PROBABLY PLEASANT? A NEURAL-PROBABILISTIC APPROACH TO AUTOMATIC MASKER SELECTION FOR URBAN SOUNDSCAPE AUGMENTATION

Kenneth Ooi, Karn N. Watcharasupat, Bhan Lam, Zhen-Ting Ong, Woon-Seng Gan

School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore
Emails: {wooi002, karn001, bhanlam, ztong, ewsgan}@ntu.edu.sg

ABSTRACT

Soundscape augmentation, which involves the addition of sounds known as "maskers" to a given soundscape, is a human-centric urban noise mitigation measure aimed at improving the overall soundscape quality. However, the choice of maskers is often predicated on laborious processes and is inflexible to the time-varying nature of real-world soundscapes. Owing to the perceptual uniqueness of each soundscape and the inherent subjectiveness of human perception, we propose a probabilistic perceptual attribute predictor (PPAP) that predicts parameters of random distributions as outputs instead of a single deterministic value. Using the PPAP, we developed a novel automatic masker selection system (AMSS), which selects optimal masker candidates based on the predicted distribution of the ISO 12913-3 *Pleasantness* score for a given soundscape. Via a large-scale listening test with 300 participants, we collected 12600 subjective responses, each to a unique augmented soundscape, to train the PPAP models in a 5-fold cross-validation scheme. Using a convolutional recurrent neural network backbone and experimenting with several variants of the attention mechanism for the PPAP, we evaluated the proposed system on a blind test set with 48 unseen augmented soundscapes to assess the effectiveness of the probabilistic output scheme over traditional deterministic systems.

***Index Terms*—** Soundscape, neural attention, soundscape augmentation, deep learning, probabilistic model

1. INTRODUCTION

It is well known that indicators based purely on the sound pressure level (SPL) of a given acoustic environment are normally insufficient in reflecting the level of annoyance and the impact to the quality of life caused by excessive noise [1]. This has led to the rise of the *soundscape* approach to noise control, which focuses on interventions that improve perceptual attributes of noise instead of simply reducing the SPL [2]. The ISO 12913 series of international standards on soundscapes aims to codify this approach by providing a circumplex model of *pleasantness* and *eventfulness*, upon which interventions can be compared, based on subjective evaluations of the surrounding acoustic environment [2–4].

Consequently, many studies have utilized soundscape augmentation techniques to alter a perception of a soundscape by adding maskers to an urban or indoor acoustic environment to optimize metrics, such as its subjectively-evaluated pleasantness or calmness. However, the choice of maskers is usually arbitrary [5], expert-guided [6], or based on post-hoc analysis [7]. An arbitrary choice of masker may be unreliable in effecting a desired perceptual change, an expert-guided choice is labor- and time-intensive, and post-hoc analyses may not be generalizable to unseen maskers and soundscapes in an unobserved context.

One way to overcome these limitations is to train a prediction model on an acoustically diverse selection of soundscapes and maskers to predict the value of some perceptual attribute, as subjectively evaluated by a person, given the raw auditory stimuli. Once trained, simulated additions of maskers to an unseen soundscape can be fed as input to the model to obtain predictions of said perceptual attribute. Then, the masker effecting the greatest increase or decrease in the attribute can be selected for in-situ augmentation as the optimal masker.

In this paper, we propose a neural approach for an automatic masker selection system (AMSS) by optimizing the *pleasantness* of an acoustic environment to augment urban soundscapes. Using a probabilistic output scheme, the models predict a distribution for the pleasantness for each soundscape-masker combination, rather than a single deterministic value, allowing for the confidence level of a prediction to be explicitly retrieved along with the predicted pleasantness.

2. RELATED WORK

Studies developing prediction models for perceptual attributes in soundscape research have primarily focused on simpler machine learning models, such as linear regression [8], support vector machines (SVM) [9], and shallow multilayer perceptrons [10]. They use input features based on acoustic measurements, psychoacoustic parameters, environmental characteristics, or some linear combination of them elucidated by principal component analysis. Lionello et al. also proposed a scaling metric for Likert scales to account for nonlinearities in the scales used to develop these systems [11].

