inter·noise 2023
CHIBA,GREATER TOKYO 20-23 AUGUST

ARAUSv2: An Expanded Dataset and Multimodal Models of Affective Responses to Augmented Urban Soundscapes

Kenneth Ooi[1], Zhen-Ting Ong[2], Bhan Lam[3], Trevor Wong[4], Woon-Seng Gan[5]
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
50 Nanyang Avenue, Singapore 639798

Karn N. Watcharasupat[6]
Center for Music Technology, Georgia Institute of Technology, USA
840 McMillan Street NW, Atlanta, GA 30332

ABSTRACT

The ARAUS (Affective Responses to Augmented Urban Soundscapes) dataset consists of a five-fold cross-validation set and independent test set of subjective perceptual responses to augmented soundscapes presented as audio-visual stimuli. However, key limitations in its original release included a disproportionate number of participants being young university students and a relatively small test set. We aim to address this by publishing ARAUSv2, which adds responses from participants to the cross-validation set from an older, non-student population, as well as responses from additional participants in a substantially larger test set consisting of new urban soundscapes recorded in a variety of settings in Singapore. The additional responses were collected in a similar fashion as the initial release, with participants rating augmented soundscapes (made by digitally adding maskers to urban soundscape recordings) on how pleasant, annoying, eventful, uneventful, vibrant, monotonous, chaotic, calm, and appropriate they were. We also present a sample of multimodal prediction models for the ISO Pleasantness and Eventfulness of the augmented soundscapes in ARAUSv2. The multimodal models use participant-linked information such as demographics and responses to psychological questionnaires, as well as visual information from the stimuli, which the baseline models presented in the initial ARAUS dataset did not utilize.

1. INTRODUCTION

Standardized soundscape datasets have been a mainstay of soundscape research, due to their ability to allow for fair and straightforward benchmarking of soundscape prediction models, as well as evaluation of soundscape intervention methods such as soundscape augmentation, where maskers are added via physical or electroacoustic means to a given soundscape to alter its perceptual parameters. These datasets can be a collection of unlabeled, general-purpose recordings in a format compatible with the soundscape standard ISO/TS 12913-2:2018 [1], like EigenScape [2], the Urban Soundscapes of the World (USotW) database [3], and the Lion City Soundscapes (LCS) dataset [4].

Alternatively, the recordings can be labeled with additional metadata, such as responses provided by human listeners on various perceptual attribute scales such as the perceived loudness, sound source dominance, or the standard affective attributes detailed in ISO/TS 12913-3:2019 [5].

[1]wooi@e.ntu.edu.sg [2]ztong@ntu.edu.sg [3]bhanlam@ntu.edu.sg [4]trev0006@e.ntu.edu.sg [5]ewsgan@ntu.edu.sg
[6]kwatcharasupat@gatech.edu

Such datasets include the International Soundscape Database [6], the Affective Responses to Augmented Urban Soundscapes (ARAUS) dataset [7], and a crowdsourced indoor soundscape dataset described by Versümer et al. [8]. These datasets include not only responses related to the perceived affective quality of the soundscapes in the datasets, but also auxiliary information that may be pertinent to garner a better understanding of the soundscape evaluations, such as that related to the ambient environment, participant demographics, and situational variables, respectively. This reflects the definition of a soundscape in ISO 12913-1:2014 as an "acoustic environment, as perceived... *in context*" [9], since the auxiliary information provides the aforementioned context.

Consequently, information from non-acoustic modalities must also be considered to develop robust models for soundscapes. For example, the visual environment can significantly affect the perception of the acoustic environment, as evidenced by a study by Hong and Jeon, where the same audio clip of traffic noise had a significantly higher mean relative preference score when combined with a visual environment of a pavement beside a road with an added green hedge blocking the view of the road, in contrast to the visual environment without the added green hedge [10]. Incidentally, a systematic review of 30 studies investigating audio-visual interactions concluded that the extent of greenery and water features were the most significant visual features affecting auditory perception [11]. A Gaussian mixture model taking as input distances from key landmarks to predict sound pressure levels (SPL) was also observed to have improved classification accuracy when segmented images from pre-trained convolutional neural networks (CNNs) were used to classify the soundscape quality, restorativeness, and pleasantness as "good", "medium", or "poor" [12].

Physiological, psychological, or demographic factors related to the listener experiencing the soundscape can also affect soundscape perception. For instance, Wang et al. found that age, in conjunction with geospatial parameters of the surrounding environment, had almost as high relative importance as acoustic factors in a random forest model developed to predict acoustic comfort [13]. Another study investigating the acceptability of aircraft noise (presented as 30-second clips) in Denali National Park and Preserve established that the listener's interest to take a commercial flightseeing tour over the area significantly increased aircraft noise acceptability. [14].

