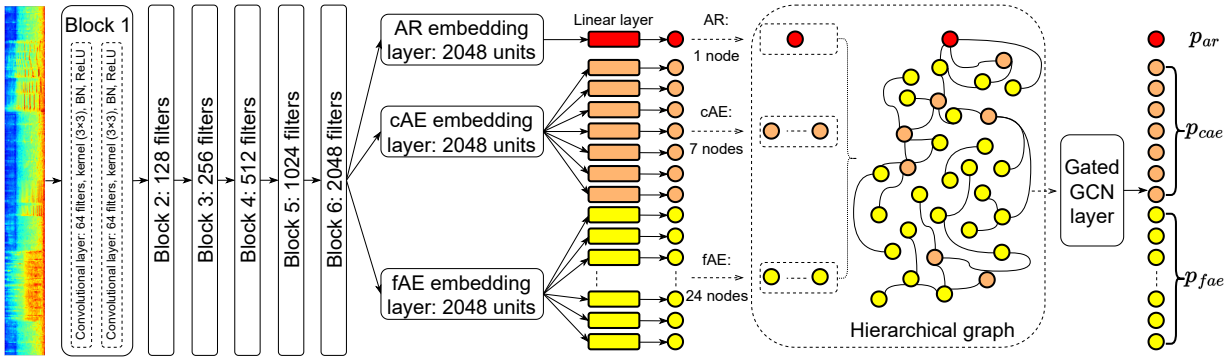

Figure 1: *The proposed hierarchical graph representation learning (HGRL) with fAE-cAE-AR (fcAR) graph.*

cAE labels are defined in Section 2.1.

The contributions are summarized as follows. 1) This paper links AEC with ARP related to human appraisal of a soundscape; 2) Inspired by the relations between fAE and human AR, this paper proposes the HGRL that models the relations of AE and their connections with AR; 3) Leveraging the knowledge from human perception and appraisal of soundscapes, this paper proposes cAE classes based on fAE and embeds them into the hierarchical graph to serve as an information exchange layer between fAE and AR; 4) This paper intuitively visualizes the learned relations between the fAE, cAE classes and AR to illustrate the ability of the hierarchical graph model to coordinate and align different levels of audio information.

2. Hierarchical graph representation learning for AEC and ARP

This section describes the dataset, the strategy of node feature extraction for AE and AR, based on which a hierarchical graph representation can be learned to perform AEC and ARP tasks.

2.1. Dataset

To the best of our knowledge, DeLTA [22] is the only publicly available dataset that 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). DeLTA comprises 2890 15-second binaural audio clips, where the training, validation, and test sets contain 2200, 245, and 445 audio clips, respectively. Based on the ontology of the event labels in AudioSet [16], 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. Semantic node feature extraction

Given an input audio clip represented in the time-domain, we first converted it into a time-frequency domain spectrogram, which is fed into convolutional blocks to extract audio representations suitable for the AEC and ARP tasks.

As illustrated in Figure 1, the employed convolutional block refers to the convolutional block structure in PANNs [12]. Each convolutional block uses a VGG-like CNN [27]. That is, the convolutional layer is repeated twice and followed by batch normalization (BN) [28], and ReLU activation functions [29]. Unlike PANN, the model proposed in Figure 1 connects 3 2048-unit embedding layers in parallel after convolutional blocks to convert high-level audio representations into separate semantic embeddings, respectively. Based on these embeddings, we can obtain the features of different nodes in the graph.

AR node. The AR embedding layer outputs a rating embedding of dimension (B, 2048), where B is the batch size. The rating embedding is mapped to an advanced rating representation by a linear layer with 64 units. Since the AR graph contains only one node, the AR representation (B, 64) is deformed to (B, 1, 64) and used as the AR node feature.

cAE nodes. To independently learn the node features of each type of cAE, the cAE embeddings with the dimension of (B, 2048) are respectively input into 7 64-unit linear layers. Among them, the output of each layer is (B, 64). After concatenating the outputs of 7 layers, a tensor with a dimension of (B, 7, 64) is obtained as the feature of the cAE nodes.

fAE nodes. The process for extracting the feature of the fAE nodes is similar to the extraction of the feature of the cAE nodes. The difference is that the fAE has 24 classes of AE, so the fAE embeddings are separately fed into 24 64-unit linear layers to capture the representation of each class of AE. Finally, the outputs of the 24 linear layers are concatenated into a (B, 24, 64) tensor as the feature of the fAE nodes.