On the other hand, a recent systematic review [12] showed that studies making use of deep neural networks to predict perceptual attribute values of soundscapes are rare, despite their prevalence in more "objective" tasks, such as sound event localization, detection, and classification. The most significant study in this respect appears to be one comparing the performances of SVM, a convolutional neural network (CNN), long short-term memory network (LSTM), and a fine-tuned VGGish network [13] on the Emo-Soundscapes dataset [14]. Emo-Soundscapes contains 1213 audio clips ranked by valence (*pleasantness*) and arousal (*eventfulness*), according to subjective paired comparisons via the Self-Assessment Manikin [14]. In [14], the fine-tuned VGGish model and the CNN trained from

This research is supported by the Singapore Ministry of National Development and the National Research Foundation, Prime Minister's Office under the Cities of Tomorrow Research Programme (Award No. COT-V4-2020-1), and the Google Cloud Research Credits Program (GCP205231017). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the view of National Research Foundation, Singapore, and Ministry of National Development, Singapore.

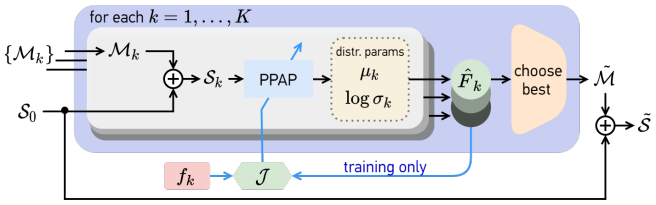

Fig. 1: Training and inference schema of the automatic masker selection system (AMSS).

scratch were respectively found to have the least mean squared errors (MSE) in predicting the valence and arousal of the audio clips. A more recent study [15] made use of similar models to classify music into one of four quadrants in a circumplex model similar to that in ISO 12913 and into the categories *neutral*, *calm*, *happy*, *sad*, *angry*, and *fearful*, using datasets totalling about a thousand audio clips. Soundscapes, however, usually contain more than just music, so it remains to be investigated if the results are generalizable to broader categories of acoustic stimuli.

In addition, these existing models in the literature are deterministic models, in the sense that the same acoustic environment is always mapped to the same predicted value without accounting for the inherent uncertainties in the subjective ratings used as ground truths. However, these ratings given by individual people are inherently random due to factors that cannot be reasonably controlled, such as the person's current stress level or noise sensitivity [16]. As such, we adopt a probabilistic approach, by training a neural network to predict a distribution across possible ratings of the output and subsequently drawing from it, allowing the model to account for the varying levels of uncertainty in each unique soundscape.

3. PROPOSED METHOD

Consider a soundscape $\mathcal{S}_0$ and a set of K candidate maskers $\{\mathcal{M}_k\}$ that can be added to the soundscape. Denote by $\mathcal{S}_k$ the *augmented* soundscape with masker $\mathcal{M}_k$. In other words, $\mathcal{S}_k$ is the soundscape $\mathcal{S}_0$ after the addition of $\mathcal{M}_k$, accounting for the response of the playback system and the real-world environment. Finding the optimal masker $\tilde{\mathcal{M}}$ to maximize the evaluation of a given perceptual (or affective) attribute f on the augmented soundscape is equivalent to finding $\operatorname{argmax}_k f(\mathcal{S}_k)$, where f could, for instance, be the rating of *pleasantness*, *eventfulness*, or some other perceptual attribute on a numerical scale. We can then augment $\mathcal{S}_0$ with $\tilde{\mathcal{M}}$ to obtain $\tilde{\mathcal{S}}$.

A naive implementation of an AMSS could, for instance, use some approximator of $f(\cdot)$ to output a predicted value $\widehat{f}_k := \widehat{f}(\mathcal{S}_k)$ of the perceptual attribute for each augmented soundscape $\mathcal{S}_k$ and pick the masker with the highest $\widehat{f}_k$. However, as mentioned earlier, this deterministic output scheme neglects the inherent uncertainty associated with subjective ratings on perceptual attribute scales. One soundscape, for instance, could be very pleasant to some but very annoying to others, while another soundscape could be nearly universally mildly pleasant to any listener, depending on the context.

