sustainability

MDPI

Article

Multilevel Annoyance Modelling of Short Environmental Sound Recordings

Ferran Orga [1,†], Andrew Mitchell [2,†], Marc Freixes [1], Francesco Aletta [2], Rosa Ma Alsina-Pagès [1,*] and Maria Foraster [3,4,5,6]

1 Grup de recerca en Tecnologies Mèdia, La Salle—URL, Quatre Camins, 30, 08022 Barcelona, Spain; ferran.orga@salle.url.edu (F.O.); marc.freixes@salle.url.edu (M.F.)
2 Institute for Environmental Design and Engineering, The Bartlett, University College London (UCL), Central House, 14 Upper Woburn Place, London WC1H 0NN, UK; andrew.mitchell.18@ucl.ac.uk (A.M.); f.aletta@ucl.ac.uk (F.A.)
3 ISGlobal, Parc de Recerca Biomèdica de Barcelona (PRBB), Doctor Aiguader, 88, 08003 Barcelona, Spain; maria.foraster@isglobal.org
4 Universitat Pompeu Fabra (UPF), 08018 Barcelona, Spain
5 CIBER Epidemiología y Salud Pública (CIBERESP), 28029 Madrid, Spain
6 PHAGEX Research Group, Blanquerna School of Health Science, Universitat Ramon Llull, 08025 Barcelona, Spain
* Correspondence: rosamaria.alsina@salle.url.edu; Tel.: +34-932-902-425
† These authors contributed equally to this work and share the first author position.

Abstract: The recent development and deployment of Wireless Acoustic Sensor Networks (WASN) present new ways to address urban acoustic challenges in a smart city context. A focus on improving quality of life forms the core of smart-city design paradigms and cannot be limited to simply measuring objective environmental factors, but should also consider the perceptual, psychological and health impacts on citizens. This study therefore makes use of short (1–2.7 s) recordings sourced from a WASN in Milan which were grouped into various environmental sound source types and given an annoyance rating via an online survey with $N = 100$ participants. A multilevel psychoacoustic model was found to achieve an overall $R^2 = 0.64$ which incorporates Sharpness as a fixed effect regardless of the sound source type and Roughness, Impulsiveness and Tonality as random effects whose coefficients vary depending on the sound source. These results present a promising step toward implementing an on-sensor annoyance model which incorporates psychoacoustic features and sound source type, and is ultimately not dependent on sound level.

Keywords: noise; annoyance evaluation; citizen; perceptive test; smart-city; annoyance modelling; wireless acoustic sensor network

Citation: Orga, F.; Mitchell, A.; Freixes, M.; Aletta, F.; Alsina-Pagès, R.M.; Foraster, M. Multilevel Annoyance Modelling of Short Environmental Sound Recordings. *Sustainability* **2021**, *13*, 5779. https://doi.org/10.3390/su13115779

Academic Editors: Cinzia Buratti, Juan Miguel Navarro and Jaume Segura-Garcia

Received: 22 April 2021
Accepted: 17 May 2021
Published: 21 May 2021

1. Introduction

Noise has been proven to have a wide impact on the social and economic aspects of citizens' lifes [1] and is regarded as one of the primary environmental health issues referenced in the new environmental noise guidelines [2]. Over the past few years, several research teams have analyzed the causes and the impact of this noise, revealing that it causes more than 48,000 new cases of ischemic heart disease and around 12,000 deaths in Europe each year [2]. Furthermore, it leads to chronic high annoyance for more than 22 million people, and sleep disturbance for more than 6.5 million people [3]. One of the main noise sources according to research is road traffic noise [4], causing psychological reactions in citizens [5] and even cardiovascular diseases [4]. Other studies analyze the effects of aircraft noise on sleep [6] and learning impairments on children [7]. Also railway noise has proven to cause annoyance due to its huge variety of sounds, e.g., rail breaks, whistles, squeels and vibrations [8,9]. Most of the literature focuses on sound level measurements and the corresponding annoyance [10], but other acoustical and psychoacoustical characteristics could be taken into account, e.g., loudness or sharpness [11],

in order to understand the degree of noise annoyance and identify the characteristics of sounds that may be more detrimental to psychological well-being and consequently for health. Such knowledge is relevant for policy makers and urban planners in order to create healthy environments.