2.3. Construction of hierarchical graphs

This paper proposes to learn a novel hierarchical graph which consists of two types of graphs as discussed next.

2.3.1. The fAE-AR (fAR) graph

As mentioned in Section 2.1, the DeLTA dataset [22] used in this paper does not contain labels on cAE. Therefore, an intuitive idea is to construct a hierarchical graph composed of fAE and AR directly. The fAE-AR (fAR) graph does not contain the cAE embedding layer and the corresponding cAE nodes shown in Figure 1, so there is no prediction vector p_{cae} of cAE either. For the fAR-based HGRL, the final loss is

$$\mathcal{L}_{\text{fAR}} = \mathcal{L}_{\text{fAE}} + \mathcal{L}_{\text{AR}} \tag{1}$$

where $\mathcal{L}_{\text{fAE}}$ is the fine-grained AE classification loss and $\mathcal{L}_{\text{AR}}$ is the AR regression loss. Given y_{fae} and y_{ar} as the true labels for fAE and AR, respectively, binary cross entropy (BCE) is used as the loss function in the fine-grained AEC, $\mathcal{L}_{\text{fAE}} = BCE(p_{fae}, y_{fae})$, and mean squared error (MSE) loss is used as the loss function in the ARP, $\mathcal{L}_{\text{AR}} = MSE(p_{ar}, y_{ar})$.

2.3.2. The fAE-cAE-AR (fcAR) graph

Based on DeLTA's 24 fAE labels, this paper proposes 7 cAE labels in Section 2.1 and assigns them to each audio clip. The cAE is the summary of fAE. Compared with a wide variety of

specific fAE, in a soundscape, people tend to perceive greater differences between sounds belonging to different cAE classes. Therefore, the cAE information can be used as an intermediate between the fAE information and the overall human perception (i.e. mood), to bridge the relations between many low-level fAEs and the top-level AR.

In the fcAR-based HGRL shown in Figure 1, we expect to investigate whether the graph-based model can learn the information about cAE nodes without the supervised labels for cAE-related output p_{cae}, and improve the performance of the model on AEC and ARP tasks. Therefore, for the fcAR-based HGRL in the unsupervised learning (UL) mode, the total loss is

$$\mathcal{L}_{\text{fcAR-UL}} = \mathcal{L}_{\text{fAE}} + \mathcal{L}_{\text{AR}} \tag{2}$$

For the fcAR-based HGRL in supervised learning (SL) mode, the cAE nodes are trained with supervision, and the total loss is

$$\mathcal{L}_{\text{fcAR-SL}} = \mathcal{L}_{\text{fAE}} + \mathcal{L}_{\text{AR}} + \mathcal{L}_{\text{cAE}} \tag{3}$$

where $\mathcal{L}_{\text{cAE}} = BCE(p_{cae}, y_{ce})$, and y_{ce} is true label for cAE.

The hierarchical graph, consisting of semantic nodes and edges, will be fed into the Gated GCN layer to further learn and update the features from the nodes and the corresponding edges by considering the information of the whole graph. The nodes features of (B, N, 64) output from the Gated GCN layer are pooled into (B, N), where N is the number of nodes in the graph. Then different nodes are directly used for the corresponding AEC and ARP tasks. Taking the fcAR-SL graph as an example, there are 32 nodes, so the final output dimension is (B, 32), the dimension of p_{ar}, p_{cae}, p_{fae} are (B, 1), (B, 7), (B, 24), respectively, and these outputs are used directly in the corresponding loss functions.

3. Experiments and results

3.1. Baseline, Experiments Setup, and Metrics

As this is the first paper performing AEC and human-related ARP regression tasks simultaneously, no reference models are available in the literature. Hence, the commonly used deep neural network (DNN), CNN, and CNN-Transformer are employed as baselines for comparison: **1)** the DNN consists of 4 fully connected (FC) layers followed by a ReLU function, the AEC and ARP layers. The number of units in each FC layer is 64, 128, 256 and 512, respectively. The output of the final FC layer is flattened and fed to the AEC and ARP layers, respectively. **2)** the CNN consists of 4 convolutional layers with (3 × 3) kernels, the AEC and ARP layers. The filters in each convolutional layer are 64, 128, 256 and 512, respectively. The output of the final convolutional layer is flattened and fed to the AEC and ARP layers, respectively. **3)** the CNN-Transformer consists of 3 convolutional layers with (3 × 3) kernels, a Transformer encoder [17], the AEC and ARP layers. Please check the project homepage (https://github.com/Yuanbo2020/HGRL) for more details.