To capture this uncertainty, we could instead treat f as a random variable having a joint distribution $p(f, \mathcal{S})$ with some "random" soundscape $\mathcal{S}$, where the observed augmented soundscape $\mathcal{S}_k$ is treated as a realization of $\mathcal{S}$. The uncertainty would then be represented in the variance of f. A prediction model in this scenario would then be attempting to model the conditional distribution $p(f|\mathcal{S} = \mathcal{S}_k)$ as some function $\widehat{F}(\mathcal{S}_k)$ of the augmented soundscape $\mathcal{S}_k$, which we shorten to $\widehat{F}_k$ for brevity.

We propose a neural approximator to output $\widehat{F}_k$, which we henceforth term the *probabilistic perceptual attribute predictor* (PPAP). In the proposed AMSS, shown in Fig. 1, each masker is passed to the aforementioned PPAP, which outputs the parameters of some predefined distribution family. In this work, we train a PPAP to output the mean μ_k and the log standard deviation $\log \sigma_k$ of a normal distribution $\mathcal{N}(\mu_k, \sigma_k^2)$, which we use to model the output distribution $\widehat{F}_k$. Once sufficient (or all) predicted distributions $\widehat{F}_k$ are computed, the "best" masker can be selected based on some predecided criteria. For example, a masker could be selected simply based on the highest μ_k, a set of deterministic values akin to $\{\widehat{f}_k\}$ could be sampled from $\{\widehat{F}_k\}$ to encourage masker exploration, or a more sophisticated criterion taking σ_k into account could be used.

Although only the observed values of f are available and the true distribution of f is unknown, the model still can be optimized by maximizing the log-probability of the ground truth given the output distribution, in a manner inspired by Bayesian optimization. Given a soundscape $\mathcal{S}_0$, a set of maskers $\{\mathcal{M}_k\}$, and ground truths $\{f(\mathcal{S}_k)\}$, the contribution to the loss function of the soundscape is given by

$$\mathcal{J}_{\text{prob.}}\left(\{\mathcal{M}_k\}; \{f(\mathcal{S}_k)\}\right) = -\frac{1}{K}\sum_k \mathcal{L}\left(f(\mathcal{S}_k); \mu_k, \sigma_k\right) \quad (1)$$

$$= \frac{1}{K}\sum_k \left[\frac{1}{2}\left(\frac{f(\mathcal{S}_k) - \mu_k}{\sigma_k}\right)^2 + \log \sigma_k\right], \quad (2)$$

where $\mathcal{L}$ is the log density function of the output distribution $\mathcal{N}(\mu_k, \sigma_k^2)$, omitting additive constants. During training, the model is optimized through batches of soundscape-masker pairs with available ground truths.

As seen in (2), the loss function using the normal distribution can be considered as a weighted MSE loss regularized by the log standard deviation. This loss function is inherently stable with respect to σ_k as the first term encourages larger σ_k while the second term encourages smaller σ_k. Of course, some other choices of the output distribution may reduce similarly to other deviation measures, such as the Laplace distribution reducing to a regularized mean absolute error, although the investigation for the choice of distribution is left for future work. Note that it is also possible to force to the model to learn deterministically by setting σ_k to some predetermined constant and training the model using the pure MSE loss between the ground truth and μ_k. For the purpose of an ablation study, we use

$$\mathcal{J}_{\text{det.}}\left(\{\mathcal{M}_k\}; \{f(\mathcal{S}_k)\}\right) = \frac{1}{2K}\sum_k \left(f(\mathcal{S}_k) - \mu_k\right)^2, \quad (3)$$

as the deterministic counterpart to (2) in this work. The deterministic loss in (3) can be thought of as (2) with a static $\sigma_k = 1$.