Several tests used in studies to evaluate the effects of environmental noise for citizens [12] can be used to design this model. This study uses real-life data and its sound characterisation, thus focusing on noise sensitivity was not the closest approach to the problem. The tests used as a basis in this work have been defined with the purpose of finding new ways of analyzing the impact of sound -usually traffic- on citizens in urban environments [13,14], in order to model the annoyance perception [15,16].

The perceptual tests were designed to measure the annoyance in people relating to different urban sounds and their characteristics [17,18], by means of short excerpts of raw acoustic audio obtained from the DYNAMAP project [19]. The most representative audio excerpts were selected, using a wide range of sound types (sirens, airplanes, people talking, dogs barking, etc.) [20,21], keeping the constants of location and sensor calibration. However, sound annoyance depends on the acoustic characterization of each sample, and it is possible to classify the acoustic excerpts depending on their characterization, which can be the basis to ask participants about their perceptions. The characterisation is based on the psychoacoustic measurements of loudness, sharpness and others defined by Zwicker [11].

The authors asked more than 100 people to conduct the perceptual tests [18]. Some preliminary results of the three tests conducted were published in [17] in which the relationship between sharpness and annoyance was analyzed by means of an A/B test [22], and later on in [18], where some of the research questions were formulated. In this paper, we aim to determine the parameters that have an effect in the individual annoyance scores. For this reason, a multilevel psychoacoustic model is trained using the results of the MUSHRA [23] test, essentially focused on annoyance evaluation by the participants over several different types of sound, while loudness and sharpness were kept constant. The results show that the differences in annoyance perception between the different demographic groups is not statistically significant and that sharpness is the main predictor for annoyance.

The paper is structured as follows: Section 2 details the state-of-the-art of annoyance modelling by means of subjective data collection. Section 3 describes the procedure followed in this work, including the dataset and the design of the perceptual test. In Section 4, the results obtained from the perceptive tests are presented and discussed, and the annoyance model is proposed. Section 5 contains for the discussion and finally, Section 6 presents the conclusions of the paper.

2. State of the Art of Annoyance Evaluation and Modelling

In this section we gather a short synthesis of the most relevant contributions of the state-of-the-art on which the design of the tests and the modelling of the perceptual annoyance have been based.

2.1. Evaluation of Annoyance

The evaluation of annoyance can be found in literature by means of the use of objective parameters related to sound and noise [10]. Nevertheless, when the goal is to measure the perception, the real annoyance experienced by people, one of the most frequently used methods is to conduct a survey to measure the degree of annoyance produced by different sounds [24–26]. Following the recommendation of the International Committee for the Biological Effects of Noise (ICBEN), this evaluation should be done in a qualitative way, using a verbal scale; this can be translated into *not at all*, *slightly*, *moderately*, *very* and *extremely*, just to give a few examples. Also an 11-point numeric scale -also from an ICBEN recommendation- can be used, where in this case, zero corresponds to *not at all* and 10 corresponds to *extremely disturbing*.

Furthermore, taking advantage of the experience in soundscapes evaluation [27] citizens can be asked about other aspects besides annoyance. To this end, a perceptual

assessment based on a Likert scale [28] could be used. This scale defines five levels of agreement with a given statement: *Strongly disagree, Disagree, Neither agree nor disagree, Agree* and *Strongly agree*. This scale was used in [17,18] to evaluate several types of noise sources according to a small group of attributes such as *loud, shrill, noisy, disturbing, sharp, exciting, calming* and *pleasant* (see the complete list of adjectives in [27]).

Borrowing from the subjective assessment of audio quality, the MUSHRA method has been also used for the evaluation of annoyance in [17,18]. MUSHRA, which stands for *MUlti Stimulus test with Hidden Reference and Anchor*, was described and designed by ITU-R under the recommendation ITU-R BS.1534-3 [23]. This recommendation gives guidelines on listening tests and subjective assessment, as well as audio quality (among other applications), assuming that the best way to evaluate audio quality is by means of subjective listening.