The log-mel energy with 64 banks [30] is employed as the acoustic feature, which is extracted by the Short-Time Fourier Transform (STFT) with a Hamming window length of 46 *ms* and a window overlap of $1/3$ [1]. Dropout and normalization are used in training to prevent over-fitting of the model [31]. A batch size of 64 and Adam optimizer [32] with a learning rate of 1e-3 are used to minimize the loss. The systems are trained on a card Tesla V100 GPU for 100 epochs. Accuracy (*Acc*), *F-score* [33], and threshold-free area under curve (*AUC*) [34] are used to evaluate the classification results. Mean absolute error (*MAE*), mean square error (*MSE*) and R2-score (*R2*) [35] are used to measure the regression results. Higher *Acc*, *F-score*, *AUC*, *R2* and lower *MSE*, *MAE* indicate better performance.

3.2. Results and Analysis

Ablation study on the AE information and AR information. The proposed HGRL involves two types of foundational information: AE and AR. The fAE graph described in Section 2.3.1 only uses fAE and AR information. To investigate the impact of the individual and joint effects of the two types of information (AE and AR) on both AEC and ARP tasks, an ablation study is conducted as reported in Table 1.

Table 1: *Ablation study of the fAR-based HGRL on the test set.*

#	Information		AEC			ARP	
	fAE	*AR*	*Acc. (%)*	*F-score (%)*	*AUC*	*MSE*	*MAE*
1	✔	✗	89.654	57.253	0.863	*N/A*	*N/A*
2	✗	✔	*N/A*	*N/A*	*N/A*	1.482	0.949
3	✔	✔	**90.895**	**59.781**	**0.874**	**1.364**	**0.917**

In #1 of Table 1, since only the AE information is used, the prediction of the AR information of the corresponding graph model is not available (N/A). Similarly, the AE information in #2 is also N/A. Compared with #1 and #2, which use a single type of information, the graph-based HGRL using AE and AR information in #3 improves the performance of AEC and ARP tasks, which shows that the proposed model can effectively combine AE information with AR information, and that jointly using the AE and AR information for the proposed model is beneficial for both tasks.

Performance of different hierarchical graphs. According to the description in Section 2.3, the hierarchical graph proposed can be subdivided into 3 types: 1) fAR composed of fAE and AR embeddings; 2) fcAR-UL consisting of fAE and AR embeddings, and unsupervised learning of cAE embeddings; 3) fcAR-SL consisting of fAE and AR embeddings, and supervised learning of cAE embeddings. Table 2 details the performance of these 3 types of hierarchical graph models on AEC and ARP tasks. The convolutional layers in the feature extraction part refer to the convolutional layers in PANNs [12], which are pre-trained on a large-scale audio dataset AudioSet [16] with 527 classes of AE. In other words, the convolutional layer weights (ConW) in PANNs contain feature information from various AE. Therefore, an intuitive idea is whether introducing this diverse feature information into the feature extraction part of the proposed HGRL could improve the learning of the overall graph model.

In Table 2, the models that transfer the ConW from PANNs into HRGL's convolutional part (the rest of HGRL is randomly initialized in training) generally outperform those that do not. Therefore, using the ConW with richer representation extracted from AudioSet benefits learning HGRL. When the models are randomly initialized without using ConW, the AEC accuracy of the fcAR-based model containing 24-class fAE and 7-class cAE is slightly inferior to that of the fAR-based model without 7-class cAE. The reasons may be 1) the supervised labels for the 7-class cAE in this paper are derived from the semantic topological map of AudioSet, and are inherently inaccurate; 2) There are overlaps between the semantic labels of the 7-class cAE, which implicitly increase the difficulty of the model in identifying different cAE; 3) The feature extraction part of the proposed HGRL is inadequate for cAE composed of multiple classes of fAE, as evidenced by using ConvW for the models involving cAE in Table 2, where the AEC accuracy of both fAE and cAE is improved. The fAR-based model does not contain cAE information, so its corresponding 7-class cAE classification accuracy is N/A. In fcAR-UL, the prediction of cAE p_{cae}

can be regarded as almost random, since there is no supervision and correction for the cAE-related output p_{cae}. Nevertheless, the performance of AR significantly improves. Finally, fcAR-SL aided by ConW initialization performs best on the ARP task.

