The Alan Turing Institute

Data Study Group Final Report:

IEDE Acoustics Group, University College London

Deep Learning Techniques for noise annoyance detection (DeLTA)

12-16 December 2022

DOI: 10.5281/zenodo.10090651

Contents

1 Executive Overview

1.1 Challenge overview

Noise annoyance is often reported as one of the main adverse effects of noise exposure on human health. Chronic high noise annoyance impacts 22 million people in Europe alone, with a broad range of public health outcomes. This Data Study Group applied sound source identification and deep learning methods on a set of urban recordings to create a model which can predict the resulting annoyance rating. The research challenge was to investigate to what degree the inclusion of sound source information can inform the optimal modelling strategy for automatically predicting noise annoyance.

1.2 Data overview

The challenge made use of the "Deep Learning Techniques for noise Annoyance detection" (DeLTA) dataset collected by the Acoustics Group at University College London (UCL). The DeLTA dataset comprises 2,980 15-second binaural audio recordings collected in urban public spaces across four cities in Europe (London, Venice, Granada, and Groningen). A remote listening experiment was distributed to a pool of pre-registered participants ($N = 1,221$). During the listening experiment, participants listened to ten recordings from the sample and were instructed to select all the sound sources they could identify within the recording (from a set of 24 possible sound sources) and then to provide an annoyance rating (from 1 to 10).

1.3 Main objectives

A growing challenge for the field of urban noise and soundscapes has been to provide a more nuanced view of the prevalence of annoying and impactful sounds across cities. To capture the effects requires the identification of different sound sources in a complex environment and the ability to automatically characterise the likelihood of annoyance. This DSG was proposed in order to extend the use of deep learning models for sound source identification to the prediction of perceptual ratings of urban

soundscapes. The challenge group applied sound source identification and deep learning methods to the recordings in the DeLTA dataset to create a series of models which can predict the resulting annoyance rating. Two key objectives were put forward:

- **To test various deep learning model structures for predicting noise annoyance.**
- **To investigate whether the inclusion of sound source information, whether from human-generated labels or from automated sound source recognition improves the ability to predict noise annoyance ratings.**

The approaches taken in this DSG and the novel developments of deep learning models to noise annoyance can help spur the creation of more advanced tools within the field.

1.4 Approach

After an in-depth exploration of the dataset which expands on the previous literature, this work was divided into two primary experiments. The first experiment takes a classical machine learning (ML) approach to predicting annoyance ratings based on the sound source labels present in the DeLTA dataset. In total, four models were built: two linear regression models and two random forest models. The primary question answered by these models was to what extent (weakly-labelled) sound sound source labels are able to predict annoyance ratings and how this prediction is improved by incorporating spectral information. These models will also act as a baseline for performance to compare the later deep learning models against.

For the second experiment, a series of 12 neural networks were trained with mel spectrograms as the primary input. Mel spectrograms provide a visual representation of the frequency and temporal information in a recording, in a way which approximates the human auditory response. Across the models, we tested a variety of architectures (including Convolutional Neural Networks (CNN), TinyCNN, Temporal Convolutional Networks (TCN), Feed forward Neural Networks (FNN), and Long Short-Term Memory (LSTM)), different spectrogram resolutions, and

different approaches for incorporating the sound source information. By investigating a wide array of modelling strategies, interesting and novel patterns can be found in what approaches do or do not work well.

1.5 Main Conclusions

Our primary conclusions of the modelling experiments are that:

1. Using higher resolution spectrograms outperforms compressed spectrograms with an approximate RMSE improvement of 16.5%.
2. In general, simpler model structures performed as well or better on this dataset than more complex models, with the TinyCNN models performing surprisingly well despite their simplicity. A TinyCNN model achieved an equivalent RMSE (1.09) to the 'best' performing model which incorporates a large-scale pretrained audio neural network (PANN) (RMSE = 1.08).
3. Including sound source information does improve the prediction accuracy, however this is often not enough to overcome the two previous factors (high-resolution spectrograms and simple model structures), demonstrating the importance of the model structure chosen.
4. When comparing like models, it appears that including the sound source information as an output target can match the performance of more complex models which make use of large pretrained audio networks.
5. The added complexity necessary to incorporate sound source labels as input features appears to drastically reduce the predictive performance, making it difficult to directly compare implicit vs explicit methods of incorporating human-generated sound source information.

1.6 Limitations

There are several limitations to the dataset. Firstly, the annoyance ratings are unbalanced (i.e. there are many more low annoyance ratings

compared to high annoyance ratings). This means that the models may under-predict high annoyance ratings. Secondly, as the initial data collection was performed in an online study, the absolute playback volume of the recordings to the participants could not be controlled. Therefore, sound level could not be used as a meaningful predictive factor. This likely limits the performance of any models, given that noise annoyance is directly related to sound level. For this reason, our analysis focused primarily on incorporating spectral characteristics.

1.7 Recommendations and further work

Coming into this challenge, there was an expectation that incorporating sound source labels as inputs into the network would lead to better predictions [Orga et al., 2021]. However, our work demonstrated that this explicit approach to incorporating sound source information added unnecessary complexity and reduced the model performance. Building on the success of jointly predicting sound source labels and annoyance ratings, more sophisticated models could be developed using our approach as a starting point. One fruitful avenue for further research could be to explore the possibility that this result also works in the inverse, i.e. including perceptual features as a joint output could improve the performance of sound source classification tasks. The simplicity and size of the TinyCNN models in particular are a promising indication that these models could be deployed on even low-power equipment, possibly even on remote monitoring sensors in a smart city context.

Although the recordings in the dataset are two-channel, we only made use of one channel in our modelling. Future work could explore the possible benefits of considering the full binaural signal. This would provide additional information including the inter-aural time and level differences. Whilst data augmentation is a common process for image-based neural networks, this was not explored for our models. This could also be beneficial for future model development, particularly if this augmentation makes further use of both channels.

2 Problem formulation

Urban soundscapes are complex environments, with overlapping sound sources each competing for our attention against an ever-shifting background. Although discussions and investigations of “urban noise” often focus only on traffic and aircraft noise, all sounds in a city contribute to the production or restoration of stress. Chronic high noise annoyance impacts 22 million people in Europe alone, with a broad range of public health outcomes, ranging from mild distress to severe and chronic physical impairment and leading to increased risks of cardiovascular and metabolic disorders [Guski et al., 2017, Śliwińska-Kowalska and Zaborowski, 2017].

These impacts have been documented using large scale maps of modelled noise levels, but noise annoyance is caused by much more than solely elevated noise levels [Yang and Kang, 2005, Zwicker and Fastl, 2007]. Some sounds are positively and some negatively perceived, and this influences annoyance to different extents. Capturing these effects requires the identification of different sound sources in a complex environment [Orga et al., 2021]. A growing challenge is in providing a more nuanced view of urban soundscapes with overlapping sound sources. Typically, noise annoyance in urban settings is estimated based on long-term measured sound levels. Recently, the focus has shifted to considering a more holistic view of the soundscape, in particular considering how different sound sources are each perceived and how they might contribute to a general degree of noise annoyance. In addition, noise level measurements from stationary sensors often do not reflect what individuals on the ground would actually be exposed to, as opposed to binaural soundscape recordings which are designed to accurately characterise the experience of urban space-users [ISO/TS 12913-2:2018, 2018, Aletta et al., 2020]. A binaural recording is an audio recording technique that employs two microphones, one for each ear, to capture sound in a way that replicates the natural cues our ears use for three-dimensional sound perception. When played back with headphones, it creates an immersive three-dimensional listening experience, replicating the sensation of being present in the original acoustic environment.

As a starting point, we consider a large-scale online active listening experiment conducted as part of the initial DeLTA project and a study by Mitchell et al. [2022a] investigating how the complexity of the soundscape - in terms of presence, number, and combination of different sound sources - would affect the perceived noise annoyance. This study found that the combination of sound sources in soundscape recordings is less important than the "soundscape complexity", while a combination of any two clearly distinguishable sound sources in a given urban soundscape appears to minimise the perceived noise annoyance, which is higher when the number of sources either increases or decreases.

This DSG was proposed in order to extend the use of deep learning models developed for sound source identification to the prediction of perceptual ratings of urban soundscapes. Two key objectives were put forward:

- **To test various deep learning model structures for predicting noise annoyance.**
- **To investigate whether the inclusion of sound source information, whether from human-generated labels or from automated sound source recognition, improves the ability to predict noise annoyance ratings.**

This report will be structured in three main sections. The first section presents an in-depth exploration of the DeLTA dataset, preparing for the model training to come later. It expands upon the soundscape complexity analysis presented in Mitchell et al. [2022a] and introduces additional context and findings. Finally, this first section presents the distributions of human responses for sound event labelling and discusses the challenges and conceptual considerations for considering ambiguity in human labelling of training datasets.

The second and third sections present the classical and deep learning predictive models trained during the DSG Challenge week, respectively. A wide range of models, each building upon one another, are trained to investigate how the model structure and methods for incorporating sound source information affect their predictive performance. We approach the given objectives by exploring three sets of annoyance rating prediction models:

1. A classical approach, to establish a baseline performance for predicting annoyance ratings. (Models **W1** - **W4**)
2. Neural Network (NN) Models which do not consider any sound source information to predict annoyance. (**C1** - **C4** , **Y0**)
3. NN Models which consider sound source information derived from:
   (a) Human labels (**R1** , **R2**)
   (b) Pretrained ID models (**Y2** - **Y4**)
   (c) Non-pretrained ID models. (**Y1** , **Y6**)

Each model variation is given a label[1] to identify it. Prior to model training, the dataset was divided into a training set (80%) of the sample and a testing set (20%). For the NN models, the training set was further divided into five folds for cross-validation.

We begin with a small set of classical ML models (linear regression and random forest) to establish a baseline against which the performance of the deep learning models can be compared. These models expand upon the results found in Mitchell et al. [2022a], which highlighted the impact of the complexity of the soundscape (i.e. the number of overlapping identifiable sounds) on the annoyance rating.

For model set two (**C1** - **C4** , **Y0**), a variety of neural network model architectures are explored, with mel spectrograms (see Section 5 for more information) as the input. These models do not include any implicit or explicit information regarding the sound source labels and attempt to predict annoyance ratings based solely on the sonic characteristics of the recordings.

Model set three (**R1** , **R2** , **Y1** - **Y6**) then builds upon set two by introducing the sound source label information, derived either from the human-generated labels in the DeLTA dataset or by incorporating sound source prediction from pretrained models. Several strategies for incorporating this information are trialled, including using embeddings from pretrained sound classification models, explicitly passing the sound

[1]These labels were created throughout the week based on the participant who was primarily in charge of building and training the model. For instance, Wingyan Yip was responsible for the classical models, which are labelled **W1** through **W4** .

event labels as input, or implicitly including the information by jointly predicting the annoyance rating and event labels.

3 Data overview

One issue with applying a deep learning approach previously has been the lack of large-scale training datasets which include high-quality urban recordings, sound source labels and, most importantly, perceptual data (i.e. annoyance ratings). The challenge makes use of the "Deep Learning Techniques for noise Annoyance detection" (DeLTA) dataset collected by the Acoustics Group at University College London (UCL). This dataset has been made publicly available under a Creative Commons 4.0 open data license on Zenodo [Mitchell et al., 2022b].

The DeLTA dataset comprises 2,980 15-second binaural audio recordings collected in urban public spaces across London, Venice, Granada, and Groningen. A remote listening experiment was distributed to a pool of pre-registered participants ($N = 1,221$). During the listening experiment, participants listened to ten 15-second-long binaural recordings of urban environments and were instructed to select all the sound sources they could identify within the recording and then to provide an annoyance rating (from 1 to 10). For the source recognition task, participants were provided with a list of 24 labels they could select from. These labels included: Aircraft, Bells, Bird tweets, Bus, Car, Children, Construction, Dog bark, Footsteps, General traffic, Horn, Laughter, Motorcycle, Music, Non-identifiable, Other, Rail, Rustling leaves, Screeching brakes, Shouting, Siren, Speech, Ventilation, and Water. Each recording was assessed by between two and four participants and on average, each recording has 3.1 identified sound sources present.

In total this dataset includes 12,210 individual ratings (from 1,221 participants) of 2,980 recordings with up to 24 source labels and one annoyance rating. The recordings were collected over numerous sessions as part of the Soundscape Indices (SSID) Project Kang et al. [2019] throughout 2019 and 2020[2], while the online listening experiment

[2]These original recordings form the International Soundscape Database (ISD) which is a publicly available dataset of binaural soundscape recordings and assessments,

was conducted between July 5th and July 23rd, 2021. The recordings were collected in urban public spaces across London, Venice, Granada, and Groningen, according to the SSID Protocol [Mitchell et al., 2020].

3.1 Data exploration and visualisation

Unbalanced dataset Mean annoyance, our main target variable, is calculated by averaging annoyance ratings from participants who rated the same sound track. As seen in Figure 1, the distribution of mean annoyance is right-skewed. This reflects an unbalanced dataset with a dominance of low annoyance ratings.