Nonetheless, acoustic information is by far the most crucial mode to consider in soundscape perception. For instance, a study by Durbridge and Murphy found that the perceived dominance of sound sources in several classes (natural, mechanical, human) had far higher correlations with valence and arousal ratings (on a 5-point self-assessment mannikin scale) from participants than a variety of metrics related to heart rate variability, which had correspondingly insignificant correlations [15]. The complexity of a given soundscape (as defined by the number of sound sources present, regardless of source type) was also found to influence noise annoyance perception in a large-scale online experiment with over 1,200 participants [16].

At present, there is much interest in the use of machine learning and deep learning techniques to craft prediction models for soundscape perception. For example, a convolutional neural network with transfer learning was used to predict the subjectively-rated annoyance of various road traffic noise samples on the 11-point scale according to ISO 15666:2021 [17]. Attention-based deep neural networks have also been explored as a method to fuse information from multiple modalities, namely from raw images and numerically-coded information related to the listener, to make predictions of the ISO Pleasantness of a given soundscape [18]. This motivated the initial publication of the ARAUS dataset (which we refer to as "ARAUSv1" in this manuscript) as described in [7], with 25,440 unique subjective perceptual responses to augmented soundscapes presented as audio-visual stimuli in 5 cross-validation folds and an independent test set.

However, due to logistical considerations in participant recruitment, ARAUSv1 has a large number of participants being young university students, which may cause models trained on it to be biased towards the predilections of this particular demographic. The test set in ARAUSv1 also consists of responses collected using only 5 participants, 6 soundscapes, and 7 non-silent maskers, which would further exacerbate any bias when using it for development.

Hence, we propose an update of the dataset from ARAUSv1 to ARAUSv2, for which we describe the data collection methodology in Section 2 and that consists of the following:

1. **ARAUSv2 test set**: two new folds (i.e., folds 6 and 7) of responses independent of each other and all ARAUSv1 cross-validation folds, which comprise a significantly larger number of responses than the ARAUSv1 test set, and
2. **ARAUSv1 extension**: additional responses to the ARAUSv1 cross-validation folds themselves from an older, non-student population.

As an illustration, we also compare a selection of multimodal models (trained on ARAUSv1 and/or ARAUSv2, as described in Section 3) as potential use cases for the updated dataset, and present the corresponding results in Section 4. Lastly, we make some concluding remarks in Section 5.

2. DATA COLLECTION METHODOLOGY

Data collection for ARAUSv2 largely follows the procedure described for ARAUSv1 in [7], so we only highlight differences and key similarities in this manuscript. As of the time of writing, data collection is still ongoing, so we present results based on a partial set of responses in ARAUSv2 here.

As with ARAUSv1, all augmented soundscapes for ARAUSv2 are presented to participants as audio-visual stimuli generated by augmenting "base" urban soundscapes with maskers from the Freesound [19] and Xeno-Canto [20] libraries via digital addition at varying soundscape-to-masker ratios (SMR). The research protocols used for data collection were approved by the Institutional Review Board of Nanyang Technological University (Ref. IRB-2020-08-035).

2.1. Base Urban Soundscapes

For the ARAUSv1 extension, the same USotW soundscapes as ARAUSv1 were used as the base urban soundscapes. However, for the ARAUSv2 test set, a subset of the 1-minute excerpts of the binaural recordings of soundscapes from the Lion City Soundscapes (LCS) dataset were used as the base urban soundscapes. The 1-minute excerpts from the LCS dataset were cropped out from extended recordings at 62 different recording locations in Singapore, and were chosen by a modified k-medoids algorithm to span the space generated by the "Pleasantness" and "Eventfulness" axes in ISO/TS 12913-3:2019 with the use of a "representative" loss function [4]. Hence, these 1-minute excerpts are suitable for use for the ARAUSv2 test set since they capture a good variety of soundscapes, which is desirable for the final evaluation of the generalizability of a given prediction model.

Each 1-minute binaural recording in the LCS was split into two halves of 30 seconds for the creation of the audio-visual stimuli in the ARAUSv2 test set. Similar to ARAUSv1, we also discarded any recording with (1) audible electrical noise, (2) measured in-situ $L_{\mathrm{A,eq}}$ values below 52 dB, and/or (3) measured in-situ $L_{\mathrm{A,eq}}$ values above 77 dB. This gave 96 urban soundscape recordings (each 30 seconds in length) for the ARAUSv2 test set from the 48 original excerpts (each 1 minute in length) shown in Table 1. The fold allocation procedure is similar to ARAUSv1, as described in Section 2.3, and there is almost an identical number of base urban soundscapes per fold in the ARAUSv2 test set and the ARAUSv1 cross-validation set.

2.2. Maskers

For the ARAUSv1 extension, the same maskers from Freesound and Xeno-canto as ARAUSv1 were used. However, for the ARAUSv2 test set, a new set of source tracks from Freesound and Xeno-canto were used to generate the 30-second maskers using the same procedure as ARAUSv1. These new source tracks are shown in Table 2, and are mutually disjoint from each other as well as all source tracks that were used in ARAUSv1, although they come from the same bird, construction, traffic, water, and wind classes. The proportion of masker recordings in each class for the ARAUSv2 test set is identical to that in ARAUSv1, and the fold allocation procedure is as described in Section 2.3.