Listening tests can be conducted in a controlled scenario (e.g., in an anechoic chamber) thus allowing the organizer to have control over all the setup. Nevertheless, this approach is expensive and time consuming. Alternatively, online listening tests have been widely used in the perceptual evaluation of audio quality or speech synthesis systems, even resorting to crowdsourcing strategies [29]. These tests can be run in parallel and anywhere, thereby reducing costs and allowing to reach a wider audience [30].

2.2. Annoyance Prediction

After the design and the execution of the perceptual tests, the resulting evaluations coming from participants are used to generate an model that can predict the annoyance value depending on the type and the parameters of the noise excerpt under study. One of the most representative examples of annoyance modelling is found in [15], where a model based on the hypothesis that annoyance is primarily determined by the detection of intruding sounds is presented. The model takes into account several measurable elements: *(i)* signal-to-noise ratio (SNR), *(ii)* indoor background level, *(iii)* the activity conducted by the listener—assuming that in the conducted tests, their main activity is not listening to events—among others. The model is obtained from the results of a test evaluating annoyance and acoustic data from a field experiment in a natural setting.

Another reference model for annoyance prediction is found in [16], where the authors model and predict road traffic-noise annoyance based on: *(i)* noise perception, *(ii)* noise exposure levels and *(iii)* demographics. The authors apply machine-learning algorithms in order to conduct the prediction and measure the error rates, which give them a good trade-off in the prediction of the traffic-noise annoyance, with a strong dependence on subjective noise perception and predicted noise exposure levels, assuming that the classical statistical approaches fail in their predictions in terms of accuracy.

A model of annoyance based on a combination of psychoacoustic metrics was proposed by Zwicker and Fastl [11]. Generated from laboratory-collected data, this model attempts to provide a method to directly calculate the relative annoyance values of single-source sounds from the psychoacoustic Loudness, Roughness, Sharpness, and Fluctuation Strength. This model has also been further expanded upon to include a term for the Tonality of the sound [31]. However, this model was developed based on laboratory studies of generated, simple sounds (i.e., not real recorded sounds) and does not take into account the semantic information associated with the real environmental sounds present in an urban environment.

In [32], the authors led us to a better understanding of the transportation noise-annoyance response, in three different and relevant approximations: *(i)* to unravel the factors that affect the annoyance response of people in reference to the mixed transportation noise, *(ii)* to contrast the noise-annoyance dependence in situations where road traffic and railway noise dominate and *(iii)* to detail the differences between those two using structural equation modelling. As expected, the results show that annoyance is largely determined by noise disturbance and the noisiness perceived by citizens. Finally, in [33] an approach to develop a road traffic noise annoyance prediction model is presented,

and it takes into account: *(i)* social aspects, *(ii)* characteristics of traffic and *(iii)* urban development. It is based on the creation of a local model, with a pilot in Istanbul (Turkey), which uses all the information gathered for the creation of the noise maps as an input, and provides annoyance levels prediction as an output, complementing the noise maps that provide no subjective indicator.

3. Methods

In this section we detail the several methods applied our experiment from the perceptual test design based on an urban sound dataset [21] to the multilevel linear regression modelling applied to obtain the annoyance prediction described as contribution in this paper.

3.1. Dataset

In order to obtain a proper representation of the acoustic environment in the design of the perceptual tests, a large quantity of recorded data is needed. The data gathered in this project belongs to different recording times and urban locations, using the Wireless Acoustic Sensor Network (WASN) deployed in Milan (Italy) in the framework of the LIFE DYNAMAP project [19,21].

Gathering the data through a WASN facilitates the collection of a wide and accurate representation of the acoustic events, because it keeps the same recording conditions in every node and allows the retrieval of data at any time of the day. The dataset used in this study has been obtained by homogeneously sampling several hours, in both weekday and weekend, with 24 sensors distributed along the urban District 9 of Milan [34]. After that, experts from the DYNAMAP developing team labelled the acoustic events of the recordings manually to obtain a 151-h dataset [21]. Due to the nature of the project, that consisted in removing events not related to traffic noise from the noise map computation, events were grouped in RTN (Road Traffic Noise) that belongs to the 83.7% of the total time of the dataset, and ANE (Anomalous Noise Event) with the 8.7% of the total time. Another class was used to include overlapping and unidentified events: COMPLX (complex) with 7.6% of the total time [20]. During the labelling process, the DYNAMAP developers found up to 26 types of anomalous events, which they decided to group in the following classes: airplane, alarm, bell, bike, bird, blind, brake, bus door, construction, dog, door, glass, horn, interference, music, people, rain, rubbish service, siren, squeak, step, thunder, tramway, train, trolley, wind, works (construction) [35].