4. VALIDATION EXPERIMENTS

To validate the proposed system, we let f be the normalized pleasantness measure as defined by the ISO 12913-3 standard [4], which we shall refer to as *ISO Pleasantness*. Specifically, we have

$$f(\mathcal{S}_k) = \frac{2(r_{\text{pl}} - r_{\text{an}}) + \sqrt{2}(r_{\text{ca}} - r_{\text{ch}} + r_{\text{vi}} - r_{\text{mo}})}{8(1 + \sqrt{2})} \in [-1, 1] \quad (4)$$

where $r_{\text{pl}}, r_{\text{an}}, r_{\text{ca}}, r_{\text{ch}}, r_{\text{vi}}, r_{\text{mo}} \in \{1, 2, 3, 4, 5\}$ are the extent to which a participant considers the augmented soundscape $\mathcal{S}_k$ to be pleasant, annoying, calm, chaotic, vibrant, and monotonous, respectively, on a 5-point Likert scale [4]. For each $\mathcal{S}_k$, we predict the

distribution $\widehat{F}_k = \mathcal{N}(\mu_k, \sigma_k^2)$, with the 'ground-truth' labels $f(\mathcal{S}_k)$ being the ISO Pleasantness rating for $\mathcal{S}_k$ given by the participant. As an ablation study, we compare models trained with our proposed method against deterministic models that predicts $f(\mathcal{S}_k)$ via μ_k directly. The validation experiments were conducted on a dataset of subjective responses to a variety of augmented soundscapes.

4.1. Dataset

Our dataset contains 12 600 unique sets of subjective responses to augmented soundscapes in a 5-fold cross-validation set, as well as 48 additional sets of responses in an independent test set, for a total of 12 648 samples. Each sample maps an augmented soundscape (as a raw audio recording) to an ISO Pleasantness value given the set of responses to the following 6-item subset of the ISO 12913-2 affective response questionnaire [3]:

> *To what extent do you agree or disagree that the present surrounding sound environment is {pleasant, chaotic, vibrant, calm, annoying, monotonous}?*

Participants responded on a 5-point scale with the labels "Strongly disagree", "Disagree", "Neither agree nor disagree", "Agree", and "Strongly agree", which were respectively coded as 1, 2, 3, 4, and 5. The coded responses were then used in (4) to compute the ground-truth labels of ISO Pleasantness for each augmented soundscape[1].

4.1.1. Augmented soundscapes

The augmented soundscapes in the 5-fold cross-validation set were made by adding 30-second excerpts of recordings from Freesound and xeno-canto as "maskers" to 30-second excerpts of binaural recordings of the soundscapes from the Urban Soundscapes of the World (USotW) database [17], a comprehensive dataset of urban soundscapes. The unaugmented soundscapes were also included as controls. All recordings used were sampled at 44.1 kHz.

Each fold has a bank of 56 maskers in the following classes: bird (16), construction (8), traffic (8), water (16), and wind (8). The classes were chosen to cover the range of sound types generally evaluated to be pleasant [7] and annoying [18]. Maskers and soundscapes in each fold are disjoint.

The augmented soundscapes in the test set were made in a similar fashion with 7 maskers[2] (8 including the unaugmented control) independent of those from the cross-validation set and 6 binaural recordings of soundscapes independent of the USotW dataset, recorded using the same Soundscape Indices Protocol [19]. The maskers were added exhaustively to all soundscapes for the test set, resulting in 48 samples in the test set. All soundscapes were calibrated according to the method described in [20] to the measured in-situ[3] A-weighted equivalent SPL ($L_{\mathrm{A,eq}}$) before adding the maskers at a constant soundscape-to-masker ratio of 0 dB for the test set, and a randomly-selected value, in dB, from $\{-6, -3, 0, 3, 6\}$ for the cross-validation set.