Table 1: Lion City Soundscapes (LCS) recordings used for ARAUSv2 test set by fold. Recordings are denoted using their index numbers in the LCS dataset. In contrast to ARAUSv1, both halves of all recordings were used.

Fold	**LCS identifiers**							
Fold 6	S0010	S0011	S0013	S0015	S0016	S0022	S0023	S0024
	S0026	S0028	S0029	S0031	S0032	S0039	S0044	S0045
	S0048	S0049	S0052	S0054	S0055	S0057	S0058	S0060
Fold 7	S0002	S0003	S0005	S0006	S0008	S0012	S0014	S0019
	S0021	S0025	S0027	S0030	S0035	S0036	S0037	S0038
	S0040	S0043	S0046	S0047	S0050	S0051	S0056	S0062

Table 2: Tracks from Freesound (denoted as "FS") and Xeno-canto (denoted as "XC") used as additional maskers in the ARAUSv2 test set by fold. Numbers after "FS" and "XC" denote the index numbers of the tracks in the Freesound and Xeno-canto databases, respectively.

	Masker track identifier						
Fold	**Bird**		**Construction**	**Traffic**	**Water**		**Wind**
6	XC484364	XC671231	FS57782	FS99812	FS611502	FS677228	FS17136
	XC588179	XC671347	FS222037	FS364813	FS651211	FS677722	FS181252
	XC599689	XC680927	FS272170	FS425624	FS654444	FS677723	FS181254
	XC615753	XC717733	FS675206	FS570535	FS655507	FS681653	FS184175
	XC622342	XC757935	FS675536	FS675315	FS656430	FS683020	FS199517
	XC644286	XC764466	FS676034	FS676802	FS670618	FS683711	FS670307
	XC649712	XC767717	FS680040	FS681455	FS673106	FS683913	FS684202
	XC655262	XC777420	FS683111	FS683325	FS674074	FS684383	FS684545
7	XC561828	XC696487	FS169163	FS172521	FS638070	FS680315	FS17292
	XC562719	XC707096	FS564724	FS418436	FS654601	FS680842	FS107934
	XC570705	XC707329	FS675147	FS537544	FS660252	FS683016	FS181250
	XC576967	XC730600	FS677261	FS679741	FS661806	FS683330	FS250138
	XC578071	XC737262	FS679055	FS682530	FS667786	FS683707	FS421127
	XC581193	XC773505	FS679739	FS682737	FS672723	FS683949	FS528944
	XC602050	XC776354	FS682593	FS683765	FS675566	FS684023	FS562433
	XC670686	XC789701	FS683019	FS684103	FS679106	FS684206	FS683858

2.3. Fold Allocation

For the ARAUSv1 extension, since the same base urban soundscapes and maskers were used as ARAUSv1, their fold allocation was identical to that in ARAUSv1. For the ARAUSv2 test set, we used a similar method as that for ARAUSv1, which involved calibration to pre-defined $L_{\mathrm{A,eq}}$ values, followed by principal component analysis (PCA) for dimensionality reduction on a selection of acoustic and psychoacoustic indicators, and finally a clustering and fold assignment via a self-

organizing map. Differences in the individual steps for the ARAUSv2 test set are as follows:

1. For the dimensionality reduction via PCA, the weights and number of components for the base soundscapes and each masker class from ARAUSv1 were directly applied to the ARAUSv2 test set (i.e., *without* refitting the PCA to the ARAUSv2 test set). This was to ensure that the base soundscapes and masker classes shared identical principal component spaces for both ARAUSv1 and ARAUSv2.
2. For the clustering and fold assignment, the number of folds was set to 2 for the ARAUSv2 test set, as opposed to the original 5 for the ARAUSv1 cross-validation set.

To validate that the abovementioned procedure resulted in similar distributions of base urban soundscapes and maskers in the ARAUSv2 test set as that in ARAUSv1, we conducted 2-sample Peacock's tests for each set of soundscapes and maskers in each class in the ARAUSv1 cross-validation set against the ARAUSv2 test set. The p-values of these tests were 0.2786, 0.6699, 0.5406, 0.2212, 0.5960, 0.6788 for the sets of base urban soundscapes, maskers in the "bird" class, maskers in the "construction" class, maskers in the "traffic" class, maskers in the "water" class, and maskers in the "wind" class. Since all p-values were above 0.05, we concluded that there is no significant difference between the base urban soundscapes and maskers in ARAUSv1 and ARAUSv2, at least based on the acoustic domain.