The most common sound classes were picked to evaluate the relationship between the event measurements and the citizens' perception of annoyance. These selected events used in the study belong to the following 9 classes: airplane, bird, brake, construction, dog, door, horn, people and siren [36]. As the selected events are the most common, those are the ones that contain the widest variety of recording conditions, including different sensor locations and recording hours [17]. The reason for that choice was double: *(i)* the availability of a wide range of examples of each type of sound to choose for the design of the tests, including the possibility of finding different samples that keep similar psychoacoustic values, and *(ii)* the fact that the most common sounds are the most reasonable to evaluate with people, as they are the most probable to generate annoyance due to their repetitiveness.

The comparison between the events is only be carried with sounds collected using the same sensor, in order to respect the same recording conditions. For this reason, if the chosen events for the perceptive tests belong to a sensor or another, depends on the availability of the classes to be compared in each sensor. In all the cases, measures were taken to ensure that the sensor containing the events has enough variety of samples with variate psychoacoustic parameters, to ensure a proper representation of each category. To satisfy these requirements, only data from four sensors have been used to make the comparisons, as they provide enough information to carry the perceptual test, i.e., hb115, hb124, hb127 and hb133 [20]. More details about the event selection process and availability study of the sensors are detailed in [17], and the time of each event in the sensors is depicted in [18].

3.2. Design of the Perceptual Tests

In order to assess the degree of annoyance produced by the aforementioned classes of sounds, an on-line test has been conducted using the Web Audio Evaluation Tool [30]. Specifically, the MUSHRA test method [23]—which was originally designed for the evaluation of audio codecs—has been adapted for that purpose. Participants were given a clear explanation of what they were asked, including detailed instructions on the operation of the test. No training phase was therefore considered. A demographics survey was included at the beginning of the test for all the 100 participants, asking for to identify age, gender, and a subjective rating of the participant's residential area (zr1- very quiet, zr2- quiet, zr3- bit noisy, zr4- noisy, zr5- very noisy).

The second part of the test consists of five sets. Each set presents a group of short acoustic events with similar values of loudness and sharpness but from different classes, and recorded in the same sensor, in order to maintain the recording conditions and location of the sounds under comparison. For each set, the participants were asked to evaluate the annoyance produced by the presented audios, ordering them in a 0–10 scale, where zero corresponds to *not at all* and 10 corresponds to *extremely disturbing* following the ICBEN recommendation. The interface was customized including a color scale to help the participants place the stimuli according to the degree of annoyance that they perceive. Each audio is represented with a green bar with a "play" icon on it and the audios are sorted randomly along the MUSHRA scale (see Figure 1). An audio is reproduced when the corresponding bar is clicked. The system ensures the participant listens to all the audios and moves all the bars before they jump to the next set of audios. The sets were presented in a random order to prevent learning biases. MUSHRA tests usually include hidden reference stimuli, which in audio or speech quality evaluation corresponds to the highest quality samples and that are used to remove outlier responses. Nonetheless, since stimuli pertaining to different classes are compared, no audio reference was included, thus avoiding biases towards a certain audio class. Moreover, the participants were asked to take the test using headphones and to keep the same volume during all the tests, to maintain the same conditions throughout the entire testing process. One hundred participants undertok this test, 59 men and 41 women, with an average age of 33. Participants were volunteers, mainly from the university and also gathered via social networks. The distribution according to residential area is the following: 9 in zr1, 37 in zr2, 35 in zr3, 18 in zr4 and 1 in zr5. The MUSHRA test allows us to *(i)* obtain an individual score of annoyance for each audio and *(ii)* carry comparisons among the different types of events contained in a set. The detail of the stimuli included in each of the five sets of the test can be found in Table 1.