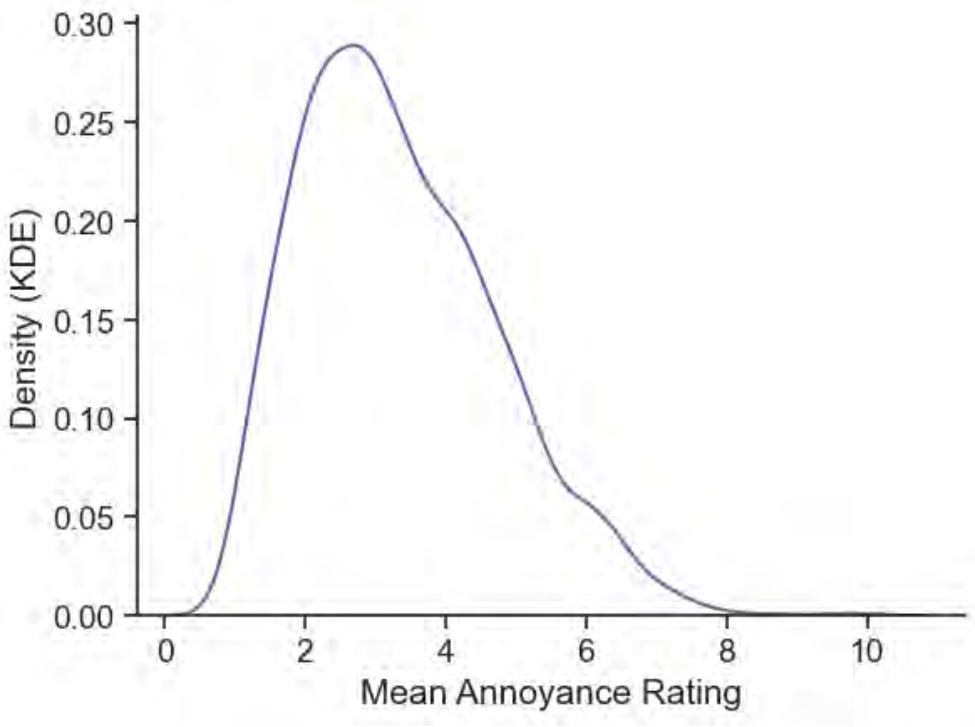


Figure 1: Density of mean annoyance ratings across the dataset. Mean annoyance is right-skewed, with mode at around 2.5.

When ratings are stratified into high and low annoyance levels with any

available on Zenodo [Mitchell et al., 2021a].

rating greater than five classified as high, we see higher variance among high ratings (Figure 2). This suggests high average ratings are likely influenced by individuals who rated highly. Highly-rated soundtracks only make up around 11% of the dataset, making the prediction of high annoyance ratings challenging.

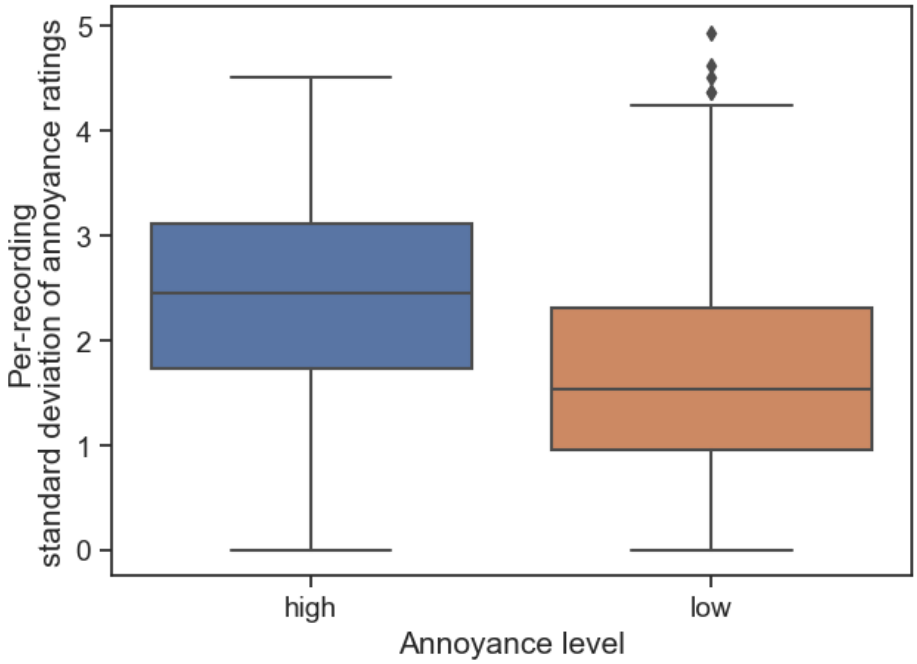


Figure 2: Variance in annoyance rating for recordings which have a high or low average annoyance rating. Highly-rated tracks have a higher variance, suggesting a high mean rating might be a result of extreme individual ratings.

Relationship between number of sounds and annoyance rating Mitchell et al. [2022a] explored the influence of the number of sound sources on annoyance rating and found that mean annoyance is minimised at $N = 2$ and increases towards $N = 8$, as plotted below in Figure 3.

We further explored this relationship by testing whether this non-linear relationship held across different types of sound sources. We added four

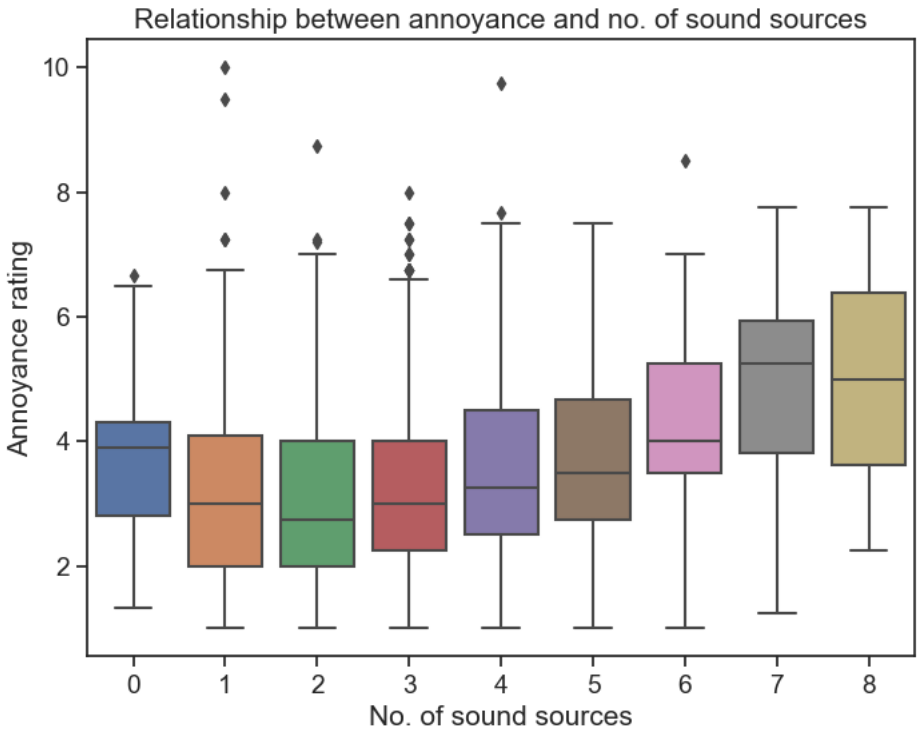


Figure 3: Distribution of annoyance ratings given the total number of sound sources in a given recording. Total number of sound sources has a non-linear relationship with mean annoyance

labels and plotted similar diagrams to see whether the relationship between sounds identified and mean annoyance changes for different kinds of sounds. The four labels are traffic, other urban (non-traffic), urban speech, and natural sounds. The original 24 source labels were sorted into one of these four categories.

A one-way ANOVA is conducted for each of the four labels because it is an appropriate significance test when comparing the annoyance ratings of multiple categories simultaneously. This test allows for the efficient assessment of whether there are statistically significant differences in annoyance ratings among the different labels. The results, reported in Table 1, indicate that the annoyance ratings for all label categories were found to be statistically significantly different, with $p < 0.01$ for all four labels.

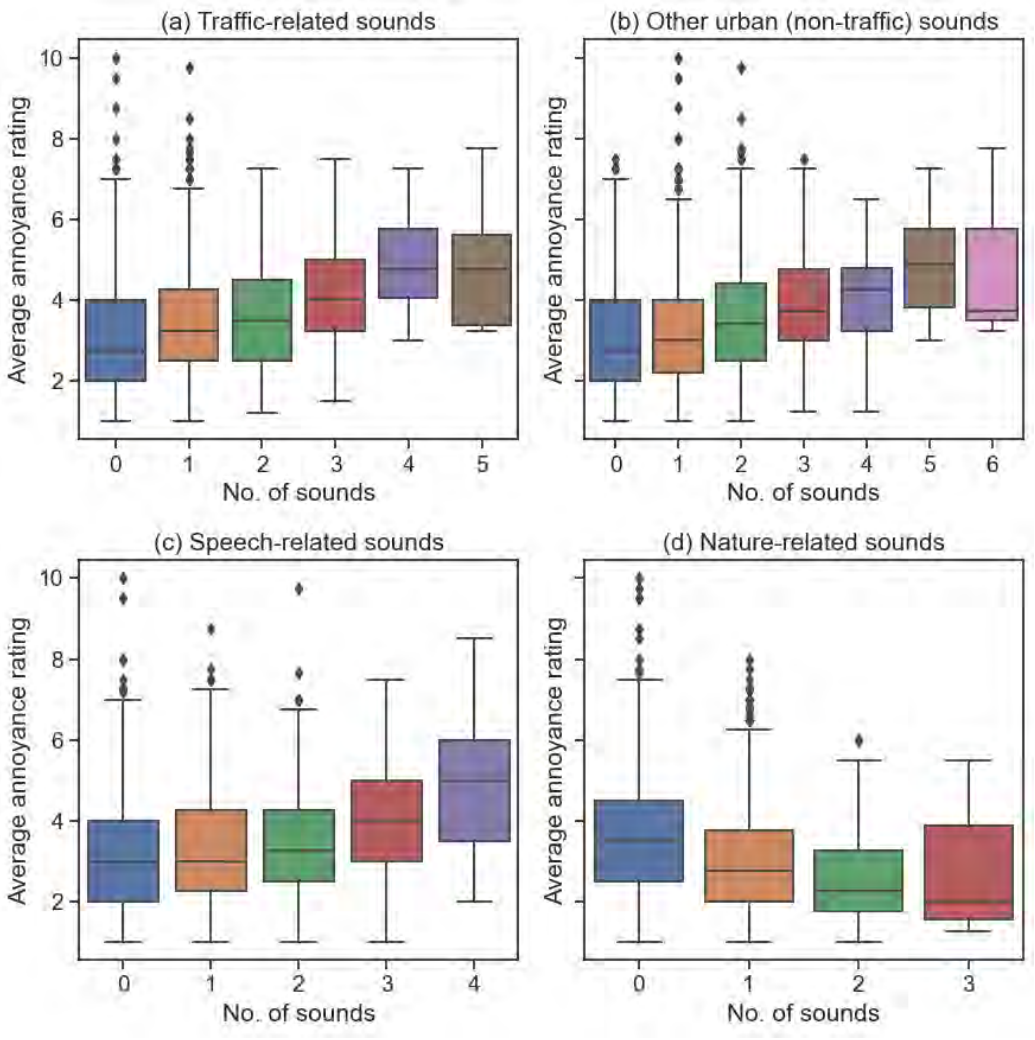


Figure 4: Distribution of annoyance rating by types of sounds. Nature-related sounds have a negative correlation with annoyance perception

As expected, more traffic sounds and other urban-related sounds like construction and footsteps correlate with higher mean annoyance. A similar pattern is found for speech-related sounds. Nature-related sounds, on the other hand, are negatively correlated with the annoyance rating. Although only four nature-related sounds appeared in the dataset, such that the negative association might just reflect the initial dip in Figure 3, the absence of a dip in the other three categories suggests it is

Table 1: Anova results for different types of sounds.

Sound label	df1	df2	F-stat	PR (F)
Traffic	5	2884	33.68	1.58e-33
Other-urban	6	2883	28.95	8.15e-34
Speech	4	2885	25.59	7.31e-21
Nature	3	2886	58.41	1.19e-36

plausible to expect a negative relationship to persist were there more nature-related sounds in the survey. In other words, a listener might find a soundscape less annoying when more nature sounds are present.

These results provide new context for the previous results on soundscape complexity reported by Mitchell et al. [2022a]. Two interpretations are possible:

1. That the previously identified relationship is indeed correct and the annoyance rating is minimised at $N = 2$ sound sources. In addition, that soundscapes with primarily nature-related sounds also tend to have only two identified sources. This would indicate that nature-dominated soundscapes also achieve a desirable level of soundscape complexity and therefore minimise annoyance through both the presence of semantically pleasant sounds and an appropriate amount of complexity.
2. That this analysis of the dataset is limited by the number of nature-related (and other semantically positive) sound sources available for participants to select. Given that the maximum number of nature-related sources is three, it is not possible to investigate the effect of high soundscape complexity with $N > 3$ semantically positive sources. This would be necessary in order to confirm or reject the hypothesis tested in Mitchell et al. [2022a].

3.2 Investigating ambiguities in source classification

Before supervising the sound source separation issue, it is worth analysing the distribution of human response for the sound events in the

DeLTA database, which will display the prevalence of certain sounds in the urban soundscape whilst also giving the degree of ambiguity in each sound label. When discussing supervised training datasets, we tend to consider the training data labels as the 'ground truth', or the true answer we are asking our model to predict [Stuart Russell, 2021]. However, this idea of a ground truth should be given extra attention when dealing with subjective data [Ellis et al., 2002]. For the DeLTA database, both the sound event labels and the annoyance ratings could be considered 'subjective' information.

For this challenge, one of our goals is to be able to predict whether a sound source is present in a soundscape or not. Ideally, the ground truth for such a task is objective knowledge for whether or not a particular source physically generated the sound. However, with the data available, this information is strictly impossible to know. For each recording in the dataset, several respondents have indicated whether they could hear certain sound sources in the soundscape. Even if all respondents agreed and indicated that exactly the same sources were present, we still cannot know objectively whether those sources actually generated the sound. In addition to this inherent subjectivity, in reality this form of data collection introduces another ambiguity when not all respondents agree on the identified sound sources. How should we treat the data when two respondents disagree about whether a sound source was actually present in the recording?