2.4. Generation of Stimuli

For both the ARAUSv1 extension and ARAUSv2 test set, the procedure to generate each 30-second audio-visual stimulus was identical to that for the ARAUSv1 cross-validation set. Namely, it was to:

1. Calibrate all base urban soundscape recordings to their in-situ $L_{\mathrm{A,eq}}$ levels measured at the time of recording.
2. Calibrate all maskers to fixed soundscaape-to-masker ratios (SMR) in the set $\{-6, -3, 0, +3, +6\}$ (i.e., an SMR of +6 would result in the masker being calibrated to an $L_{\mathrm{A,eq}}$ higher than that of the base urban soundscape by 6 dB).
3. Randomly choose a base urban soundscape and masker (including the possibility of silence) from the same fold, as well as a corresponding SMR.
4. Augment the calibrated base urban soundscape with the calibrated masker recording at the chosen SMR via element-wise addition in the time domain to obtain the audio, and crop the 0°-azimuth, 0°-elevation field of view of the 360°-video to obtain the video for the final stimulus.

However, the ARAUSv2 test set used the soundscapes in 1 as base urban soundscapes and the maskers in 2 instead, which formed a disjoint set with ARAUSv1.

2.5. Participant Recruitment

The participant recruitment procedure for ARAUSv2 was similar to that for ARAUSv1, and participants were also screened to have the same satisfactory hearing ability as that for ARAUSv1 (defined as a mean threshold of hearing at most 20 dB in both ears for participants aged under 30 and a mean threshold of hearing at most 30 dB for participants aged 30 and above). However, for the ARAUSv1 extension, only participants who indicated (prior to recruitment) that they were not a student were included. This was to mitigate the limitation of having responses from a largely young, non-student population in ARAUSv1. No such restriction was imposed for participants in the ARAUSv2 test set. This gave rise to the demographic breakdown of participants in ARAUSv1 and ARAUSv2 as shown in Tables 3 and 4, respectively. As of the time of writing, the ARAUSv2 dataset contains responses from 709 participants, with a projected increase by the end of data collection.

Table 3: Demographic breakdown of participants in ARAUSv1

	Original						Extension (at time of writing)				
Fold	**Test**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Sample size	5	120	120	120	120	120	15	15	15	15	15
# female	4	62	56	69	68	69	8	9	8	8	11
# male	1	58	64	51	52	51	7	6	7	7	4
Mean age	22.4	27.4	26.7	26.3	27.3	26.3	43.7	47.1	46.2	39.2	42.1
Std. dev. of age	5.5	10.2	9.8	10.4	11.0	8.7	12.1	12.2	10.1	11.2	9.7
Median age	21	24	24	23	23	24	38	49	48	37	43
Maximum age	32	63	71	68	65	60	66	69	61	62	60
Minimum age	18	19	18	18	18	18	29	27	31	25	29
% student	80.0	68.3	71.7	80.0	70.8	75.0	0	0	0	0	0
% non-student	20.0	31.7	28.3	20.0	29.2	25.0	100	100	100	100	100

Table 4: Demographic breakdown of participants in ARAUSv2 (at the time of writing)

	Cross-validation					Test		Overall
Fold	**1**	**2**	**3**	**4**	**5**	**6**	**7**	
Sample size	135	135	135	135	135	17	17	709
# female	70	65	77	76	80	10	10	388
# male	65	70	58	59	55	7	7	321
Mean age	29.2	28.9	28.5	28.6	28.1	30.1	30.0	28.7
Std. dev. of age	11.6	11.9	12.1	11.6	10.1	10.8	11.4	11.4
Minimum age	19	18	18	18	18	20	20	18
Median age	24	24	24	24	24	25	24	24
Maximum age	66	71	68	65	60	57	62	71
% student	60.7	63.7	71.1	62.9	66.7	41.1	64.7	64.4
% non-student	39.3	36.3	28.9	37.1	33.3	58.9	35.3	35.6
% non-student	60.7	63.7	71.1	62.9	66.7	41.1	64.7	64.4

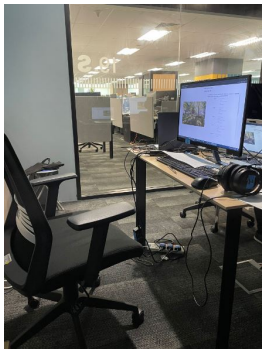

Figure 1: Photo taken at Meeting Room 19.S, located on the 19th floor of HDB Hub, Singapore.

2.6. Listening Conditions

For both the ARAUSv1 extension and ARAUSv2 test set, the listening conditions were identical to that for ARAUv1. However, due to logistical considerations, some participants for the ARAUSv1 extension listened to the stimuli at a quiet meeting room shown in Figure 1.

2.7. Questionnaires and Consistency Checks

The same affective response questionnaire (ARQ) and participant information questionnaire (PIQ) as ARAUSv1 was used for ARAUSv2, and the same consistency checks were performed for data quality before accepting responses obtained via the data collection procedure into ARAUSv2. Each participant evaluated 42 stimuli generated via the method in Section 2.4 given the following prompt:

> Imagine that you are standing at the location shown in the video, listening to the sound environment playing through the headphones.