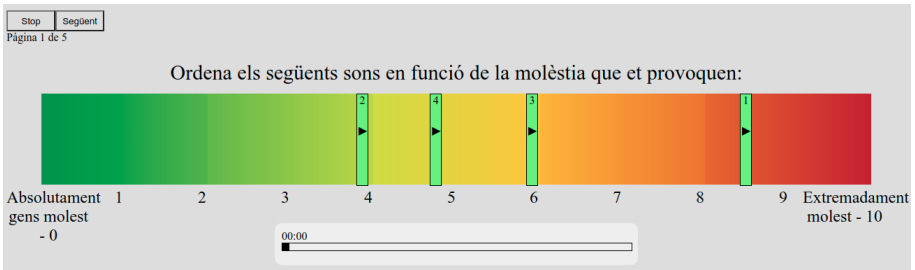

Figure 1. Screenshot of the MUSHRA test conducted to assess the annoyance provoked by different sounds. Title: sort the following sounds according to the caused annoyance. The scale ranges from *not annoying at all* to *extremely annoying*.

Table 1. Psychoacoustic parameters calculated for the 27 stimuli used in the listening experiment.

Sensor	Label	Psychoacoustic Parameters				
		Loudness (N_5 sone)	Sharpness (acum)	Roughness (asper)	Tonality (tuHMS)	Impulsiveness (iu)
hb133	peop	15.1	1.46	0.032	0.204	0.270
hb133	door	16.8	1.43	0.029	0.113	0.354
hb133	dog	13.1	1.22	0.033	0.373	0.266
hb133	brak	16.0	1.76	0.030	0.326	0.241
hb133	bird	12.6	1.73	0.024	0.283	0.214
hb133	airp	13.0	1.27	0.060	0.438	0.231
hb127	sire	17.7	1.56	0.045	1.540	0.178
hb127	peop	16.1	1.62	0.035	0.410	0.417
hb127	horn	18.1	1.56	0.028	0.666	0.260
hb127	door	19.8	1.72	0.037	0.037	0.479
hb127	brak	19.0	1.95	0.034	0.251	0.281
hb127	sire	20.1	1.73	0.046	1.670	0.288
hb127	peop	22.0	1.96	0.036	0.322	0.452
hb127	horn	19.9	2.16	0.034	1.290	0.336
hb127	brak	21.0	1.81	0.030	1.170	0.275
hb127	airp	24.4	1.65	0.056	0.172	0.446
hb115	wrks	20.3	1.97	0.054	0.227	0.267
hb115	trck	24.4	1.60	0.033	0.040	0.276
hb115	sire	19.5	1.46	0.054	0.861	0.333
hb115	peop	25.1	1.79	0.032	0.411	0.331
hb115	horn	22.3	2.00	0.032	0.806	0.155
hb115	door	26.3	1.62	0.038	0.045	0.397
hb115	brak	20.6	1.93	0.034	0.216	0.313
hb115	wrks	24.6	1.92	0.064	0.447	0.317
hb115	sire	26.6	1.77	0.044	0.626	0.290
hb115	horn	29.5	2.35	0.039	0.486	0.262
hb115	door	31.3	1.88	0.048	0.223	0.402

3.3. Psychoacoustic Data Analysis

The dataset resulted in 27 audio-recordings of identified sound events with durations ranging between 1.01 and 2.69 s. The calibrated audio files were imported in the ArtemiS Suite software (v. 11.5, HEAD acoustics GmbH) and the following psychoacoustic parameters were computed: *loudness*, *sharpness*, *roughness*, *tonality*, and *impulsiveness* [11]; values for these parameters are reported in Table 1. The rationale for selecting a relatively large set of psychoacoustic metrics is that they are often used as indicators to predict perceptual constructs (such as annoyance) in perceptual studies, as shown in recent soundscape literature [37,38]. Fluctuation Strength, which could otherwise be included in this list of psychoacoustic parameters as in Zwicker's annoyance model, was not included as the length of the recordings are too short to obtain a valid value. Loudness was calculated according to the DIN 45631/A1 standard for time-varying sounds, in a free-field [39]. As recommended by the standard, in order to avoid the under-estimation of evaluated loudness which is seen when using the arithmetic average of the loudness curve, the N_5 value (the 5% percentile value of the time-dependent loudness curve) is used as the single value of loudness. Sharpness was calculated according to DIN 45692, in a free-field [39]. With this sharpness method, the absolute loudness of the sound is not accounted for, so there should not be a duplication of information across the loudness and sharpness metrics. Roughness was calculated according to the hearing model by Sottek [40], with the option to skip the first 0.5 s in order to not distort the single value. Impulsiveness was also calculated

according to the hearing model by Sottek, with a 0.5 s skip interval. Finally, tonality was calculated according the ECMA-74 (17th edition), which is based on the hearing model of Sottek, with a frequency range of 20 Hz to 20 kHz [41].