[1] Ground-truth labels, log-mel spectrograms of the augmented soundscapes, and further details on our participants and the dataset are provided at doi: 10.21979/N9/YSJQKD

[2] The 7 maskers were excerpted from xeno-canto track IDs 640568 (bird "B1") and 568124 (bird "B2"), as well as Freesound track IDs 586168 (construction "Co"), 587219 (traffic "Tr"), 587000 (water "W1"), 587759 (water "W2"), and 587205 (wind "Wi")

[3] We would like to thank the late Prof. Bert De Coensel and the WAVES Research Group at Ghent University, Ghent, Belgium, for generously providing the $L_{\mathrm{A,eq}}$ values for the binaural tracks in the USotW database.

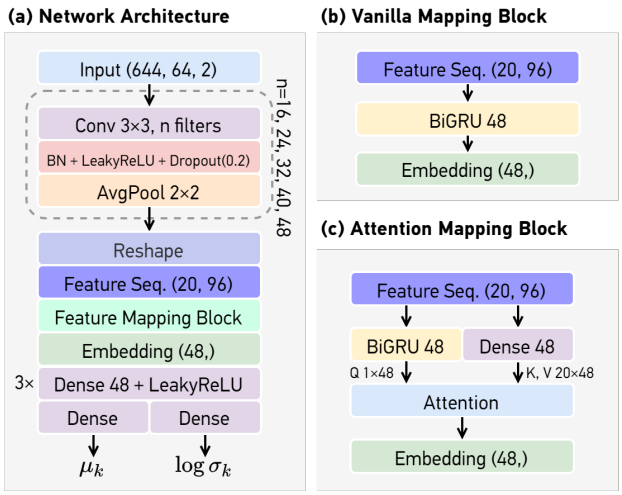

Fig. 2: Model architecture. The feature mapping block in (a) is implemented via either (b) or (c).

4.1.2. Subjective responses

To obtain subjective responses to the augmented soundscapes, we recruited 300 participants to each rate 42 unique, randomly-chosen augmented soundscapes from a fold of the validation set. We also recruited 5 other participants to each rate the 48 augmented soundscapes in the test set. Each augmented soundscape in the validation and test set is therefore is rated by one and five participants, respectively. All participants listened to the calibrated augmented soundscapes on a pair of circumaural headphones (Beyerdynamic Custom One Pro), powered by an external sound card (Creative SoundBlaster E5). After listening to each augmented soundscape, they answered the 6-item questionnaire described in Section 4.1, and the ISO Pleasantness was calculated from their responses according to (4).

4.2. Model and training

In this work, we use log-mel spectrograms of the augmented soundscapes as inputs. The log-mel spectrograms were extracted using a 4096-sample Hann window with 50 % overlap and compressed to 64 mel bins. Fig. 2(a) shows the base convolutional recurrent neural network (CRNN) architecture used throughout the paper. We experimented with four different feature mapping blocks, namely a vanilla mapping block using a bidirectional gated recurrent unit (BiGRU), an additive attention block [21], a dot-product attention block [22], and a multi-head attention block with 4 heads [23]. The vanilla block is shown in Fig. 2(b). All attention-based blocks share the same general flow shown in Fig. 2(c). The last layers of the model consist are dense layers which finally output μ_k and $\log \sigma_k$. In the deterministic ablation models, $\log \sigma_k$ is ignored.

All models were trained using a 5-fold cross-validation scheme, with 10 models per fold, to a total of 50 models. Each fold uses the same 10 seeds for the models. All models were trained up to 100 epochs using an Adam optimizer with a learning rate of 5×10^{-5}. For each model, the model weights with the best validation loss are used for evaluation in both the cross-validation set and the test set.