As with ARAUSv1, the ARQ included the items concerning the perceived affective quality of the stimulus from part 2 of the Method A questionnaire in ISO/TS 12913-2:2018:

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

The responses to these items respectively gave us the values $r_{\text{pl}}, r_{\text{ev}}, r_{\text{ch}}, r_{\text{vi}}, r_{\text{ue}}, r_{\text{ca}}, r_{\text{an}}, r_{\text{mo}} \in \{1, 2, 3, 4, 5\}$, which were used to compute the *normalized* ISO Pleasantness $P \in [-1, 1]$ according to Equation 1 and ISO Eventfulness $E \in [-1, 1]$ according to Equation 2:

$$P = \frac{2(r_{\text{pl}} - r_{\text{an}}) + \sqrt{2}\,(r_{\text{ca}} - r_{\text{ch}} + r_{\text{vi}} - r_{\text{mo}})}{8\left(1 + \sqrt{2}\right)}, \tag{1}$$

$$E = \frac{2\,(r_{\text{ev}} - r_{\text{ue}}) + \sqrt{2}\,(r_{\text{ch}} - r_{\text{ca}} + r_{\text{vi}} - r_{\text{mo}})}{8\left(1 + \sqrt{2}\right)}, \tag{2}$$

Upon completion of the ARQ for all 42 stimuli, we administered the PIQ, which asked participants to provide the same set of basic demographic information and responses to standard psychological questionnaires as ARAUSv1.

3. MODELS

With the newly-collected responses for ARAUSv2, we attempted to design a preliminary set of multimodal models for the prediction of normalized ISO Pleasantness and ISO Eventfulness, making use of the both the acoustic and visual information in the audio-visual stimuli, as well as the PIQ responses in the dataset. The multimodal models fall into two classes: Linear models and attention-based deep neural networks (DNNs).

3.1. Linear models

For the linear models, we use the elastic net models with the acoustic and psychoacoustic parameters described in ARAUSv1 [7] as features corresponding to the acosutic domain, and combine them in a stepwise fashion with as generalized linear regression models with features corresponding to the participant-linked and visual domains. In addition, we explore the efficacy of ignoring the acoustic features altogether, using only the participant-linked and/or visual features to predict the normalized ISO Pleasantness and ISO Eventfulness. All linear models were trained using ARAUSv2 with the data collected as of the time of writing.

For the participant-linked features, we followed the coding given in Appendix B of [7] for ordered categorical variables (e.g., age, score on Weinstein's noise sensitivity scale [21]) but with appropriate normalization of the intervals to $[0, 1]$. Binary categorical variables (e.g., gender) were left as values in $\{0, 1\}$, whereas all other unordered categorical variables (e.g., occupational status, ethnicity) were re-coded as $n - 1$ binary dummy variables for each such variable, where $n > 2$ is the number of possible choices for such a variable [22]. For instance, the variable `occupation` was originally coded as {0: others, 1: student, 2: employed, 3: retired, 4: unemployed}, so we re-coded the variable as {0000: others, 0001: student, 0010: employed, 0100: retired, 1000: unemployed}.

For the features corresponding to the visual domain, we first read each frame of the video corresponding to each augmented soundscape (as generated using the process described in Section 2.4) as a still image with hue-saturation-value (HSV) channels. In this representation, the hue (H) channel can be considered as a color wheel taking angular values (in degrees) in the range $[0, 360)$, so we computed the fraction of each image with H values between 180 and 240 (inclusive), between 76 and 150 (inclusive), and at least 340 or at most 10, as the fraction of the image with relatively "blue", "green", and "red" hues, respectively. This was inspired from the observation that visual features, such as green infrastructure, could influence the affective perception of a given soundscape [23], and a sample of the segmented output is shown in Figure 2. The median value of the fractions with "blue", "green", and "red" hues across all frames in each base urban soundscape was taken as the final set of visual features used for the linear models.

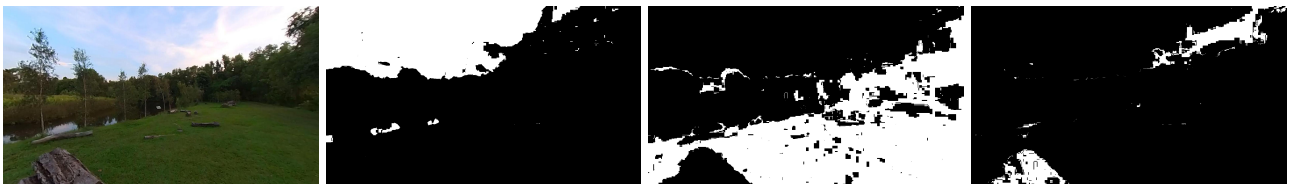

Figure 2: First frame of video corresponding to soundscape S0005 (extreme left) after segmentation in HSV color space to "blue" (center left), "green" (center right), and "red" (extreme right) hues.