3.4. Multi-Level Linear Regression Modelling

The analysis for this study utilizes multi-level linear regression modelling (MLM), with a random intercept and a random slope, using backward step feature selection. MLM's are commonly used in psychological research for repeated measures studies [42,43] and for applied prediction models [44,45]. Multi-level modelling allows for the incorporation of nested and non-nested group effects within the structure of the model, where the coefficients and intercepts for the independent variables are allowed to vary across groups. For this study, the data ~~is~~ are grouped into two non-nested sets to form a two-level model: by repeated measures per respondent ('user') and by sound type ('label'). In order to take into account the repeated measures across participants, and to correct for the participant's mean annoyance level, the 'user' variable is included in the second-level as a random intercept. We then include the psychoacoustic features as label effects, with coefficients which are allowed to vary across the sound type labels. The psychoacoustic features are also included as fixed effects in the first level, which do not vary across either the user or label groups.

The initial model structure, as written in Wilkinson-Rogers notation [46], is thus:

$$\begin{aligned} \text{annoyance} \sim \text{Loudness} + \text{Roughness} + \text{Sharpness} + \text{Tonality} + \text{Impulsiveness} \\ +(1 \mid \text{user}) + (1 + \text{Loudness} + \text{Roughness} + \text{Sharpness} + \text{Tonality} + \text{Impulsiveness} \mid \text{label}) \end{aligned} \quad (1)$$

Feature Selection

The MLM is initially fitted with all of the potential features included within both levels. In order to reduce the complexity of the model, a backwards step feature selection process is applied to both levels of the model. This process involves fitting the full model which includes all of the potential independent features (i.e., Equation (1)). The feature with the highest p-value (least significant) is then removed from the candidates and the model is refit. This process is repeated until all features meet the predefined significance threshold of $p < 0.05$. For a two-level model, first backward elimination of the second level is performed, followed by backward elimination of the first-level (or fixed) part.

If more than one feature is selected in the first-level, then the variance inflation factor (VIF) is calculated in order to check for multicollinearity, with a pre-determined threshold of $VIF < 5$. Any features which remain after the backwards stepwise selection and exceeded this threshold were investigated and removed if they were highly collinear with the other features. Once the feature selection process is completed, the final model with only significant features of interest included is fit and the table of the model coefficients is printed along with plots of the random effects and standardized estimates terms. Finally, quantile plots of the residuals and random effects are examined to confirm they are normally distributed [47].

The input and output features are z-scaled prior to the analysis and model building by subtracting the mean and dividing by the standard deviation in order to directly compare the coefficient values of independent variables measured on different scales [47]. The model fitting and feature selection was performed using the 'step' function from 'lmerTest' (v. 3.1.3) [48] in the R statistical software (v. 4.0.5) [49]. The summaries and plots were created using the 'sjPlot' package (v. 2.8.7) [50] and the multi-level R^2 values were calculated using 'MuMIn' (v. 1.43.17) [51].

4. Results

4.1. Differences in Annoyance between Groups

The average annoyance score of all users across all stimuli was $M = 0.58$ ($SD = 0.05$). Since some basic demographic information about the 100 participants of the perceptual test was known, it seemed logical to explore possible differences in annoyance scores

between different groups/levels of stratification of the sample, mostly for descriptive purposes. Therefore, Areas of residence and Gender were considered as factors in this analysis. Gender was treated as a binary variable (F/M), while Areas of residence was treated as a five-level categorical variable based on people's self-reported character of the area where they typically reside (range: 1–5; very quiet–very noisy). One-way repeated measures ANOVA was deemed to be the most appropriate approach to take into account the multiple responses that each of the 100 participants provided for the different recordings (N = 27). A first analysis was then conducted to determine whether there was a statistically significant difference in annoyance between Areas of residence: no statistically significant differences were observed in this case F (4.95) = 1.374, p = 0.249. Likewise, a second one-way repeated measure ANOVA was carried out to check whether statistically significant differences in annoyance existed between females and males: no statistically significant effect was observed in this case either F (1.98) = 0.714, p = 0.400. Such small differences between groups can indeed be observed in Figure 2.