In Mitchell et al. [2022a], they addressed this question by taking a majority vote approach - if a majority of respondents for a given recording indicated that a source was present, then it is considered present. But this still raises the question - How much agreement is there about the presence of sources and which sources have the most ambiguity? This could also become a fine-tuning training objective for the sound classification model[3], such that when the sound source label prediction for annoyance ratings are propagated, it is in accordance with human perception. The goal would be to tune the predicted probabilities for each sound source to match the real-world percentage of agreement about the presence of a source.

We calculate the probability of the occurrence for each sound event in

[3]Note, unfortunately our experiments with the sound classification model did not progress enough to pursue this so this idea is speculation at this point.

participants' response. This is effectively the degree of agreement for whether each sound label is present in a given recording. We then examine the average agreement given across all recordings which (may) contain that sound. The function is as follows:

$$P(C_s) = \frac{1}{N_s} \sum_i \frac{n_{si}}{n_i}$$

where N_s is the number of recordings with s sound source identified by at least one participant, n_i is the total number of participants who assessed recording i, and n_{si} is the number of participants who identified sound source s in recording i. Therefore a $P(C_s)$ of 0.25 would indicate that 25% of respondents indicated that the sound label was present in the recording.

The result in Figure 5 shows a steady increase in agreement across the sound sources, ranging from around 50% for broadband sounds such as rustling leaves and ventilation to around 80% for bells, bird tweets, and music. Compared to the other sound events, speech is the most salient in the complex environment with an average agreement across recordings of 87%.

We follow the categorisation of the sound events mentioned in Section 2.2 and plot the probability as shown in Figure 6 for each of the categories. One thing worth noticing is that we separate the speech from other human voice, so there are five categories we use in the current analysis.

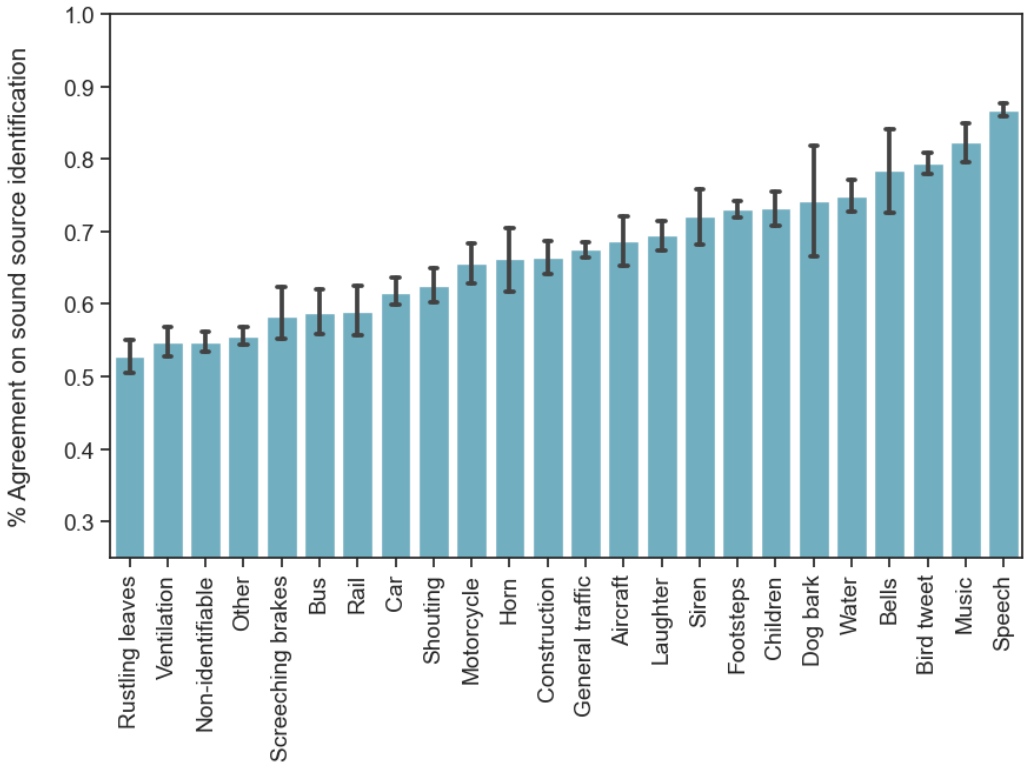


Figure 5: The degree of agreement among human labellers for each sound file in DeLTA dataset.

The result presented in Figure 6 confirms the preceding results that the speech is the most salient with the least ambiguity regarding its presence. By contrast, the traffic noise seems to be the sound source with the least agreement in participants' perception. The labelling of other sound sources such as other human voice, other urban sound and nature seem to be more ambiguous as indicated by a wider distribution of the response.

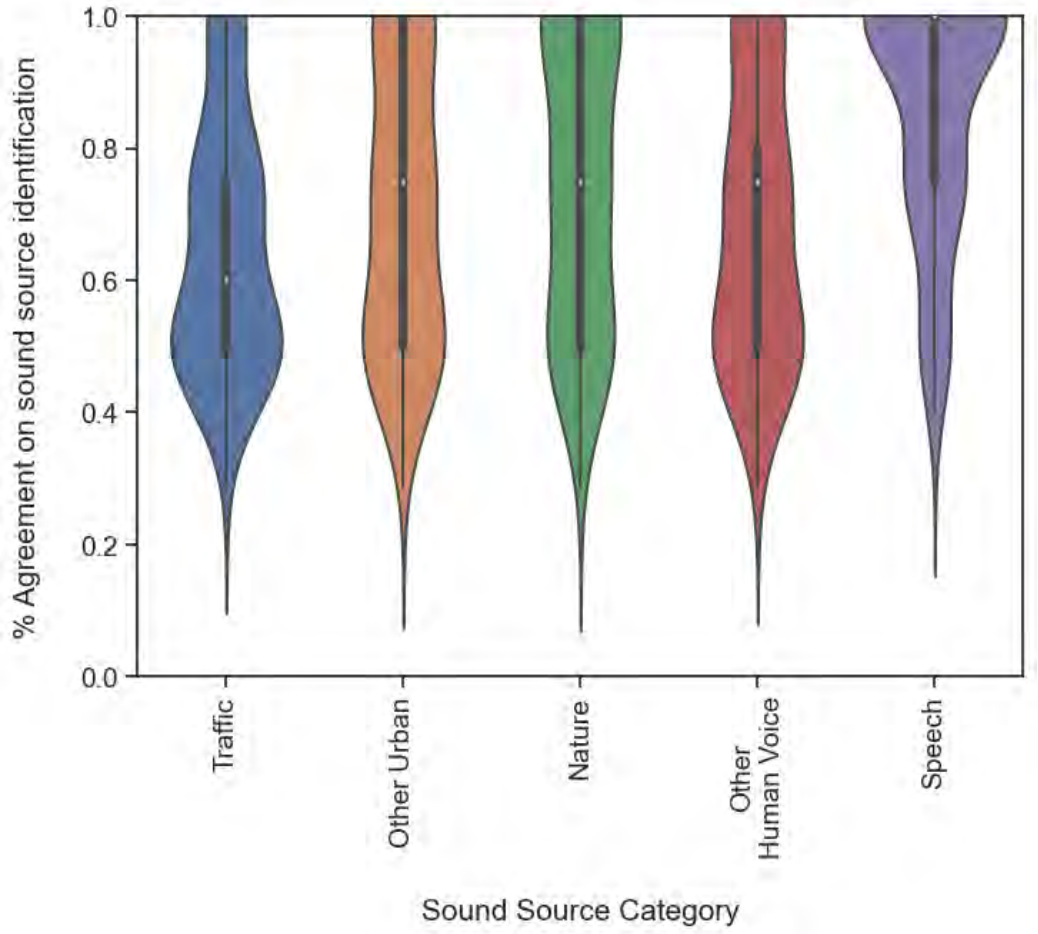


Figure 6: The distribution of human response for the the five sound source types.

The result so far indicates a high ambiguity issue of sound sources labelling in human perception. We propose that this ambiguity can be decomposed into two aspects of variation:

1. Salience: the sound source was present, but either noticed or not noticed.
2. Cross-over: The same sound source was heard between participants, but has been identified as having a different label(s).

Accounting for this ambiguity was developed as a target goal as we developed our models, however this avenue of investigation could not be explored in the experiments. This ambiguity in human labelling should be kept in mind when interpreting our results and when considering other sound source identification models.

4 Experiment 1: Establishing a baseline with classical ML models

We begin by training a set of classical ML models which can provide a baseline level of performance for the deep learning models in Experiment 2 to be compared against. These models expand on the results found in Mitchell et al. [2022a] and the new analysis presented in Section 3.1 to predict the annoyance rating from the sound source identification data in the DeLTA surveys. Initially these models include only the source IDs, then a spectral feature derived from the audio files is introduced.

Figure 7 presents an overview of the classical models which were tested.

4.1 Linear Regression Models - W1 and W2

Since the results in Section 3.1 indicated linear relationships between the number of different sound sources within each source category and mean annoyance, we begin with an Ordinary Least Squares (OLS) regression as a baseline model.

We created a multivariate linear regression model to predict the mean annoyance for each recording *i*, given as $Mean\_annoyance$. For each category of sources as defined in Section 3.1, the number of sound sources within that category identified in the recording is given by the $cnt\_$ variables. To add additional sonic information, two spectral features are also included. The recording is split into one second chunks and the spectral centroid at each second is calculated. The spectral centroid is an estimate of the 'centre of gravity' of the spectrum, in this case taken across the whole spectral range, and has been previously used in applications such as speech processing [Le et al., 2011] and music

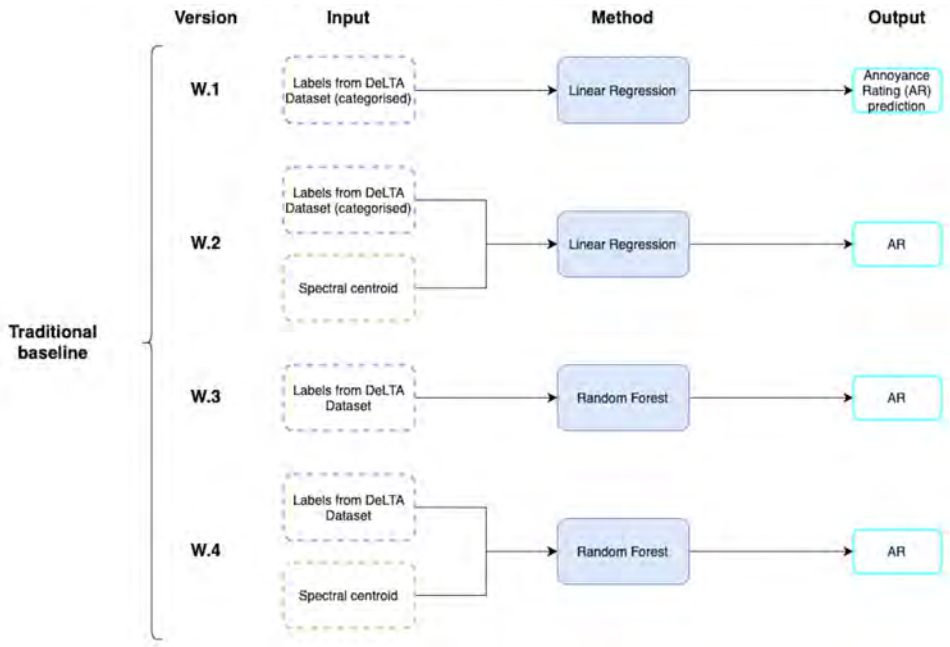


Figure 7: Classical models

perception [Schubert et al., 2004]. This time series of spectral centroids results in two features - a mean spectral centroid ($mean\_spec\_cent$) and standard deviation of the spectral centroid (std_spec_cent) for each recording *i*. Spectral centroid is often associated with the ”brightness” of a sound as it describes the balance of high and low frequencies, hence mean and standard deviation gives an idea of the general ’brightness’ and disparity in ’brightness’ in the sound track. We estimated two models, one without spectral characteristics (**W1**) and another with spectral characteristics included (**W2**).

We then define a multivariate linear regression model including intercept (α) and error terms (ϵ):

Table 2: Summary of results from Model **W.1**

	Coef	**Std. Err**	**p**	**[0.025, 0.975]**
Intercept	1.6522	0.095	0.000	[1.466, 1.839]
cnt_total_sources	-0.1695	0.044	0.000	[-0.255, -0.084]
np.power(cnt_total_sources, 2)	0.0194	0.005	0.000	[0.009, 0.03]
cnt_traffic	0.1040	0.020	0.000	[0.066, 0.142]
cnt_speech	0.1329	0.025	0.000	[0.084, 0.181]
cnt_other_urban	0.0589	0.034	0.087	[-0.008, 0.126]
cnt_nature	-0.0743	0.019	0.000	[-0.112, -0.036]
N	R^2	**Adj.** R^2	$RMSE_{Test}$	$RMSE_{Train}$
2242	0.122	0.120	1.325	1.334

$$\begin{aligned}\text{Mean\_annoyance}_\text{i} =& \alpha + \beta_1 \text{cnt\_total\_sources}_\text{i} + \beta_2 \text{cnt\_total\_sources}_\text{i}^2 + \\ &+ \beta_3 \text{cnt\_traffic} + + \beta_4 \text{cnt\_other\_urban}_\text{i} \\ &+ \beta_5 \text{cnt\_speech}_\text{i} + \beta_6 \text{cnt\_nature}_\text{i} + \beta_7 \text{mean\_spec\_cent}_\text{i} \\ &+ \beta_7 \text{std\_spec\_cent}_\text{i} + \epsilon,\end{aligned}$$

Since the distribution of mean annoyance is right-skewed (Figure 1), we applied a box-cox transformation to $mean\_annoyance$ improve model fit. Variable $cnt\_total\_sources$ is modelled with an additional quadratic term as a non-linear association is identified between the number of sound sources and mean annoyance in Figure 3. Since the number of sound sources are measured with a much smaller scale than sonic characteristics, all independent variables are z-scale standardised.