Table 2: *Performance of the proposed HGRL on the test set.*

PANNs ConvW	Hierarchical Graph		AEC *Acc.* (%)		ARP		
			24 fAE	7 cAE	*MSE*	*MAE*	*R2*
Without	fAR		90.895	*N/A*	1.364	0.917	0.296
	fcAR	UL	90.171	59.230	1.093	0.818	0.436
		SL	90.459	82.009	1.076	0.817	0.444
With	fAR		91.751	*N/A*	1.176	0.861	0.393
	fcAR	UL	**91.770**	52.901	1.079	0.822	0.443
		SL	91.713	**85.915**	**1.049**	**0.802**	**0.458**

Comparison with other models. Table 3 presents the results of various models on the DeLTA test set. The DNN composed of fully connected multilayer perceptrons has the simplest structure and the worst performance. The result of CNN is better than that of DNN, which illustrates the effectiveness of the convolutional layer for feature extraction. It is worth noting that the CNN-Transformer, which incorporates one Transformer [17] encoder with multi-head attention and residual structure, achieves better results on the *F-score* [33] of the balance of precision and recall, but for AEC accuracy, it is inferior to CNN. The reason may be that the dataset used in this paper is not large enough, and CNN-Transformer is overfitting with the training set, resulting in poor performance on the test set. Previous work [36] also shows that Transformer-based models generally outperform CNN-based models on large-scale datasets, but not on small datasets. Furthermore, Table 3 presents the results of PANNs, which achieve state-of-the-art CNN-based performance on AudioSet. Since the model in this paper performs both AEC and ARP tasks, we replace the last layer of PANNs with a parallel AEC layer and ARP layer, where the AEC layer contains 24 units with the sigmoid activation function, and the ARP layer contains 1 unit with the linear activation function. The results of PANNs in a fixed mode illustrate the importance of convolutional weights containing 527 classes of audio event knowledge for the AEC task. In fine-tuning and fixed modes, PANNs with large-scale audio event knowledge from AudioSet are highly accurate for AEC but poor for ARP. Ultimately, the proposed hierarchical graph constructed with the AE information achieves competitive results in both AEC and ARP tasks.

In-depth analysis. In the fcAR graph, it is worth exploring whether the model learns the relations between the introduced 7-class cAE, the existing 24-class of fAE, and the AR with corresponding supervision information. To answer this question, Figure 2 visualises the Pearson correlation coefficient (PCC) [37] of AE probabilities and corresponding AR outputs on the test set by the fcAR-SL instead of fcAR-UL. The prediction about cAE in fcAR-UL is randomised. The results in Figure 2

Table 3: *Comparison of different models on DeLTA dataset.*

Model	AEC		ARP		
	F-score(%)	*Acc.*(%)	*MSE*	*MAE*	*R2*
DNN	53.986	90.049	1.733	1.011	0.105
CNN	55.050	90.750	1.675	0.997	0.135
CNN-Transformer	58.667	88.942	1.445	0.966	0.254
PANNs (Fixed)	53.753	91.058	1.262	0.880	0.348
PANNs (Fine-tuning)	63.860	**91.882**	1.162	0.858	0.400
HGRL-fAR	67.428	91.751	1.176	0.861	0.393
HGRL-fcAR-UL	**68.269**	91.770	1.079	0.822	0.443
HGRL-fcAR-SL	67.911	91.713	**1.049**	**0.802**	**0.458**