3.2. Deep neural networks

For the attention-based DNNs, we use the models described in [18], with the log-mel spectrograms of the base urban soundscape and masker coupled with the log-gain of the masker as inputs corresponding to the acoustic domain, the coded PIQ responses described in Section 3.1 as features corresponding to the participant-linked domain, and a randomly selected still frame from the video as input corresponding to the visual domain. Early fusion of the different domains was used, where the embeddings computed from the input features were stacked as different channels to a convolutional "feature augmentation block" performing soundscape augmentation in the feature domain. Since the DNNs were configured to necessitate acoustic inputs, we trained them in four different configurations, using: acoustic-only inputs; acoustic and visual inputs; acoustic and participant-linked inputs; and acoustic, participant-linked, and visual inputs. These models were trained and validated on ARAUSv1, but tested using the ARAUSv2 test set as of the time of writing.

4. RESULTS AND DISCUSSION

Tables 5 and 6 respectively show the results of the models described in Section 3 when predicting the normalized ISO Pleasantness and ISO Eventfulness, as mean squared error (MSE) values across the predictions in their training, validation, and test sets. For reference, we also present the results of a "dummy" model that predicts the mean value of the ground-truth ISO Pleasantness (or ISO Eventfulness) labels in their training sets for all stimuli, which any other model making meaningful use of the input features should theoretically outperform.

Table 5: Mean squared errors (± standard deviation over 10 differently-seeded runs where applicable) of ISO Pleasantness models described in Section 3. Smaller values are desirable, as denoted by the down arrow (↓). For training and validation, the linear models used ARAUSv2 whereas the DNNs used ARAUSv1. However, for testing, all models used the ARAUSv2 test set. For the modalities, "A", "P", and "V" represent the use of acoustic, participant-linked, and visual features, respectively.

		Mean squared error (↓)		
Model type	Modalities	Train	Validation	Test
Dummy	—	0.1536	0.1537	0.1812
Elastic net (acoustic) + linear regression (non-acoustic)	A	0.1334	0.1339	0.1481
	A+V	0.1306	0.1320	0.1443
	A+P	0.1284	0.1315	0.1366
	A+P+V	0.1282	0.1321	0.1377
	P	0.1511	0.1538	0.1725
	V	0.1527	0.1544	0.1854
	P+V	0.1503	0.1545	0.1760
DNN (early fusion)	A	0.1079 ± 0.0029	0.1217 ± 0.0009	0.1384 ± 0.0021
	A+V	0.1089 ± 0.0019	0.1199 ± 0.0009	0.1426 ± 0.0051
	A+P	**0.1071** ± 0.0036	0.1204 ± 0.0010	**0.1342** ± 0.0018
	A+P+V	0.1082 ± 0.0031	**0.1194** ± 0.0012	0.1395 ± 0.0025

Table 6: Mean squared errors (± standard deviation over 10 differently-seeded runs where applicable) of ISO Pleasantness models described in Section 3. Smaller values are desirable, as denoted by the down arrow (↓). For training and validation, the linear models used ARAUSv2 whereas the DNNs used ARAUSv1. However, for testing, all models used the ARAUSv2 test set. For the modalities, "A", "P", and "V" represent the use of acoustic, participant-linked, and visual features, respectively.

		Mean squared error (↓)		
Model type	Modalities	Train	Validation	Test
Dummy	—	0.1574	0.1576	0.1501
Elastic net (acoustic) + linear regression (non-acoustic)	A	0.1320	0.1329	0.1672
	A+V	0.1271	0.1294	0.1818
	A+P	0.1275	0.1314	0.2013
	A+P+V	0.1252	0.1304	0.1882
	P	0.1554	0.1586	0.1555
	V	0.1518	0.1544	**0.1431**
	P+V	0.1499	0.1555	0.1480
DNN (early fusion)	A	0.1121 ± 0.0020	0.1224 ± 0.0009	0.2231 ± 0.0111
	A+V	**0.1120** ± 0.0025	0.1228 ± 0.0012	0.1755 ± 0.0139
	A+P	0.1121 ± 0.0031	0.1226 ± 0.0012	0.2343 ± 0.0159
	A+P+V	0.1123 ± 0.0022	**0.1223** ± 0.0009	0.1850 ± 0.0104

Most models for the prediction of normalized ISO Pleasantness performed better than the corresponding dummy model in the training, validation, and test sets. The only exceptions were in the cross-validation set for the linear models using only participant-linked features (MSE of 0.1538), only visual features (MSE of 0.1544), and only participant-linked and visual features (MSE of 0.1545), with the dummy model achieving an MSE of 0.1537, as well as in the test set for the linear model using only visual features (with an MSE of 0.1854 against an MSE of 0.1812 for the dummy model).