The results for the model without spectral traits (**W1**) and with spectral traits (**W2**) are reported in Table 2 and Table 3, respectively. The full model outputs and diagnostics are reported in Section A in the Appendix. R^2 and Adjusted R^2 are easily calculated for linear regression models and are reported in Tables 2 and 3 alongside the root-mean-square error (RMSE) for comparison with the models which follow.

The residuals are normally distributed with approximately constant variance (shown in Figures 8 and 9), thus the conditions for inference in multiple regression are met.

Table 3: Summary of results from Model **W.2**

	Coef	**Std. Err**	**p**	**[0.025, 0.975]**
Intercept	1.5154	0.091	0.000	[1.338, 1.693]
cnt_total_sources	-0.0980	0.042	0.018	[-0.179, -0.017]
np.power(cnt_total_sources, 2)	0.0119	0.005	0.019	[0.002, 0.022]
cnt_traffic	0.0699	0.019	0.000	[0.033, 0.107]
cnt_speech	0.0990	0.024	0.000	[0.053, 0.145]
cnt_other_urban	0.1035	0.033	0.002	[0.039, 0.168]
cnt_nature	-0.1218	0.019	0.000	[-0.158, -0.085]
mean_spec_cent	0.1942	0.012	0.000	[0.171, 0.218]
std_spec_cent	-0.0449	0.011	0.000	[-0.067, -0.022]
N	R^2	**Adj.** R^2	$RMSE_{Test}$	$RMSE_{Train}$
2242	0.213	0.211	1.217	1.254

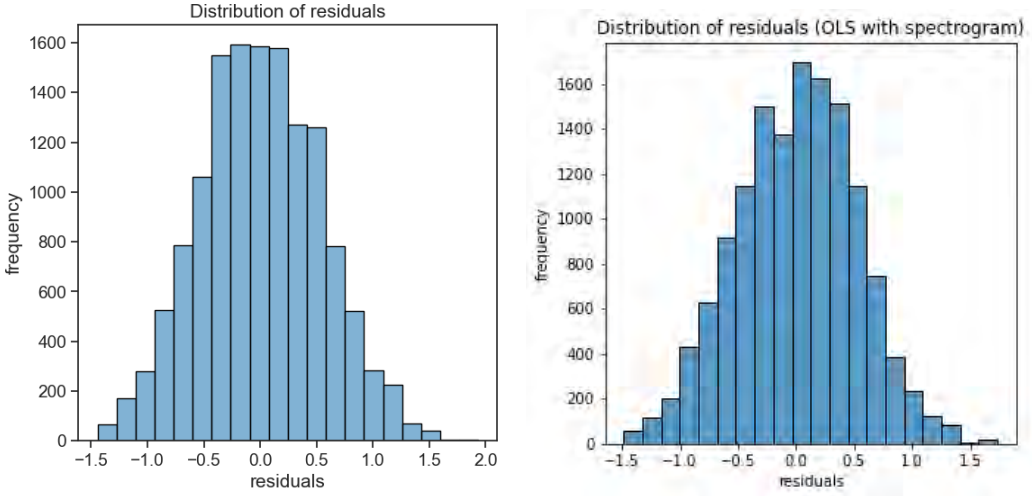
All variables are significant at 5% level. $mean\_spec\_cent$ has the strongest positive impact on mean annoyance, suggesting higher general frequency soundtracks tend to be more annoying. On the other hand, cnt_nature has the strongest moderating effect on annoyance, as expected from the box plots in Figure 4a.

When spectral traits are added for Model **W2** , the adjusted R-squared increases from 0.120 (see Tables 2 and 3) to 0.211. Since $mean\_spec\_cent$ is also the strongest positive predictor of annoyance rating, **our regression results suggest sonic characteristics have significant explanatory power on annoyance rating.**

Model performance is also better with spectral traits (**W2**). The test RMSE is 1.217, while the test RMSE for the model without spectral traits (**W1**) is 1.325. The train RMSE of the model with spectral traits (**W2**) is 1.254, while the one without is 1.334 (**W1**).

Since the dataset is unbalanced, as shown in Figure 2, we evaluated the OLS performance for high and low annoyance levels using AR = 5 as a threshold. Figure 10 plots the residual distribution among high and low annoyance ratings in the evaluation set. AR for soundtracks with high actual annoyance ratings are generally under-predicted.

While the OLS results show that sonic characteristics are also important predictors, the low R-squared indicates that we need more information to



(a) Model **W1** (b) Model **W2**

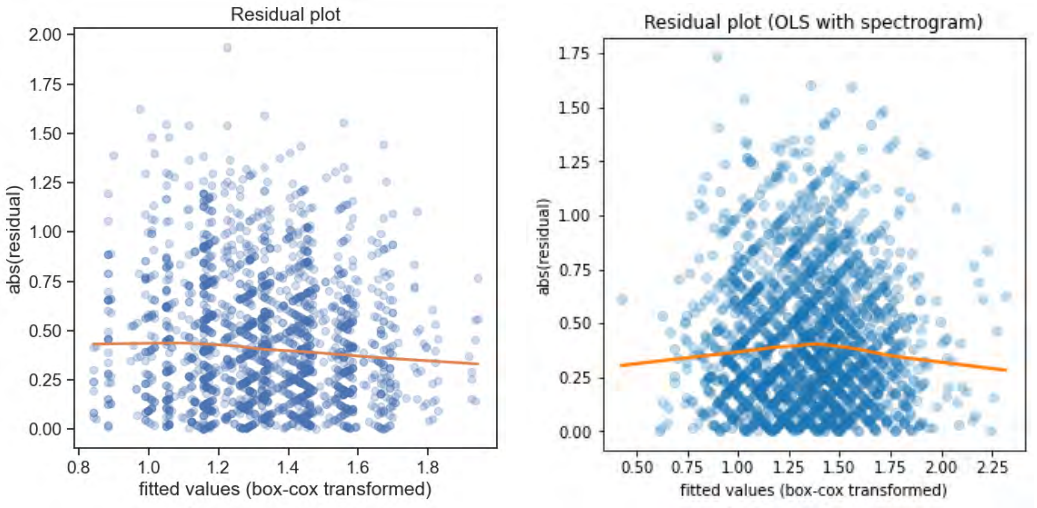
Figure 8: Distribution of Residuals for OLS Models

explain the variability in annoyance ratings. Other spectral traits or perceptive elements not captured by the survey might explain more variability in the data. These spectral features may represent elements which can be extracted within the deep learning approaches presented later.

4.2 Random Forest Models - W3 and W4

OLS assumes a linear and additive relationship between the predictors and the target variable. To capture a possibly complex and non-linear relationship between predictors and variables, we applied random forest models, an ensemble machine learning method that operates by combining the predictions of multiple decision trees. Random forest offers good predictive performance and interpretability through feature importance, and has been previously used in predictive soundscape modelling [Lionello et al., 2020].

In contrast to the preceding models, we also included all the sound source labels present, together with variables described in the regression model. We applied a random search with a five-fold cross-validaion (CV) to identify



(a) Model **W1** (b) Model **W2**

Figure 9: Fitted values scatterplots for OLS Models

Table 4: Summary of Random Forest results

Model	$RMSE_{Train}$	$RMSE_{Test}$
W3	1.153	1.231
W4	1.032	1.171

the optimal number of trees in the random forest. As with the OLS models in the previous section, we trained two random forest models, one with spectral characteristics (**W4**) and one without (**W3**).

For the random forest with spectral characteristics, the best model has 344 trees. The training and testing RMSE results are shown in Table 4

The variables used in the OLS model are also top predictors in the random forest, shown in Figure 11. The mean spectral centroid is the most importance predictor. The standard deviation of the spectral centroid, while with a relatively low coefficient magnitude in the OLS, is the second most important predictor in the random forest. The number of total sound sources is also among the top predictors, providing additional confirmation of the initial results from Mitchell et al. [2022a].

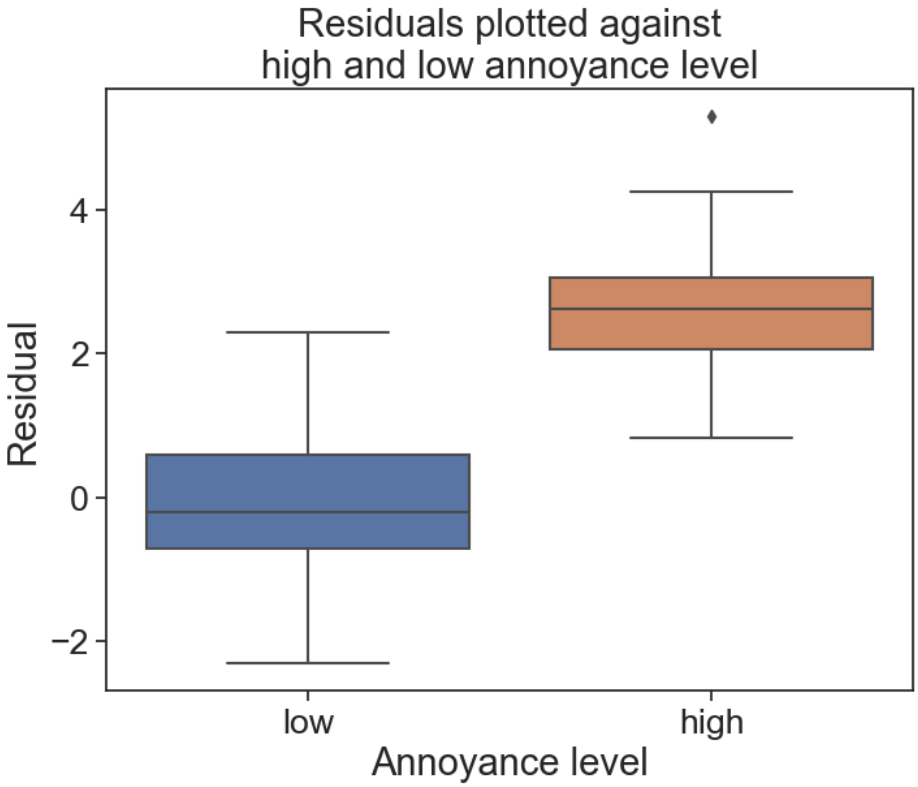


Figure 10: Distribution of residuals for **W2** across high and low annoyance level ratings. High annoyance ratings are underpredicted.

Bird tweet and construction stand out as two distinct sound sources with particularly high feature importance. Since bird tweet is labelled a nature sound and construction is labelled an 'other urban' sound in the regression, the magnitude of the coefficient values on cnt_nature and cnt_other_urban in Table 3 might have been driven by these two sounds.

The classical models thus show that sonic characteristics are as, if not more important, than the number of sound sources. The best performing of the classical models is **W4** (random forest with spectral characteristics) with a test RMSE of 1.171. This result leads into the following deep learning models and our primary question. Since sonic characteristics have been confirmed as important and likely other spectral characteristics

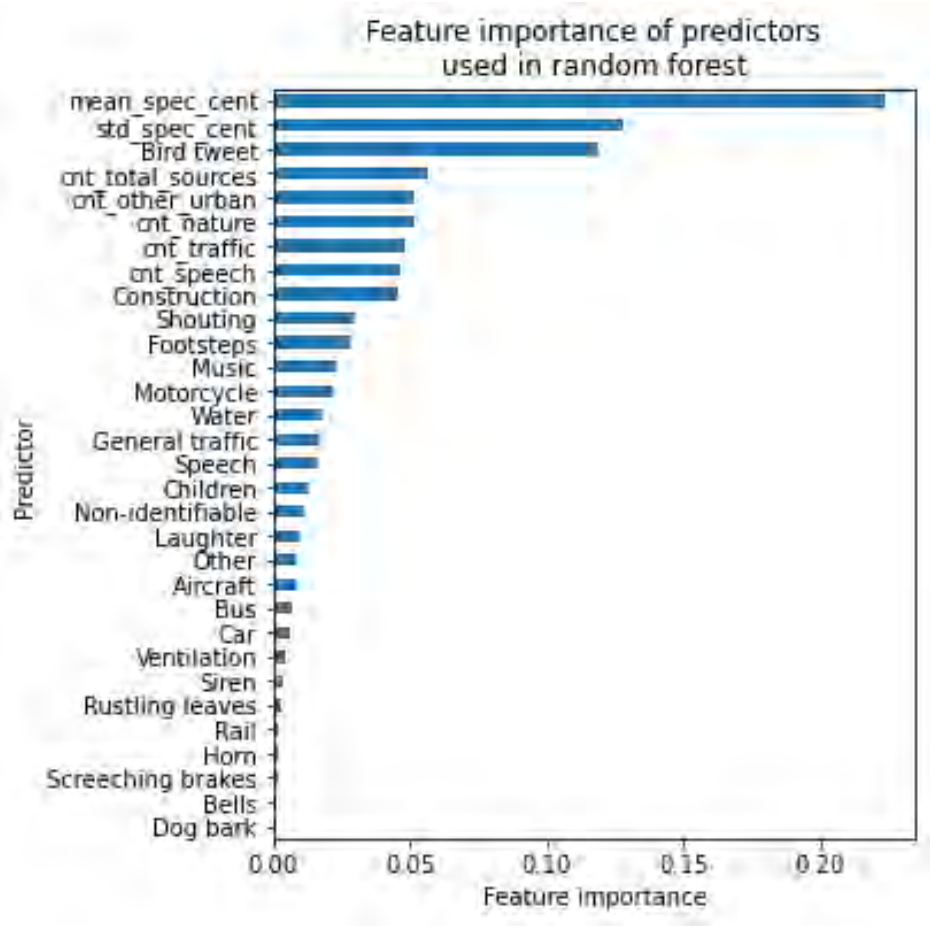


Figure 11: Feature importance of predictors used in random forest model.

can be extracted, how much additional predictive performance will the models achieve when sound source information is included?