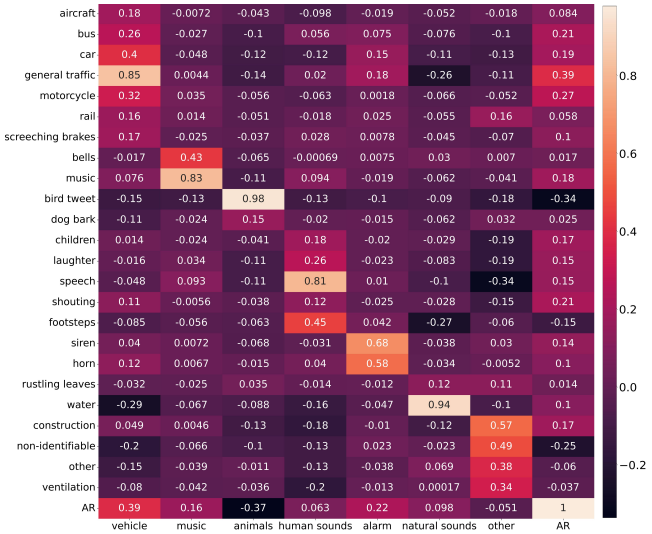

Figure 2: *Correlations of outputs by HGRL on the test set.*

show that the fcAR-based HGRL automatically aligns fAE with the cAE classes to which they belong, even if the introduced cAE information in training is inaccurate and the corresponding relations are implicit. Furthermore, the model also well coordinated the relations between 7-class cAE and AR, and between 24-class fAE and AR. Among cAEs, the most annoying AEs are the *vehicle* and *alarm* sounds. Conversely, *animals* sounds are the least likely to be annoying. Among fAEs, the most annoying AE is the *general traffic* sound, and the least annoying is the *bird tweet*. These trends are consistent with human intuitions in real life. The proposed HGRL automatically learns the relations between fAE, cAE and AR implicit in the dataset based on graph representations. In summary, the proposed HGRL successfully captures the relations between the two different grains of AE information and further aligns them with the AR information for the joint AEC and ARP tasks.

4. Conclusions

This paper presented a method for associating the audio events-related AEC task to the human mood-related ARP task, using a HGRL model consisting of fAE, cAE and AR embeddings, inspired by the human perception of AE in real-life soundscapes. Experiments on the DeLTA dataset show that: 1) The HGRL using AE and AR information can effectively combine these 2 types of information and improve the performance on both AEC and ARP tasks. 2) Compared to the fAR-based HGRL, the fcAR-based HGRL using additional cAE information and the pretrained convolutional weights achieves better results on both AEC and ARP tasks. 3) The correlation-based analysis shows that the HGRL successfully integrates the AE information with the AR information for the joint AEC and ARP tasks, while capturing the relations between two different grains of cAE and fAE information implied in the dataset and further aligning them with the AR information.

5. ACKNOWLEDGEMENTS

The WAVES and AIRO Research Groups received funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme.

6. References

[1] Y. Hou, B. Kang, W. Van Hauwermeiren, and D. Botteldooren, "Relation-guided acoustic scene classification aided with event embeddings," in *Proc. of International Joint Conference on Neural Networks*, 2022, pp. 1–8.

[2] J. Ren, X. Jiang, J. Yuan, and N. Magnenat-Thalmann, "Sound-event classification using robust texture features for robot hearing," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 447–458, 2016.

[3] R. T. Buxton, M. F. McKenna, M. Clapp, E. Meyer, E. Stabenau, L. M. Angeloni, K. Crooks *et al.*, "Efficacy of extracting indices from large-scale acoustic recordings to monitor biodiversity," *Conservation Biology*, vol. 32, no. 5, pp. 1174–1184, 2018.

[4] E. Margaritis, J. Kang, F. Aletta, and Ö. Axelsson, "On the relationship between land use and sound sources in the urban environment," *Journal of Urban Design*, vol. 25, no. 5, pp. 629–645, 2020.

[5] H.-D. Kim, J.-S. Choi, and M.-S. Kim, "Human-robot interaction in real environments by audio-visual integration," *International Journal of Control, Automation, and Systems*, vol. 5, no. 1, pp. 61–69, 2007.

[6] E.-L. Tan, F. A. Karnapi, L. J. Ng, K. Ooi, and W.-S. Gan, "Extracting urban sound information for residential areas in smart cities using an end-to-end IoT system," *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 14 308–14 321, 2021.

[7] K. D. Kryter, *The Handbook of Hearing and the Effects of Noise: Physiology, Psychology, and Public Health.* Academic Press, 1994.