5. RESULTS AND DISCUSSION

Table 1 summarizes the results of the validation experiments described in Section 4 on both the 5-fold cross validation set and the

Model	Params.	Cross-Validation Set			Test Set		
		Deterministic	Probabilistic	Improv. (%)	Deterministic	Probabilistic	Improv. (%)
(Label mean)	—	0.1556	—	—	0.1181	—	—
PPAP w/o attention	76K	0.1262±0.0009	0.1255±0.0011	0.5	0.1106±0.0114	0.1020±0.0096	7.8*
PPAP w/ additive att.	85K	0.1264±0.0015	0.1251±0.0014	1.0*	0.1030±0.0137	0.0970±0.0080	5.8*
PPAP w/ dot-prod. att.	85K	0.1259±0.0012	0.1250±0.0013	0.7*	0.0950±0.0131	0.0906±0.0091	4.6
PPAP w/ multi-head att.	123K	0.1244±0.0009	0.1230±0.0010	1.1*	0.0999±0.0147	0.0996±0.0104	0.3

Table 1: Mean fold MSEs of the PPAP models (± standard deviation) over the 10 runs tested for each setting. Each run consists of five models, each on a different fold of the cross-validation set but with the same initial conditions. For all models, the contributions to the MSEs were calculated as $(\mu_k - f_k)^2$. Asterisks $(*)$ denote statistically significant improvements ($p < 0.05$).

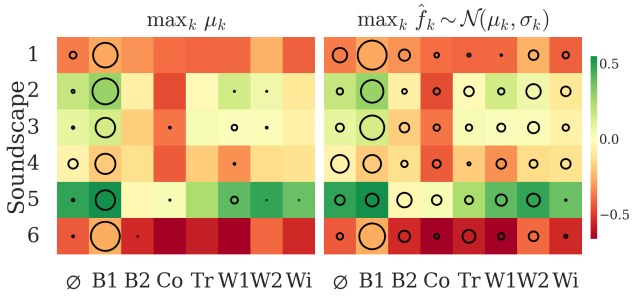

Fig. 3: Masker selection using the 50 models (dot-product variant) on the test set, with a naive maximum μ_k selection scheme (left) and a random sampling scheme to encourage masker exploration (right). The selection criterion for each base soundscape is shown on top of each subplot. The cell color indicates the average *ISO Pleasantness* rated by the test-set participants. The circle size indicates the number of models selecting the masker. See Footnote 2 for masker descriptions.

independent test set. As a reference, we also provide the results of a trivial, deterministic "label mean model", where the mean of labels in the training set is used as the prediction for all stimuli in the validation and test set. All other investigated models performed better than the label mean model, thus indicating meaningful feature extraction by the models. The test set MSEs were also lower than that of the validation set MSEs. This is likely due to the fact that, compared to the validation set (12 600 samples, 280 maskers), the test set was smaller and less diverse (48 samples, 7 maskers).

Furthermore, for the four architectures tested, the models with probabilistic output performed better than the identical models with deterministic output for both the validation and test set. The percentage reduction in MSE for each model is shown in Table 1 as well. We can observe that the CRNN with the multi-head attention block and the vanilla CRNN respectively experienced the greatest improvement in validation and test set MSE of 1.1 % and 7.8 %.

To quantify the significance of these reductions, we performed two-sided Wilcoxon signed-rank tests between the deterministic models and probabilistic models for both the validation set MSEs and test set MSEs. The reductions that were significant at a 5 % significance level are marked with asterisks in Table 1. The reductions in the test set MSEs for the vanilla CRNN and PPAP with additive attention are statistically significant and provide evidence to support the generalizability of our proposed method to unseen data. This may be because the inherent randomness contributed by the test set participants differs from participants in the training set, and the PPAP generalizes over this randomness better. Nonetheless, a further ablation study would be required to validate this hypothesis.

However, the lowest improvement in test set MSE was actually observed in the PPAP with multi-head attention block. This may be due to the fact that the PPAP with multi-head attention block has about 50 % more parameters than the other three models (about 120K against about 80K), which could have caused the trained models (both deterministic and probabilistic) to be overfitted to our comparatively smaller dataset of 12 600 training and validation samples.

Figure 3 shows the maskers selected by the AMSS system based on the 50 PPAP models with dot-product attention using two different masker selection criteria. In the naive selection scheme, most models select the first bird masker (B1) across all base soundscapes. This is consistent with the consensus in previous soundscape studies [24–27], which generally observed more pleasant responses to birdsong maskers. However, the second bird masker (B2) might have generally received lower ratings by participants and correspondingly low vote counts by the models due to it being considered less pleasant *in the context* of the test set soundscapes, which has been observed in a previous study [29].