5 Experiment 2: Deep Learning Methods

The primary goal of this DSG was to investigate various methods of employing deep learning for predicting noise annoyance. Specifically, we wished to trial ways of incorporating sound source information - whether from human labels derived from a survey or from automated source recognition - to augment the annoyance prediction. This section reports

the results of 11 deep learning (DL) models trained to predict annoyance ratings.

Inputs To generate the inputs for the deep learning models, we first calculated a mel spectrogram for each recording. To capture both the frequency information of a signal and its variation over time, a spectrogram computes a Fast Fourier Transform (FFT) of windowed sections of the signal (demonstrated in Figure 12).

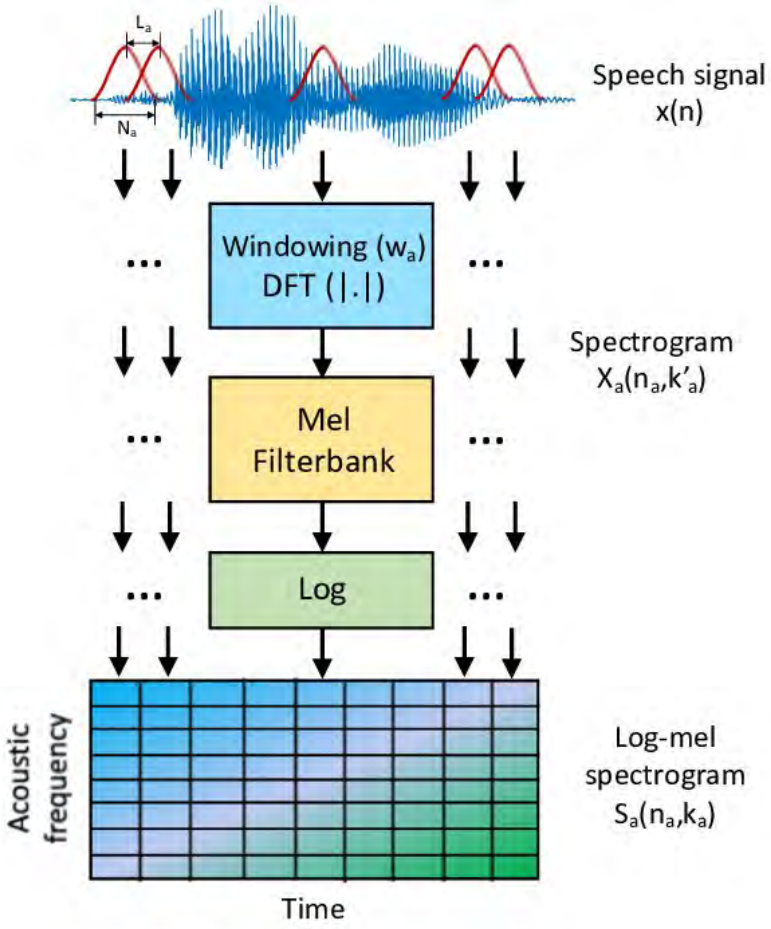


Figure 12: Demonstration of the calculation process for a spectrogram. Image from Gallardo-Antolín and Montero [2021].

By combining these windowed spectra, we can create a spectrogram which visualises the temporal pattern of the spectral content of the full recording. The mel scale provides a linear scale for the human auditory system (i.e. where each step in frequency on the mel scale is judged to be equal in distance from one another). This provides a spectrogram which theoretically better approximates the human auditory system and is commonly used in Audio Classification systems [Hershey et al., 2017, Thornton, 2019]. An example of a mel spectrogram computed for one of the DeLTA recordings is given in Figure 13.

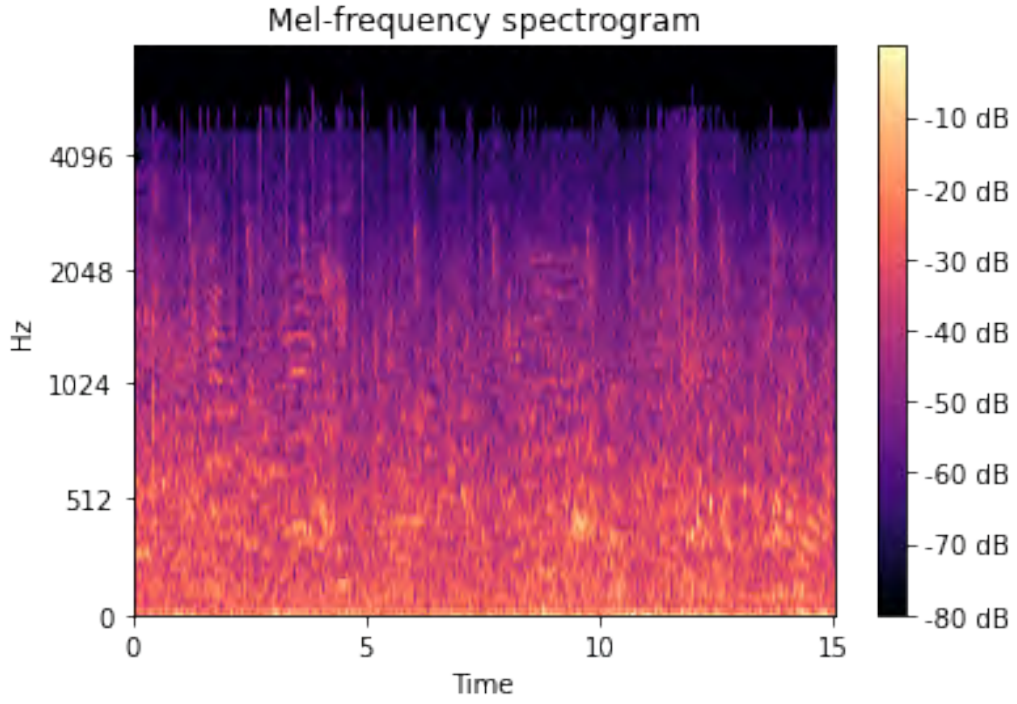


Figure 13: Mel spectrogram computed for one of the DeLTA recordings (2CV11_1.wav).

By default, the mel spectrograms are calculated with 128 mel frequency bins and a hop length of 512, giving a fairly high resolution spectrogram. For the calculation of mel spectrograms, hop length denotes the number of samples between successive frames, determining the temporal resolution and the amount of overlap between those frames in the audio

signal. For some of the models presented below, the possibility for low resolution spectrograms to be used was also tested, with the goal of decreasing the input size and speeding up training times. For these, the spectrograms were resized via interpolation to 32 by 128.

In addition to the mel spectrogram, those models which include sound source information also have the 24 binary source labels as inputs.

5.1 Models without sound source information

This section presents the results from the first set of deep learning models, which do not include any information about the sound source. The models from this set are shown in Figure 14, with structures shown in Appendix C. Each of the models trained is described below.

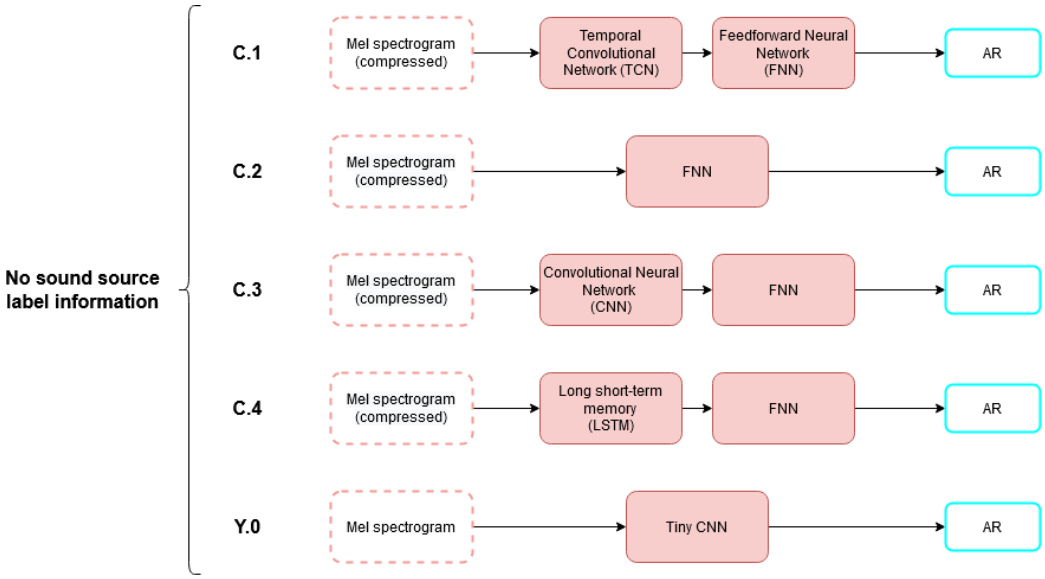


Figure 14: No Sound source information used

C1 While many studies have applied Convolutional Neural Networks to address noise and acoustics issues, this model is using the Temporal Convolutional Network (TCN). Like a CNN, the TCN can treat the time-series information in the spectrogram as an image, however TCN's consider the time-dependence of the x-axis of the image. The TCN

architecture in particular carries certain features that allow for greater effectiveness in time-series analysis, as compared to generic CNNs. Model **C1** uses a TCN on a compressed mel spectogram, feeds the result to the Feed-Forward Neural Network (FNN) and outputs an Annoyance Rating (AR). (See Appendix B for detailed implementation notes on TCN)

C2 Model uses the standard FNN on the compressed mel spectogram and outputs an AR.

C3 Model uses a Convolutional Neural Network (CNN) applied to compressed mel Spectogram. Then the FNN is applied to produce an AR.

C4 Model applies an Long Short-Term Memory (LSTM) network to compressed mel spectograms, then feeds the output feature vector to an FNN and outputs AR.

Y0 This model explores whether the annoyance rate can be successfully predicted based only on a simple convolutional neural network. The model receives a high resolution mel spectogram, and trains a TinyCNN to output an AR.

5.2 Models including sound source information

Three approaches to incorporate the sound source information are explored. The first directly inputs the human-generated labels from the DeLTA dataset. The second approach replaces these human-generated labels with labels generated by a pretrained, publicly available sound source recognition model (PANN) [Hershey et al., 2017]. These first two approaches both attempt to explicitly inform the model about the sound sources present by including source labels as direct inputs. The goal of the PANN-based models is to investigate the possibility of using automated source recognition to remove the 'human-in-the-loop' aspect. This would allow the model to be used in an automated context.

The third approach includes the human-generated labels from DeLTA as outputs only. We consider this to be *implicitly* including the sound source information. By training the models to jointly predict the AR and to predict the sound sources in the recording, the latent embedding of the models will be trained to include information about the sources based on the mel spectrogram.

5.2.1 Human-generated Labels

Figure 15 shows the structures of the models based on human-generated labels.

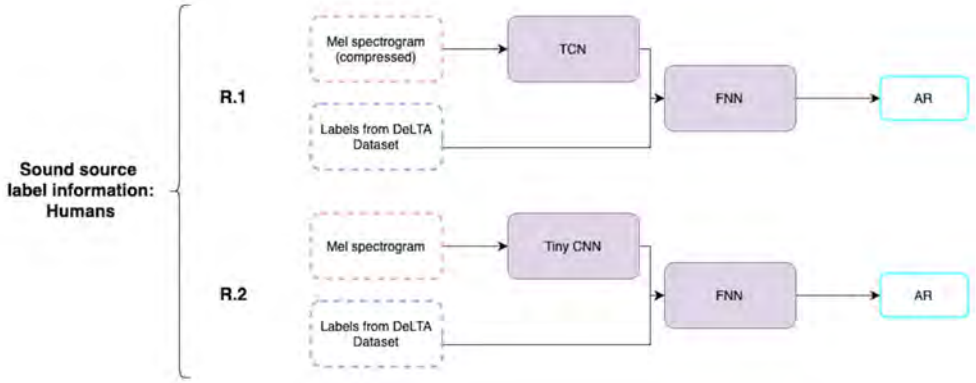


Figure 15: Sound source including models with human generated labels.

R1 Model applies TCN to mel spectograms and combines the result with the human-generated labels and feeds that to the FNN to produce AR.

The rationale behind the model is that adding sound source information as additional inputs to the model would add more value to the model. This was done in a framework such that the sound source data is stacked with latent feature representation of the spectrograms instead of plain stacking the data with the spectrogram. The spectrogram is treated as an image and direct stacking of source data with image data was not explored in this study.

Although we added only 24 features of source information (i.e. the binary sound source labels), the architecture selected to do this used a fully

connected FNN to predict the Annoyance rating. In this model, the sound source labels indicated the presence or absence of a sound source in the given sound. The FNN model used stacked latent spectrogram features that come out of TCN feature extraction and the one-hot encoded sound source labels as inputs.

Our speculative reasoning for the models performing poorly is due to the additional complexity of the architecture used to stack the two data formats. We attribute this to the complexity of the model since simple model that does not include sound source data as inputs is able to perform better and does not overfit on training data.

R2 Model applies TinyCNN to mel spectograms, combines the result with the human-generated labels, and feeds that to the FNN to produce AR.

This model is very similar in architecture to model **R1** as we use the spectrogram and sound source information as inputs to the model. There are two fundamental differences between these two models. Firstly, we use the TinyCNN as a feature embedding network instead of the TCN architecture as in other experiments the TinyCNN model had better performance, as will be demonstrated in Section 5.3. This is shown in models **Y1** and **Y4** .