[8] Y. Hao, J. Kang, and H. Wörtche, "Assessment of the masking effects of birdsong on the road traffic noise environment," *The Journal of the Acoustical Society of America*, vol. 140, no. 2, pp. 978–987, Aug. 2016.

[9] E. Ratcliffe, B. Gatersleben, and P. T. Sowden, "Bird sounds and their contributions to perceived attention restoration and stress recovery," *Journal of Environmental Psychology*, vol. 36, pp. 221–228, dec 2013.

[10] C. Trudeau, D. Steele, and C. Guastavino, "A tale of three misters: The effect of water features on soundscape assessments in a Montreal public space," *Frontiers in Psychology*, vol. 11, p. 3214, 2020.

[11] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2013, vol. 22.

[12] 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*, vol. 28, pp. 2880–2894, 2020.

[13] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *Proc. of International Joint Conference on Neural Networks*, 2017, pp. 3461–3466.

[14] W. Wei, H. Zhu, E. Benetos, and Y. Wang, "A-CRNN: A domain adaptation model for sound event detection," in *Proc. of ICASSP*, 2020, pp. 276–280.

[15] Y. Gong, Y. A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. of INTERSPEECH*, 2021, pp. 571–575.

[16] J. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," in *Proc. of ICASSP*, 2017, pp. 776–780.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez *et al.*, "Attention is all you need," in *Proc. of International Conference on Neural Information Processing Systems*, 2017, pp. 5998–6008.

[18] Y. Hou and D. Botteldooren, "Event-related data conditioning for acoustic event classification," in *Proc. of INTERSPEECH*, 2022, pp. 1561–1565.

[19] Y. Wang and F. Metze, "A first attempt at polyphonic sound event detection using connectionist temporal classification," in *Proc. of ICASSP*, 2017, pp. 2986–2990.

[20] Y. Hou, Z. Liu, B. Kang, Y. Wang, and D. Botteldooren, "CT-SAT: Contextual transformer for sequential audio tagging," in *Proc. of INTERSPEECH*, 2022, pp. 4147–4151.

[21] D. Oldoni, B. De Coensel, M. Boes, M. Rademaker, B. De Baets, T. Van Renterghem, and D. Botteldooren, "A computational model of auditory attention for use in soundscape research," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 852–861, Jul. 2013.

[22] A. Mitchell, M. Erfanian, C. Soelistyo, T. Oberman, J. Kang, R. Aldridge, J.-H. Xue, and F. Aletta, "Effects of soundscape complexity on urban noise annoyance ratings: A large-scale online listening experiment," *International Journal of Environmental Research and Public Health*, vol. 19, no. 22, p. 14872, 2022.

[23] M. Lionello, F. Aletta, and J. Kang, "A systematic review of prediction models for the experience of urban soundscapes," *Applied Acoustics*, vol. 170, p. 107479, 2020.

[24] A. Mitchell, T. Oberman, F. Aletta, M. Kachlicka, M. Lionello, M. Erfanian, and J. Kang, "Investigating urban soundscapes of the COVID-19 lockdown: A predictive soundscape modeling approach," *The Journal of the Acoustical Society of America*, vol. 150, no. 6, pp. 4474–4488, Dec. 2021.

[25] Y. Hou, S. Song, C. Yu, Y. Song, W. Wang, and D. Botteldooren, "Multi-dimensional edge-based audio event relational graph representation learning for acoustic scene classification," *arXiv preprint arXiv:2210.15366*, 2022.

[26] Y. Li, R. Zemel, M. Brockschmidt, and D. Tarlow, "Gated graph sequence neural networks," in *Proc. of International Conference on Learning Representations*, 2016.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Prof. of International Conference on Learning Representations*, 2015.

[28] J. Bjorck, C. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," in *Proc. of International Conference on Neural Information Processing Systems*, 2018, pp. 7705–7716.

[29] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *The Annals of Statistics*, vol. 48, no. 4, pp. 1875–1897, 2020.

[30] A. Bala, A. Kumar, and N. Birla, "Voice command recognition system based on MFCC and DTW," *International Journal of Engineering Science and Technology*, vol. 2, no. 12, pp. 7335–7342, 2010.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Prof. of International Conference on Learning Representations*, 2015.

[33] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[34] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[35] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 2021.

[36] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–41, 2022.

[37] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise Reduction in Speech Processing*, pp. 1–4, 2009.