Notably, the linear models that did not utilize any acoustic information performed significantly worse than those that did, which indicates that information from the acoustic modality is still the most important for the prediction of normalized ISO Pleasantness. However, the participant-linked and visual features described in Section 3.1 were not completely useless as well, as evidenced by the decreased test set MSEs for the linear models using only the participant-linked features (0.1725), and only participant-linked and visual features (0.1760), against the dummy model (0.1812). The linear models also performed better When the participant-linked and visual features were coupled with information from the acoustic modality, with the MSE of the acoustic-only linear model decreasing from 0.1481 on the ARAUSv2 test to 0.1443 when visual features were added, 0.1366 when participant-linked features were added, and 0.1377 when both visual and participant-linked features were added. This indeed reflects the definition of a soundscape as an "acoustic environment as perceived... by a person or people, in context", since the participant-linked features provide relevant information related to listener perception, and the visual features provide relevant information related to the context of evaluation.

A similar phenomenon can be observed in the validation set performance in the DNN models for normalized ISO Pleasantness, with the models using information from the non-acoustic modalities performing (on average) better than those using only information from the acoustic modalities. This does not completely translate over to the ARAUSv2 test set, but the best-performing model among those investigated in this study there is the DNN using both acoustic and participant-linked information (MSE of 0.1342 ± 0.0018), which still indicates the usefulness of capturing multimodal information in the modeling of normalized ISO Pleasantness. This observation is further substantiated by the fact that best-performing models in the train, validation, and test sets are all multimodal DNN models as well. The use of raw images for the DNN models using information from the visual modality, however, could have impacted their relative performance in the test set against the corresponding acoustic-only models (MSE of 0.1426 ± 0.0051 for acoustic + visual, and MSE of 0.1395 ± 0.0025 for acoustic + participant-linked + visual, against MSE of 0.1384 ± 0.0021 for acoustic-only) due to the relatively small dataset of images in ARAUSv2 as opposed to standard datasets (e.g., ImageNet) used in the field of computer vision. This could conceivably be mitigated by using pre-trained models on such large datasets in a transfer learning paradigm, where the weights are fine-tuned to the frames in the videos in the ARAUSv2 stimuli, and is a possible avenue for further exploration.

Conversely, most models for the prediction of normalized ISO Eventfulness performed better than the corresponding dummy model only in the training and validation sets but not the test set. Only the linear models using only visual features (MSE of 0.1431) and both participant-linked and visual features (MSE of 0.1480) outperformed the dummy model (MSE of 0.1501) in the ARAUSv2 test set. This hints at the possibility of a significant domain mismatch in the case of normalized ISO Eventfulness between the train/validation sets and the test sets proposed in ARAUSv2, despite the *a priori* Peacock's tests in Section 2.3 indicating otherwise. This could possibly be mitigated by changing the order of the train/validation/test folds for the purpose of crafting ISO Eventfulness models, but is outside the scope of the current study. Nonetheless, based on the results observed on the validation set, we can observe that there exist only minor differences in model performance between the acoustic-only linear models for normalized ISO Eventfulness (MSE of 0.1329) and those also using information from other modalities (MSE of 0.1294 with visual features, 0.1314 with participant-linked features, and 0.1304 with both visual and participant-linked features). Similarly, there are

no significant differences in the performance of the DNN models for normalized ISO Eventfulness, regardless of which modalities were used as input. Hence, this suggests a possible redundancy in the information provided from the participant-linked and visual modalities with respecto the information available in the acoustic modality, at least in the features and input methods chosen for this study. This may also hint at the possibility of the normalized ISO Eventfulness being more of a context-independent characteristic of the soundscape, since additional contextual information from the visual environment and the listener did not appear to improve the performance of the normalized ISO Eventfulness models.

5. CONCLUSION

In conclusion, we described a set of recent updates to the ARAUS dataset that consists of additional responses collected using similar procedures, but using audio-visual stimuli that are mutually disjoint from those in the pre-existing version. While data collection is still in progress, but analysis and modeling using the newly-collected responses as of the time of writing shows promise in allowing the updated dataset (ARAUSv2) to mitigate the original limitations concerning a small test set and a disproportionate representation of young university students.

Future work involving ARAUSv2 could include the investigation of how best to use it to develop generalizable models for the prediction of (normalized) ISO Eventfulness, possibly by having a different train/validation/test fold breakdown than that for ISO Pleasantness. The dataset could also be further expanded using synthetically-generated but realistic maskers, such as those belong to the wind and water class, by a technique involving the interpolation of control vectors in a latent space as described in [24]. This would allow for a larger, infinite variety of maskers that could be used as part of the audio-visual stimuli in the dataset, instead of the present technique of picking and excerpting tracks from the Freesound and Xeno-Canto databases.

Data and replication code using ARAUSv2 will be made available as an update to the existing ARAUSv1 repositories at https://doi.org/10.21979/N9/9OTEVX and https://github.com/ntudsp/araus-dataset-baseline-models when data collection is completed.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation, Singapore, and Ministry of National Development, Singapore under the Cities of Tomorrow R&D Program (CoT Award: COT-V4-2020-1). 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.