The second fundamental difference in the model is the representation of the sound source information. Building on the results of the preceding models and the investigations into the ambiguity of human labelling presented in Section 3.2, this model was altered to use the degree of agreement for a sound source ID rather than a simple binary of source presence. The source labels are migrated from one hot encoding, which indicates only presence to a more proportional representation. Here the presence of data is proportional to the number of survey participants that identified a particular sound, as in Section 3.2 where the ambiguities of human labelling are discussed.

5.2.2 Pretrained IDs

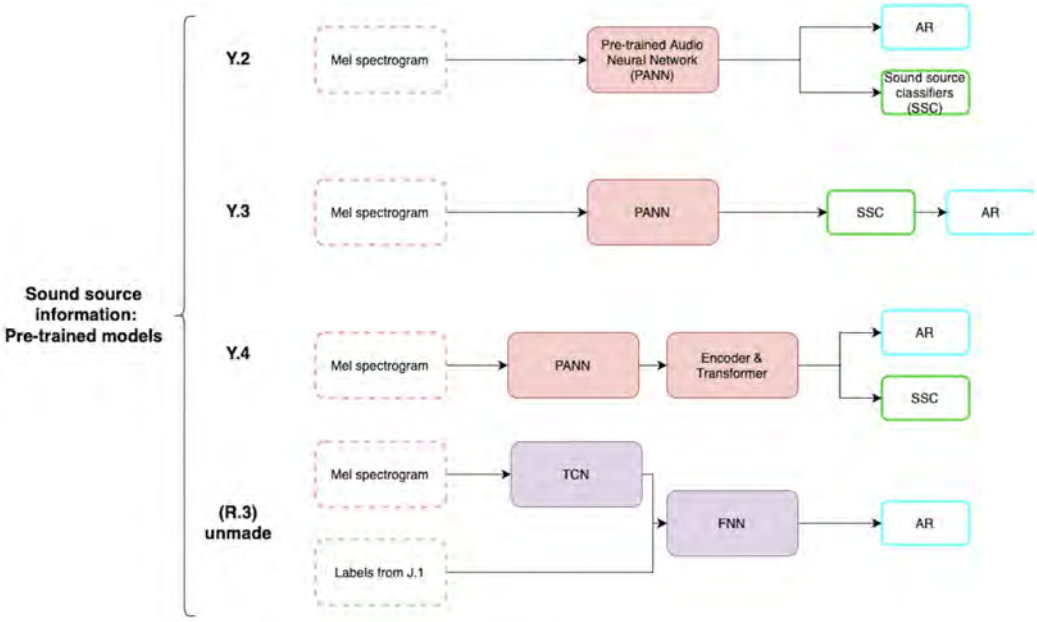


Figure 16: Sound Source including models with usage of Pretrained models

Y2 This model is using the PANN model (Pre-trained Audio Neural Network) [Kong et al., 2020], pretrained using sound source information from the Google Audio Set [Gemmeke et al., 2017a]. Audio Set is a hierarchically structured set of audio even categories such that each category (i) provides a comprehensive set that can be used to describe the audio events encountered in real-world recordings; (ii) can be distinguished by a ”typical” listener. The Audio Set dataset contains 632 audio event categories, arranged in a hierarchy with a maximum depth of six levels and a collection of 1,789,621 labelled 10-sec excerpts from YouTube videos. The PANN is applied to the mel spectrograms to perform sound source recognition and outputs both sound source classification and AR.

Y3 Similar to **Y2** , this model is using the PANN, pretrained using sound source information from Audioset. The PANN is applied to mel to produce

sound source classification vector which is then fed to a 1-layer FNN to output the annoyance rate. The sound source classification is also given as an output.

Y4 The model applies the PANN to the mel spectogram, then the resulting partial sound source information is fed to the separated multi-head attention encoder blocks to form the CnnT (CNN-Transformer). The model outputs both the sound source classification and an AR.

J.1 Throughout the DSG week, one of the experiment teams worked towards a model framework for the sound source recognition task, attempting to resolve the possible cross-over ambiguity between sound source labels (see Section 3.2). Unfortunately this work could not be completed meaning results and outcomes cannot be reported for this model. Figure 17 demonstrates the proposed framework for this model.

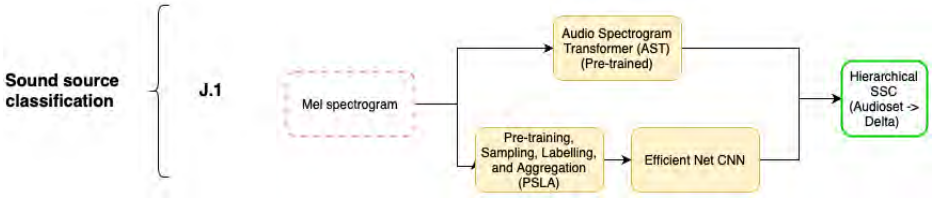


Figure 17: Sound source classification model

R3 This model was intended to combine the structure of **R1** with a bespoke the sound source recognition model **J.1**. The goal was to build on the progress made in Section 3.2, train (or tailor) our own separate model to predict sound source labels, then feed this into **R1** , replacing the human labels from DeLTA. Unfortunately, since the source recognition model could not be completed, **R3** was also not completed.

5.2.3 Sound source label as output (implicit source information)

This final set of prediction models takes a different approach to incorporating the sound source information. Rather than directly feeding

either human-generated or automatically identified source labels, we instead include the sources as a target to be jointly predicted alongside the AR. In this way, the trained model would implicitly include information about the sound sources embedded within the hidden layers of the CNN. This has the added benefit that the trained model itself is able to generate predicted sound source labels from just an input spectrogram.

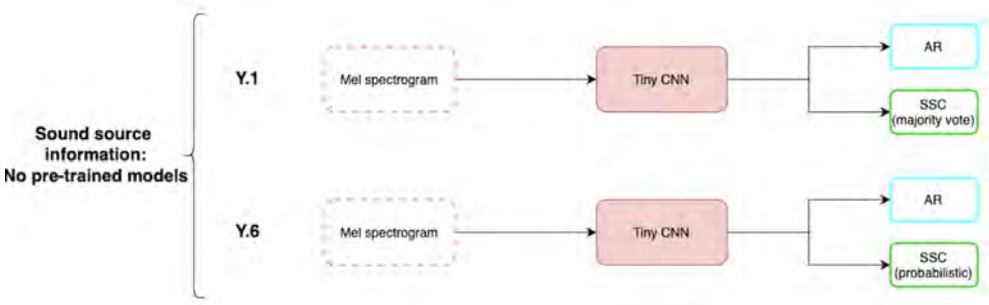


Figure 18: Models which implicitly include sound source information by jointly predicting AR and SSC.

Y1 Model receives a mel spectrogram, uses a TinyCNN to output both an AR and a one-hot encoded vector with 24 values corresponding to the sound source classification.

Y6 This model is very similar to **Y1** except that the output is a vector of the sound source identification proportion for each label, as derived in Section 3.2. Thus, the model is trained using data which reflects the ambiguity of human labelling, rather than encoding the labels as definite yes/no binary answers.

5.3 Model comparison.

Each of the models described above is trained on the 80% training set and tested with the 20% holdout set. For all models, the Root Mean Squared Error (RMSE) is used as the testing score for the Annoyance Rating (AR) which has a range from 1-10. For those models which also output sound source classification (SSC) predictions, the area under the ROC curc (AUC) is reported as the test score [Fawcett, 2006], calculated

Table 5: Comparison of models

Category	Model	AR test (RMSE)	AR train (RMSE)	SSC test (AUC)	SSC train (AUC)
Set 1: Classical baseline	W.1	1.28	1.34	-	-
	W.2	1.22	1.25	-	-
	W.3	1.23	1.15	-	-
	W.4	1.17	1.03	-	-
Set 2: No source info	C.1	1.30	1.31	-	-
	C.2	3.87	3.78	-	-
	C.3	1.26	1.29	-	-
	C.4	1.50	1.33	-	-
	Y.0	1.13	1.03	-	-
Set 3a: Human labels	R.1	1.30	0.97	-	-
	R.2	1.29	0.68	-	-
Set 3b: PANN	Y.2	1.08	0.95	0.86	0.88
	Y.3	1.10	0.97	0.88	0.97
	Y.4	1.12	0.18	0.91	0.99
Set 3c: Implicit source info	Y.1	1.09	0.97	0.88	0.92
	Y.6	1.07	0.16	0.90	1.00

using the `metrics.auc_roc_score()` function from `scipy`. AUC provides an uncomplicated metric which has several benefits over simple accuracy. AUC decouples model performance from class skew (uneven distribution of class instances across the classes) and error costs (i.e. the tradeoffs between true positives and false positives) and have been commonly used for scoring multi-class classification models [Hand and Till, 2001]. Table 5 reports the training and testing scores for all of the models. Figure 19 plots a comparison of the various types of models tested.

Spectrogram resolution The very first conclusion that could be made from the model comparison is that the high resolution spectrogram considerably outperform a compressed spectrogram (e.g. **Y1** $-$ **Y6** models are better than **C1** $-$ **C4**). In all but one case, those models which included high resolution spectrograms out-performed those which used the compressed spectrograms. What is particularly interesting is that even though **W2** and **W4** could also be said to incorporate 'highly compressed spectrogram' information (i.e. the spectral characteristics used are single-dimension representations of the full spectrogram), in

general they still outperform the compressed spectrogram neural network models, even those which also include sound source labels.

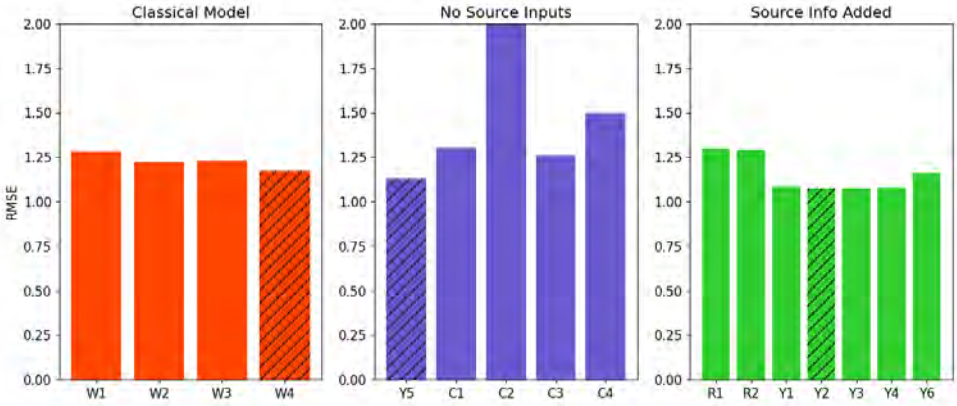


Figure 19: AR RMSE results (lower is better).

Inclusion of sound source information To further investigate the differences among model architecture and the inclusion of sound source information, we will limit our discussion to the high-resolution spectrogram models: **R2** , **Y0** , **Y1** , **Y2** , **Y3** , **Y4** , and **Y6** . A comparison of the AR performance of **Y0** to **Y1** -**Y6** demonstrates a clear improvement when some form of sound source information is included. Although **Y0** (RMSE=1.13) outperforms **R1** (RMSE=1.30) and **R2** (RMSE=1.30) which explicitly included human-generated labels as inputs, it is at least marginally outperformed by all of the comparable high-resolution models with sound source information (RMSE $<$1.12).

In particular, we can compare the performance of **Y0** and **Y1** which share nearly identical architectures. Both are a TinyCNN with the only difference being the number of output nodes: one linear regression node for AR in **Y0** and one linear regression and 24 linear classification nodes in **Y1** . By implicitly including sound source information through joint prediction, **Y1** (RMSE=1.09) shows a marked improvement over **Y0** (RMSE=1.13).

Model complexity Although **R2** also made use of the high-resolution spectrograms, its performance is significantly worse than the **Y** models.

As noted earlier, our speculation is that this is due to the complexity of the model architecture and the use of the FNN. The model structure itself is similar to **C1** , making use of a TCN and an FNN and we would expect that the performance would improve since **R1** includes the human-generated sound source labels. This is indeed true in the training performance, with **R1** achieving a training RMSE of 0.97 compared to the **C1** training RMSE of 1.31. However, both models have identical testing performance, indicating the increased complexity of **R1** likely caused it to overfit to the training data.

A similar issue may contribute to the discrepancy in training and testing scores for model **Y6** . While **Y1** and **Y6** have identical architectures, binary SSC labels are used for **Y1** whereas probabilistic SSC labels derived from Section 3.2 are used as the outputs for **Y6** . While **Y6** slightly outperforms **Y1** for the AR task on the test set, its training RMSE score is significantly lower, which would indicate that it has overfit to the training data and may be less generalisable. It is possible this overfitting issue is due to some additional complexity in the training process, however more work is required to thoroughly determine whether any benefit may be gained by using the probabilistic source labels to reflect the ambiguity in human-generated labelling.

Best performing models Overall, the best performing models were **Y1** , **Y2** and **Y6** , all achieving a testing RMSE below 1.10. Although **Y6** technically achieves the best performance of these three, it has a wide difference between the training and testing score, indicating it may be overfitted. As such, it would appear that **Y1** and **Y2** are the best models out of the 12 neural networks that were trained. Of these, it is particularly interesting that **Y1** , although it is significantly simpler overall and does not incorporate a pretrained sound recognition model, nearly matches the performance of the other more complex model. This is especially interesting when considering that **Y1** outperforms **Y2** on the audio classification task.

There are very few existing models within the literature whose performance could be directly compared with the models presented here, so it is difficult to establish what is considered a 'strong' or 'weak' RMSE within this task. This report is the first piece of work using the DeLTA

dataset and, more broadly, attempting to predict individual sound perception ratings is a relatively recent development in the field. While predicting the annoyance due to a single source (e.g. traffic or electric vehicles) or predicting community annoyance levels are common tasks which achieve very high accuracy, predicting individual annoyance perception for complex realistic soundscapes with multiple sound sources is a much more challenging task which has only recently seen attention. Given this, we cannot yet state absolutely whether the prediction scores achieved here are sufficient for practical applications, but when compared against other attempt within the soundscape perception prediction field (see Ooi et al. [2023, 2022], Mitchell et al. [2021b]) the results achieved are promising.