In the random selection scheme, the system still tends to select B1 which gives a good pleasantness score for most of the soundscapes, but is now exploring other maskers significantly more than the naive scheme. This can be useful in a real-time system where human feedback can be obtained in-situ [28] to adaptively adjust the masker selection or improve future models. It can also be seen that with Soundscape 5, where the *Pleasantness* of the unaugmented soundscape is already relatively high, the exploration rate is higher than other base soundscapes, allowing a more diverse acoustic experience with only a small compromise on the pleasantness level.

6. CONCLUSION

In this work, we proposed an automatic masker selection system (AMSS) for human-centric urban soundscape augmentation, using a probabilistic perceptual attribute predictor (PPAP). The proposed PPAP was implemented using a convolutional recurrent neural network and trained to output predictions in a probabilistic manner. This allowed it to simultaneously predict the perceptual attribute of a soundscape while accounting for the inherent randomness in human subjective perception of acoustic stimuli. Via a large-scale listening test with more than 300 participants and more than 12K unique soundscapes, we validated the effectiveness of our PPAP in predicting the pleasantness of augmented soundscapes, including those generated from unseen soundscapes and maskers. Future works on the proposed AMSS include in-situ implementations to assess its ecological validity, as well as investigation of the proposed method on other perceptual attributes (f), such as *eventfulness* or *calmness*, because our proposed method is not specific to any particular attribute. Indeed, since the primary assumption underpinning the PPAP is that of random ground-truth labels, one may also conceivably apply it to any context where predictions of subjective evaluations are desired.

7. REFERENCES

[1] J. Kang, F. Aletta, T. Oberman, M. Erfanian, M. Kachlicka, M. Lionello, and A. Mitchell, "Towards soundscape indices," in *Proc. Int. Congr. Acoust.*, 2019, pp. 2488–2495.

[2] International Organization for Standardization, "ISO 12913-1 Acoustics. Soundscape Part 1: Definition and conceptual framework," 2014.

[3] ——, "ISO/TS 12913-2 Acoustics. Soundscape Part 2: Data collection and reporting requirements," 2018.

[4] ——, "ISO/TS 12913-3 Acoustics. Soundscape Part 3: Data analysis," 2019.

[5] T. M. Leung, C. K. Chau, and S. K. Tang, "On the study of effects on different types of natural sounds on the perception of combined sound environment with road traffic noise," in *Proc. 45th Int. Congr. Expo. Noise Control. Eng.*, 2016, pp. 1764–1770.

[6] J. Y. Hong, B. Lam, Z. T. Ong, K. Ooi, W. S. Gan, J. Kang, S. Yeong, I. Lee, and S. T. Tan, "A mixed-reality approach to soundscape assessment of outdoor urban environments augmented with natural sounds," *Build. Environ.*, vol. 194, no. February, p. 107688, 2021.

[7] T. van Renterghem, K. Sun, K. Filipan, K. Vanhecke, T. de Pessemier, B. de Coensel, W. Joseph, and D. Botteldooren, "Interactive soundscape augmentation of an urban park in a real and virtual setting," *Proc. Int. Congr. Acoust.*, pp. 899–903, 2019.

[8] K. Sun, K. Filipan, F. Aletta, T. van Renterghem, T. de Pessemier, W. Joseph, D. Botteldooren, and B. de Coensel, "Classifying urban public spaces according to their soundscape," *Proc. Int. Congr. Acoust.*, pp. 6100–6105, 2019.

[9] T. Giannakopoulos, M. Orfanidi, and S. Perantonis, "Athens Urban Soundscape (ATHUS): A Dataset for Urban Soundscape Quality Recognition," *Proc. MultiMedia Model. 25th Int. Conf.*, pp. 338–348, 2019.