REFERENCES

1. International Organization for Standardization, *ISO 12913-2 Acoustics - Soundscape - Part 2: Data collection and reporting requirements*. Geneva, Switzerland: International Organization for Standardization, 2018.
2. M. Ciufo and D. Thomas, "EigenScape: A Database of Spatial Acoustic Scene Recordings," *Appl. Sci.*, vol. 7, 2017.
3. B. De Coensel, K. Sun, and D. Botteldooren, "Urban Soundscapes of the World: Selection and reproduction of urban acoustic environments with soundscape in mind," in *Proc. Inter-Noise*, 2017.
4. K. Ooi, *et al.*, "Lion City Soundscapes (LCS): Recording a Dataset of Characteristic Soundcapes of Singapore," pp. 1–8, 2023. [Online]. Available: https://doi.org/10.21979/N9/AVHSBX
5. International Organization for Standardization, *ISO 12913-3:2019 - Acoustics - Soundscape - Part 3: Data analysis*. Geneva, Switzerland: International Organization for Standardization,

2019.

6. A. Mitchell, *et al.*, "The International Soundscape Database: An integrated multimedia database of urban soundscape surveys – questionnaires with acoustical and contextual information," 2021. [Online]. Available: https://doi.org/10.5281/Zenodo.5914762
7. K. Ooi, *et al.*, "ARAUS: A Large-Scale Dataset and Baseline Models of Affective Responses to Augmented Urban Soundscapes," *IEEE Transactions Affect. Comput.*, pp. 1–25, 2023.
8. S. Versümer, J. Steffens, and F. Rosenthal, "Extensive crowdsourced dataset of in-situ evaluated binaural soundscapes of private dwellings containing subjective sound-related and situational ratings along with person factors to study time-varying influences on sound perception – research data," University of Applied Sciences Düsseldorf, Germany Institute of Sound and Vibration Engineering, Duesseldorf, Germany, Tech. Rep. April, 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7858848
9. International Organization for Standardization, *ISO 12913-1:2014 - Acoustics - Soundscape - Part 1: Definition and conceptual framework*. Geneva, Switzerland: International Organization for Standardization, 2014.
10. J. Y. Hong and J. Y. Jeon, "Designing sound and visual components for enhancement of urban soundscapes," *The J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2026–2036, 2013.
11. Y. Hasegawa and S. K. Lau, "Audiovisual bimodal and interactive effects for soundscape design of the indoor environments: A systematic review," *MDPI Sustain.*, vol. 13, no. 1, pp. 1–30, 2021.
12. R. Yue, *et al.*, "A visualized soundscape prediction model for design processes in urban parks," *Build. Simul.*, vol. 16, no. 3, pp. 337–356, 2023.
13. J. Wang, *et al.*, "What Constitutes the High-Quality Soundscape in Human Habitats? Utilizing a Random Forest Model to Explore Soundscape and Its Geospatial Factors Behind," *Int. J. Environ. Res. Public Heal.*, vol. 19, no. 21, 2022.
14. L. A. Ferguson, *et al.*, "How much noise is too much? Methods for identifying thresholds for soundscape quality and ecosystem services," *Appl. Acoust.*, vol. 209, p. 109388, 2023.
15. S. Durbridge and D. T. Murphy, "Assessment of soundscapes using self-report and physiological measures," *Acta Acustica*, vol. 7, no. 5, pp. 12–25, 2023.
16. A. Mitchell, *et al.*, "Effects of Soundscape Complexity on Urban Noise Annoyance Ratings: A Large-Scale Online Listening Experiment," *Int. J. Environ. Res. Public Heal.*, vol. 19, no. 22, 2022.
17. J. Wang, *et al.*, "Deep Learning-Based Road Traffic Noise Annoyance Assessment," *Int. J. Environ. Res. Public Heal.*, vol. 20, no. 6, p. 5199, 2023.
18. K. Ooi, *et al.*, "Autonomous Soundscape Augmentation With Multimodal Fusion Of Visual And Participant-linked Inputs," in *IEEE ICASSP 2023*, 2023.
19. F. Font, G. Roma, and X. Serra, "Freesound Technical Demo," in *Proc. 2013 ACM Multimed. Conf.*, 2013, pp. 411–412.
20. B. Planqué and W.-P. Vellinga, "Xeno-canto: a 21st century way to appreciate Neotropical bird song," *Neotropical Bird.*, vol. 3, no. January, pp. 17–23, 2008.
21. N. D. Weinstein, "Individual differences in reactions to noise: A longitudinal study in a college dormitory," *J. Appl. Psychol.*, vol. 63, no. 4, pp. 458–466, 1978.
22. W. Mendenhall and T. Sincich, "Multiple Regression Models," in *A Second. Course Stat. Regres. Analysis*, 7th ed. Pearson Education, Inc., 1966.
23. F. Stevens, D. T. Murphy, and S. L. Smith, "Soundscape auralisation and visualisation: A cross-modal approach to soundscape evaluation," in *DAFx-18*, 2018, pp. 133–140.
24. C. Gupta, *et al.*, "Towards Controllable Audio Texture Morphing," in *IEEE ICASSP 2023*, 2023.