5.4 Conclusions

Based on these results, our primary conclusions are that:

1. Using higher resolution spectrograms outperforms compressed spectrograms.
2. In general, simpler model structures performed better on this dataset than more complex models, with the TinyCNN models performing surprisingly well despite their simplicity.
3. Including sound source information does improve the prediction accuracy, however this is often not enough to overcome the two previous factors, demonstrating the importance of the model structure chosen.
4. When comparing like models, it appears that including the sound source information as an output target can match the performance of more complex models which make use of large pretrained audio networks. In addition, the added complexity necessary to incorporate sound source labels as input features appears to drastically reduce the predictive performance, making it difficult to directly compare implicit vs explicit methods of incorporating human-generated sound source information.

6 Future work and research avenues

With the exception of the classic modelling approaches, in depth examination of the model prediction output was not undertaken beyond assessment of the training and testing RMSE scores. Further development of these models would benefit from exploration of this. This may help identify whether the models are capable of consistently identifying high annoyance sounds. Whilst data augmentation is a common process for image-based neural networks, this was not explored for our models. Due to the recent success of data augmentation in the speech and vision domains, augmentation methods for this audio data would be a useful avenue or further work [Park et al., 2019]. Several augmentation methods drawn from the domain of image recognition which operate on the mel spectrogram could be applied, to expand the training dataset:

- Creating a mask in the x-axis (time-domain) or in the y-axis (frequency masking).
- Warping or skewing the image
- Adding white noise - this could be added either as audio white noise to the recording before calculating the mel spectrogram or as visual noise to the spectrogram image.

Although the recordings in the dataset are two-channel, we only made use of one channel in our modelling. Future work could explore the possible benefits of considering the full binaural signal. This would provide additional information including the inter-aural time and level differences.

Coming into this challenge, there was an expectation that incorporating sound source labels as inputs into the network would lead to better predictions. However, our work demonstrated that this explicit approach to incorporating sound source information added unnecessary complexity and reduced model performance. Building on the success of jointly predicting sound source labels and annoyance ratings, more sophisticated models could be developed using our approach as a starting point. One fruitful avenue for further research could be to explore the possibility that this result also works in the inverse i.e. including

perceptual features as a joint output could improve the performance of sound source classification tasks.

Team members

Participants

Emmeline Brown is a fourth-year PhD student affiliated with the Centre for Computational Medicine, UCL; Centre for Advanced Biomedical Imaging, UCL; and Moorfields Eye Hospital. Her research aims to apply deep learning and mechanistic modelling methods to ophthalmological imaging data for the development of solutions that could be deployed in clinic. Previously she worked as a research assistant in bioinformatic analysis in Parkinson's disease genome-wide association studies at Queen's Square Institute of Neurology, UCL. She contributed to the project by implementing and training machine learning models, and drafting the final presentation.

Ratneel Deo (Facilitator) Ratneel is a PhD scholar at the University of Sydney. He studies coral classification from fossil and benthic imagery via unsupervised and supervised deep learning methods. He contributed to this project by supporting the challenge team in scheduling and managing project timeline in his capacity as a facilitator.

Yuanbo Hou is a PhD student at Ghent University in the Wireless, Acoustics, Environment, & Expert Systems (WAVES) Group. His research focuses on deep learning methods for acoustic scene classification and audio event detection. He contributed to this project by implementing and training several neural network models.

Jasper Kirton-Wingate is a 1st year PhD student at Edinburgh Napier University as part of the COG-MHEAR programme. His research interests are in multi-modal, personalised sound enhancement modelling algorithms for the hearing impaired.

Jinhua Liang is a 2nd-year Ph.D. student at the Centre for Digital Music, Queen Mary University of London. His research goal is to recognise environmental sounds with limited annotations by investigating few-shot learning, transfer learning, and self-/semi-supervised learning. He

contributed to this project by developing tools for audio signal processing and design the architecture of deep neural networks.

Alisa Sheinkman is a third-year PhD candidate studying at the School of Math at the University of Edinburgh. Her thesis develops advances in Bayesian Deep Learning with a focus on efficient inference schemes for Bayesian deep models and appropriate priors for the data-driven design of the network architecture. She contributed to this project by advancing the technical discussions and drafting the final report.

Christopher Soelistyo recently finished a PhD at University College London, and will soon begin a postdoctoral position at the Alan Turing Institute. His research aimed to utilise deep learning to discover the biophysical determinants of cell fate in cell competition, by first training a deep neural network to predict cell fate, then investigating the input features used by the model in forming its predictions. He contributed to the project by proposing, developing and training several of the models used for the source recognition and annoyance prediction tasks.

Hari Sood (Facilitator) is a Research Application Manager at the Alan Turing Institute, focused on finding use-cases and real-world implementation for the Institute's Data Safe Haven that follow open and reproducible research principles. He was a facilitator for the group, ensuring the project was always moving forwards, people were heard and able to contribute, and ideas were synthesized down into a coherent narrative.

Arin Wongprommoon is a fourth-year PhD student at the Centre for Engineering Biology, University of Edinburgh. His research aims to characterise the yeast metabolic cycle, a biological oscillator in budding yeast, doing so by using single-cell fluorescence microscopy, image segmentation, analysis of oscillatory time series, and characterisation of such time series using unsupervised and supervised machine learning approaches. He contributed to the project by managing software development and operation practices (DevOps), illustrating data augmentation algorithms, and drafting the final presentation.

Kaiyue Xing works as Laboratory Manager at School of Education, Communication and Language Science, Newcastle University. She graduated from the University of Manchester with a PhD degree in

Linguistics in 2022. Her research focuses on the tongue movement of Mandarin rhotic sounds with Ultrasound Tongue Image techniques and explores the correlation between articulation and acoustics. In the current project, she contributes to the visualisation of ambiguity of the source resources in human response.

Wingyan Yip is a Business Intelligence Specialist at Soldo Ltd. She has a strong interest in policy evaluation with population data. She currently works on product and marketing analytics, focusing on using inference approaches to guide acquisition and retention strategies. She contributed to the current project by creating baseline models (classical) and engineering features through data exploration.

Principal Investigator and Challenge Owner

Andrew Mitchell is a Research Fellow in Urban Soundscape Modelling at University College London. His research advances the application of engineering approaches to urban soundscape design through the development of novel machine learning models for predicting soundscape perception. Through his work on the Deep Learning Techniques for noise Annoyance detection (DeLTA) Project, led by Dr Francesco Aletta, Andrew acted as the Challenge Owner for this DSG by providing the datasets and proposing the initial challenge. In addition, Andrew acted as the Academic Principal Investigator, shaping the research questions, leading and organising the technical discussions, and drafting the final report.

Funding

This work was supported by UCL Health of the Public, via the Small Grants Scheme 2020-2021, project: ”Deep Learning Techniques for noise Annoyance detection” (DeLTA). This work was supported by Towards Turing 2.0 under the EPSRC Grant EP/W037211/1 The Alan Turing Institute, particularly the Postdoctoral Enrichment Awards.

References

Francesco Aletta, Tin Oberman, Andrew Mitchell, Huan Tong, and Jian Kang. Assessing the changing urban sound environment during the COVID-19 lockdown period using short-term acoustic measurements. *Noise Mapping*, 7(1):123–134, January 2020. ISSN 2084-879X. doi: 10.1515/noise-2020-0011. URL https://www.degruyter.com/view/journals/noise/7/1/article-p123.xml.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *Arxiv*, 2018. doi: 10.48550/ARXIV.1803.01271. URL https://arxiv.org/abs/1803.01271.

Daniel P. W. Ellis, Brian Whitman, Adam Berenzweig, and Steve Lawrence. The quest for ground truth in musical artist similarity. *Proceedings: Third International Conference on Music Information Retrieval*, 2002. doi: 10.7916/D80R9ZRG.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, jun 2006. doi: 10.1016/j.patrec.2005.10.010.

Ascensión Gallardo-Antolín and Juan Montero. On combining acoustic and modulation spectrograms in an attention lstm-based system for speech intelligibility level classification. *Neurocomputing*, 05 2021. doi: 10.1016/j.neucom.2021.05.065.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017a. doi: 10.1109/ICASSP.2017.7952261.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017b.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv*, 2021. doi: 10.48550/ARXIV.2104.01778. URL https://arxiv.org/abs/2104.01778.

Rainer Guski, Dirk Schreckenberg, and Rudolf Schuemer. WHO environmental noise guidelines for the european region: A systematic review on environmental noise and annoyance. *International Journal of Environmental Research and Public Health*, 14(12):1539, dec 2017. doi: 10.3390/ijerph14121539.

David J. Hand and Robert J. Till. A simple generalisation of the Area Under the ROC Curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001. doi: 10.1023/a:1010920819831.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017. doi: 10.1109/ICASSP.2017.7952132.

ISO/TS 12913-2:2018. Acoustics – Soundscape – Part 2: Data collection and reporting requirements, 2018. International Organization for Standardization, Geneva, Switzerland, 2018.

Jian Kang, Francesco Aletta, Tin Oberman, Mercede Erfanian, Magdalena Kachlicka, Matteo Lionello, and Andrew Mitchell. Towards soundscape indices. In *Proceedings of the 23rd International Congress on Acoustics*, volume integrating 4th EAA Euroregio 2019 : 9-13 September 2019, pages 2488–2495, Aachen, September 2019. RWTH Aachen University. doi: 10.18154/RWTH-CONV-239249. URL https://www.researchgate.net/publication/335661596{_}Towards{_}soundscape{_}indices.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. doi: 10.1109/taslp.2020.3030497.

Phu Ngoc Le, Eliathamby Ambikairajah, Julien Epps, Vidhyasaharan Sethu, and Eric HC Choi. Investigation of spectral centroid features for cognitive load classification. *Speech Communication*, 53(4):540–551, 2011.

Matteo Lionello, Francesco Aletta, and Jian Kang. A systematic review of prediction models for the experience of urban soundscapes. *Applied Acoustics*, 170, June 2020. ISSN 0003-682X. doi: 10.1016/j.apacoust.2020.107479. URL https://linkinghub.elsevier.com/retrieve/pii/S0003682X20305831.

Andrew Mitchell, Tin Oberman, Francesco Aletta, Mercede Erfanian, Magdalena Kachlicka, Matteo Lionello, and Jian Kang. The Soundscape Indices (SSID) Protocol: A Method for Urban Soundscape Surveys–Questionnaires with Acoustical and Contextual Information. *Applied Sciences*, 10(7):2397, April 2020. ISSN 2076-3417. doi: 10.3390/app10072397.

Andrew Mitchell, Tin Oberman, Francesco Aletta, Mercede Erfanian, Magdalena Kachlicka, Matteo Lionello, and Jian Kang. The International Soundscape Database: An integrated multimedia database of urban soundscape surveys – questionnaires with acoustical and contextual information, October 2021a. URL https://doi.org/10.5281/zenodo.5578572.

Andrew Mitchell, Tin Oberman, Francesco Aletta, Magdalena Kachlicka, Matteo Lionello, Mercede Erfanian, and Jian Kang. Investigating urban soundscapes of the COVID-19 lockdown: A predictive soundscape modeling approach. *The Journal of the Acoustical Society of America*, 150(6):4474–4488, December 2021b. doi: 10.1121/10.0008928.

Andrew Mitchell, Mercede Erfanian, Christopher Soelistyo, Tin Oberman, Jian Kang, Robert Aldridge, Jing-Hao Xue, and Francesco Aletta. Effects of soundscape complexity on urban noise annoyance ratings: A large-scale online listening experiment. *International Journal of Environmental Research and Public Health*, 19(22), 2022a. doi: 10.3390/ijerph192214872. URL https://www.mdpi.com/1660-4601/19/22/14872.

Andrew Mitchell, Mercede Erfanian, Christopher Soelitsyo, Tin Oberman, and Francesco Aletta. Delta (deep learning techniques for noise annoyance detection) dataset, 2022b. URL https://doi.org/10.5281/zenodo.7158056.

Kenneth Ooi, Karn N. Watcharasupat, Bhan Lam, Zhen-Ting Ong, and Woon-Seng Gan. Probably pleasant? a neural-probabilistic approach to automatic masker selection for urban soundscape augmentation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2022. doi: 10.1109/icassp43922.2022.9746897.

Kenneth Ooi, Zhen-Ting Ong, Karn N. Watcharasupat, Bhan Lam, Joo Young Hong, and Woon-Seng Gan. ARAUS: A large-scale dataset and baseline models of affective responses to augmented urban soundscapes. *IEEE Transactions on Affective Computing*, pages 1–17, 2023. doi: 10.1109/taffc.2023.3247914.

Ferran Orga, Andrew Mitchell, Marc Freixes, Francesco Aletta, Rosa Ma Alsina-Pagès, and Maria Foraster. Multilevel Annoyance Modelling of Short Environmental Sound Recordings. *Sustainability*, 13(11):5779, May 2021. ISSN 2071-1050. doi: 10.3390/su13115779. URL https://www.mdpi.com/2071-1050/13/11/5779.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA, sep 2019. doi: 10.21437/interspeech.2019-2680. URL https://doi.org/10.21437%2Finterspeech.2019-2680.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification, 2015. URL https://doi.org/10.7910/DVN/YDEPUT.

Monika Rychtáriková and Gerrit Vermeir. Soundscape categorization on the basis of objective acoustical parameters. *Applied Acoustics*, 74(2): 240–247, February 2013. ISSN 0003-682X. doi: 10.1016/j.apacoust.2011.01.004.