[10] L. Yu and J. Kang, "Modeling subjective evaluation of soundscape quality in urban open spaces: An artificial neural network approach," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1163–1174, 2009.

[11] M. Lionello, F. Aletta, A. Mitchell, and J. Kang, "Introducing a Method for Intervals Correction on Multiple Likert Scales: A Case Study on an Urban Soundscape Data Collection Instrument," *Front. Psychol.*, vol. 11, no. January, 2021.

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

[13] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 131–135.

[14] J. Fan, F. Tung, W. Li, and P. Pasquier, "Soundscape emotion recognition via deep learning," *Proc. 15th Sound Music. Comput. Conf.*, pp. 289–294, 2018.

[15] E. Koh and S. Dubnov, "Comparison and analysis of deep audio embeddings for music emotion recognition," in *Proc. 4th AAAI Workshop Affect. Content Analysis*, 2021, pp. 15–22.

[16] F. Aletta, T. Vander Mynsbrugge, P. Thomas, K. Filipan, D. Botteldooren, M. Petrovic, P. de vriendt, D. Van de Velde, and P. Devos, "The relationship between noise sensitivity and soundscape appraisal of care professionals in their work environment: a case study in Nursing Homes in Flanders, Belgium," *Proc. 11th Eur. Congr. Expo. Noise Control. Eng.*, 2018.

[17] B. De Coensel, K. Sun, and D. Botteldooren, "Urban Soundscapes of the World: Selection and reproduction of urban acoustic environments with soundscape in mind," *Proc. 46th Int. Congr. Expo. Noise Control. Eng.*, 2017.

[18] C. K. Chau, T. M. Leung, J. M. Xu, and S. K. Tang, "Modelling noise annoyance responses to combined sound sources and views of sea, road traffic, and mountain greenery," *J. Acoust. Soc. Am.*, vol. 144, no. 6, pp. 3503–3513, 2018.

[19] A. Mitchell, T. Oberman, F. Aletta, M. Erfanian, M. Kachlicka, M. Lionello, and J. Kang, "The soundscape indices (SSID) protocol: A method for urban soundscape surveys- Questionnaires with acoustical and contextual information," *Appl. Sci.*, vol. 10, no. 7, pp. 1–27, 2020.

[20] K. Ooi, Y. Xie, B. Lam, and W. S. Gan, "Automation of binaural headphone audio calibration on an artificial head," *MethodsX*, vol. 8, no. February, p. 101288, 2021.

[21] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015.

[22] M. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2015, pp. 1412–1421.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. Conf. Neural Inf. Process. Syst.*, pp. 5999–6009, 2017.

[24] B. D. Coensel, S. Vanwetswinkel, and D. Botteldooren, "Effects of natural sounds on the perception of road traffic noise," *J. Acoust. Soc. Am.*, vol. 129, no. 4, pp. EL148–EL153, 2011.

[25] Y. Hao, J. Kang, and H. Wörtche, "Assessment of the masking effects of birdsong on the road traffic noise environment," *J. Acoust. Soc. Am.*, vol. 140, no. 2, pp. 978–987, 2016.

[26] W. Zhao, H. Li, X. Zhu, and T. Ge, "Effect of birdsong soundscape on perceived restorativeness in an Urban Park," *Int. J. Environ. Res. Public Heal.*, vol. 17, no. 16, pp. 1–15, 2020.

[27] J. Y. Hong, Z.-T. Ong, B. Lam, K. Ooi, W.-S. Gan, J. Kang, J. Feng, and S.-T. Tan, "Effects of adding natural sounds to urban noises on the perceived loudness of noise and soundscape quality," *Sci. Total. Environ.*, vol. 711, p. 134571, 2020.

[28] F. A. Karnapi, B. Lam, K. Ooi, Y.-T. Lau, K. Watcharasupat, T. Wong, W.-S. Gan, J. Hong, S. Yeong, and I. Lee, "Development of a feedback interface for in-situ soundscape evaluation," in *Proc. 50th Int. Congr. Expo Noise Control. Eng.*, 2021.