Emery Schubert, Joe Wolfe, Alex Tarnopolsky, et al. Spectral centroid and timbre in complex, multiple instrumental textures. In *Proceedings of*

the international conference on music perception and cognition, North Western University, Illinois, pages 112–116. sn, 2004.

Mariola Śliwińska-Kowalska and Kamil Zaborowski. WHO environmental noise guidelines for the European region: A systematic review on environmental noise and permanent hearing loss and tinnitus. *International Journal of Environmental Research and Public Health*, 14 (10), 2017. ISSN 1660-4601. doi: 10.3390/ijerph14101139.

Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labelling of venus images. *Advances in neural information processing systems*, 7, 1994.

Peter Norvig Stuart Russell. *Artificial Intelligence: A Modern Approach, eBook, Global Edition*. Pearson ITP, April 2021. URL https://www.ebook.de/de/product/40661130/stuart_russell_peter_norvig_artificial_intelligence_a_modern_approach_ebook_global_edition.html.

BZJLS Thornton. Audio recognition using mel spectrograms and convolution neural networks. Technical report, UCSD, 2019.

W. Yang and J. Kang. Acoustic comfort evaluation in urban open public spaces. *Applied Acoustics*, 66(2):211–229, February 2005. doi: 10.1016/j.apacoust.2004.07.011.

Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: facts and models*. Springer, Berlin ; New York, third ed. edition, 2007. ISBN 978-3-540-23159-2. doi: 10.1007/978-3-540-68888-4.

Appendices

A OLS model table

A.1 Model W.1

Table 6: OLS results for model without spectrogram traits

Dep. Variable:	annoyance_t		R^2		0.122
Model:	OLS		**Adj.** R^2		0.120
Method:	Least Squares		**F-statistic:**		
Prob (F-statistic):	**Log-Likelihood:**		-1834.9		
No. Observations:	2242		**AIC:**		3685
Df Residuals:	2235		**BIC:**		3724
Df Model:	6				
	coef	**std err**	**t**	**P $>$ |t|**	**[0.025, 0.975]**
Intercept	1.6522	0.095	17.385	0.000	[1.466, 1.839]
cnt_total_source	-0.1695	0.044	-3.895	0.000	[-0.255, -0.084]
np.power(cnt_total_sources, 2)	0.0194	0.005	3.633	0.000	[0.009, 0.030]
cnt_traffic	0.1040	0.020	5.317	0.000	[0.066, 0.142]
cnt_speech	0.1329	0.025	5.362	0.000	[0.084, 0.181]
cnt_other_urban	0.0589	0.034	1.714	0.087	[-0.008, 0.126]
cnt_nature	-0.0743	0.019	-3.844	0.000	[-0.112, -0.036]
Omnibus:	10.328		**Durbin-Watson:**		1.450
Prob(Omnibus):	0.006		**Jarque-Bera (JB):**		0.120
Skew:	0.101		**Prob(JB):**		0.0105
Kurtosis:	2.762		**Cond. No.**		132

A.2 Model W.2

Table 7: OLS results for model with spectrogram traits

Dep. Variable:	annoyance_t		R^2		0.213
Model:	OLS		**Adj.** R^2		0.211
Method:	Least Squares		**F-statistic:**		75.77
Prob (F-statistic):	8.14e-111		**Log-Likelihood:**		-1711.5
No. Observations:	2242		**AIC:**		3441
Df Residuals:	2233		**BIC:**		3492
Df Model:	8				
	coef	**std err**	**t**	**P $>$ |t|**	**[0.025, 0.975]**
Intercept	1.5154	0.091	16.734	0.000	[1.338, 1.693]
cnt_total_source	-0.098	0.042	-2.360	0.018	[-0.179, -0.017]
np.power(cnt_total_sources, 2)	0.0119	0.005	2.339	0.019	[0.002, 0.022]
cnt_traffic	0.0699	0.019	3.731	0.000	[0.033, 0.107]
cnt_speech	0.0990	0.024	4.201	0.000	[0.053, 0.145]
cnt_other_urban	0.1035	0.033	3.169	0.002	[0.039, 0.168]
cnt_nature	-0.218	0.019	-6.572	0.000	[-0.158, -0.085]
mean_spec_cent	0.1942	0.012	16.106	0.000	[0.171, 0.218]
std_spec_cent	-0.0449	0.011	-3.932	0.000	[-0.067, -0.022]
Omnibus:	4.306		**Durbin-Watson:**		1.401
Prob(Omnibus):	0.116		**Jarque-Bera (JB):**		3.936
Skew:	-0.050		**Prob(JB):**		0.140
Kurtosis:	2.821		**Cond. No.**		133

B Temporal Convolutional Neural Network.

The fundamental idea behind the TCN is that time-series information can be treated as an image, where ”time” occupies one of the image axes. For example, if we have a multi-variable time-series input of shape $(t \times c)$ - where t is the number of time-steps and c is the number of variables - then this can be treated as a 2D mono-channel image where t and c are the width and height of the image respectively. The framing of input information as an image immediately raises the possibility for the use of CNNs. However, the TCN architecture in particular carries certain features that allow for greater effectiveness in time-series analysis, as compared to generic CNNs.

The first is that the convolutions used are causal, meaning that in the course of a convolution operation, information flows only from the past to

the future and not the other way round. In concrete terms, this means that, assuming the ”time” axis is the x-axis, a convolution applied to a pixel with coordinates x_0, y_0 will involve only those pixels for whom $x \leq x_0$.

The second is that the TCN is able to accept an input with some length L and produce an output with exactly the same length. This aspect of the TCN leverages the technique of using ”fully convolutional” networks, which produce an output of the same size as the input. Thus at the core of the TCN architecture are 1D fully convolutional networks that use causal convolutions.

One remaining problem is that in traditional CNNs, the receptive field of each node in a convolutional layer is limited by the size of its associated kernel filter. This presents an issue for time-series tasks in which information must ideally be integrated across the entire length of the input. CNNs would typically solve this ”distance” problem using pooling operations, but this would jeopardise the ability of the TCN to produce an output with the same length as the input. The way that TCNs circumvent this issue is to use dilated causal convolutions. This means that in each successive layer of the TCN, the convolutional filters target pixels that are spaced increasingly further apart, even though the actual size of the filter remains the same. This approach allows the TCN to capture the entire history of the time-series input.

A TCN model can thus be characterised by a dilation vector, in which each successive element is the dilation factor for each successive convolutional layer - these would typically increase by a factor of 2.

One final feature of TCNs is the use of residual blocks developed to circumvent the problem of accuracy degradation that has been known to plague CNNs with a very high number of convolution layers. More specifically, it had been recognised that as network depth increases, the maximum accuracy achievable actually reaches a plateau and then decreases. The fact that training accuracy displays this trend reveals that the problem does not lie in over-fitting. Whatever the root of the problem, the general finding is that sometimes shallower networks can achieve better performance than their deeper counterparts, imposing a limit on the network complexity attainable before performance deteriorates.

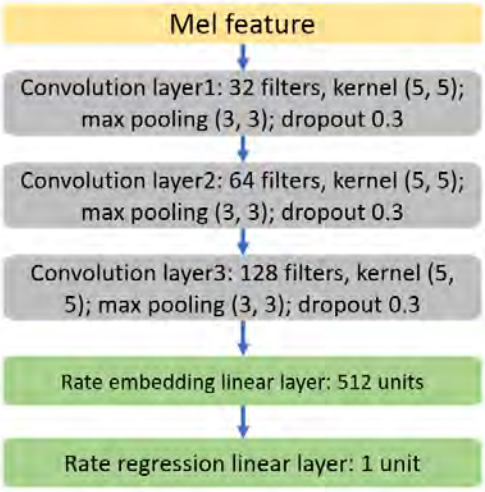
The input $\mathbf{x}$ is processed to yield some desired output $\mathcal{H}'(\mathbf{x})$ via a two steps. First, the input is processed through the convolutional layers to yield $\mathcal{F}(\mathbf{x})$, then the original input $\mathbf{x}$ is transmitted via a unit-weight skip connection to downstream of the convolutional layers, where it is then added to $\mathcal{F}(\mathbf{x})$ to form $\mathcal{H}(\mathbf{x})$, which is then passed through an activation function ϕ to yield $\mathcal{H}'(\mathbf{x})$. Hence, we have

$$\mathcal{H}'(\mathbf{x}) = \phi(\mathcal{H}(\mathbf{x})) = \phi(\mathcal{F}(\mathbf{x}) + \mathbf{x}), \tag{1}$$

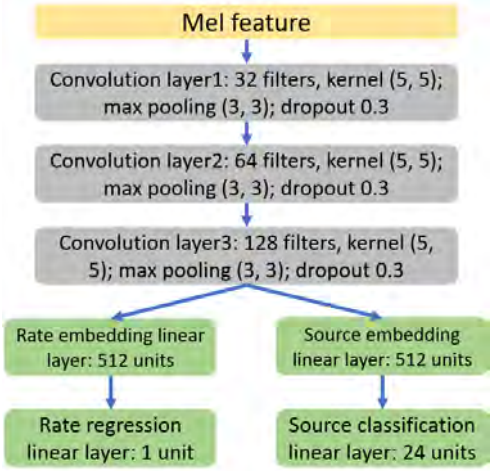
where $\mathcal{F}(\mathbf{x})$ can be seen to be the difference, or residual between the non-activated target output $\mathcal{H}(\mathbf{x})$ and the original input $\mathbf{x}$. For this reason, the skip connection is often called a ”residual connection” and the entire structure is often called a ”residual block”.

The main rationale behind the development of residual blocks was that for deep CNNs, it would be easier for the convolutional layers to learn the residual mapping $\mathcal{F}(\mathbf{x})$, compared to the original target mapping $\mathcal{H}(\mathbf{x})$. Presumably, this is due to the fact that the residual connections enable the model to simultaneously leverage the strengths of both shallow networks and deep networks.

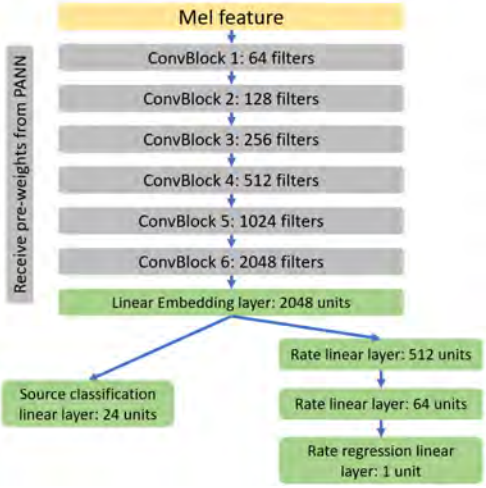
C Y0 through Y6 network structures



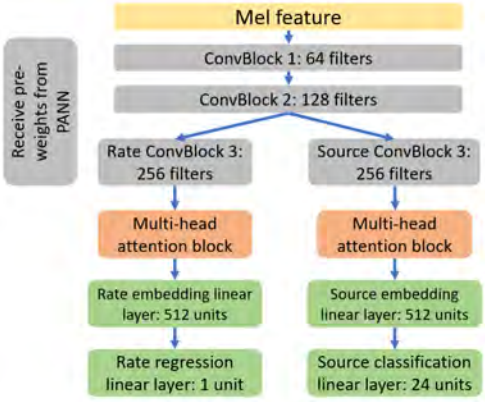
(a) **Y0**



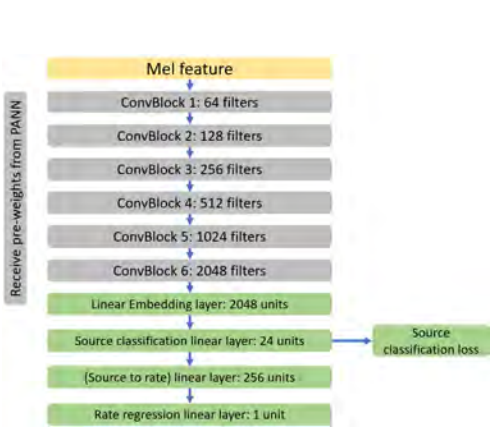
(b) **Y1**



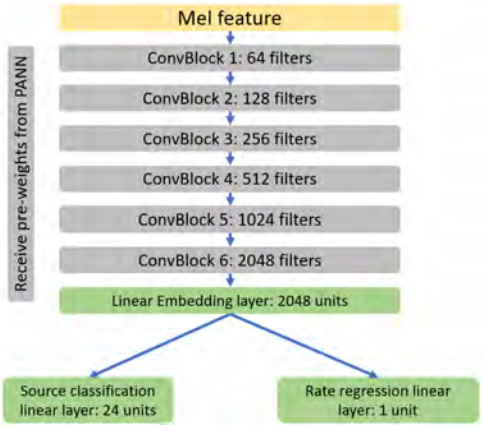
(c) **Y2**



(d) **Y3**


Mel feature
ConvBlock 1: 64 filters
ConvBlock 2: 128 filters
ConvBlock 3: 256 filters
ConvBlock 4: 512 filters
ConvBlock 5: 1024 filters
ConvBlock 6: 2048 filters
Linear Embedding layer: 2048 units
Source classification linear layer: 24 units
Source classification loss
(Source to rate) linear layer: 256 units
Rate regression linear layer: 1 unit
Receive pre-weights from PANN

(a) **Y4**


Mel feature
ConvBlock 1: 64 filters
ConvBlock 2: 128 filters
ConvBlock 3: 256 filters
ConvBlock 4: 512 filters
ConvBlock 5: 1024 filters
ConvBlock 6: 2048 filters
Linear Embedding layer: 2048 units
Source classification linear layer: 24 units
Rate regression linear layer: 1 unit
Receive pre-weights from PANN

(b) **Y6**

The
Alan Turing
Institute

turing.ac.uk

@turinginst