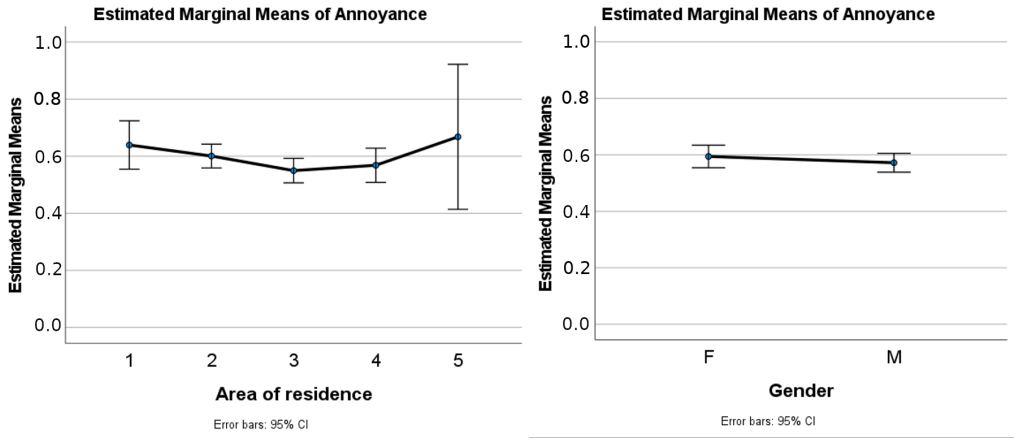

Figure 2. Estimated Marginal Means for Annoyance as a function of Areas of residence (**left**) and Gender (**right**).

4.2. Annoyance Model

The modelling process returned some interesting results about the parameters that have an effect in predicting the individual annoyance scores. In the context of the multi-level linear regression modelling, the included variables were assumed to have an effect at two levels: the first level (i.e., fixed effect(s)), and the second level, where annoyance score intercepts are allowed to vary as a function of users (i.e., the 100 participants), and where each feature of interest is allowed its own coefficient as a function of labels (i.e., the 7 types of sounds). Sharpness came up as the main predictor with a strong statistical significance in the fixed-effect level, as reported in Table 2. This implies that, regardless of any other factors, the sharper the sounds, the more annoying these are perceived to be.

The second-level effects presented in Figure 3 show that level- and loudness-based acoustic parameters do not play a significant role in predicting annoyance when considering other psychoacoustic factors and specific sound sources. The variables selected by the feature selection algorithm within the type of sound (label) level include: Impulsiveness, Roughness, Tonality. Among those, the effects of Impulsiveness, Tonality and type of sound are relatively small, while Roughness appears to be more important. For instance, when other effects are controlled, the sound type "horn" seems to be less annoying, the rougher it is; while for the types of sound "bird" and "siren", higher Roughness values will lead to higher annoyance scores. Looking at the model from the point of view of the types of sound, one could observe that "horns" tend to be more annoying than other sounds if they are more impulsive, while "people" or "birds" or "brakes" result in more annoying scores compared to other sounds if they tonal component is more prominent. Overall, for this model, the marginal and conditional R^2 values are 0.08 and 0.64, accordingly. Marginal

R^2 provides the variance explained by the fixed effects only, and conditional R^2 provides the variance explained by the whole model, i.e., both fixed effects and second-level effects.Thus, the majority of variance is explained by second-level factors, while a smaller portion (8%) is covered by Sharpness alone.

Table 2. Random intercept-random slope multi-level model of psychoacoustic annoyance, accounting for repeated measures (user) and sound source type (label) within the second level. Coefficients and confidence intervals given are for z-scaled data.

	Annoyance		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.02	−0.13–0.16	0.811
Sharpness	0.33	0.25–0.40	**<0.001**
Random Effects			
σ^2	0.47		
τ_{00user}	0.28		
$\tau_{00label}$	0.02		
ICC	0.39		
N_{user}	100		
N_{label}	10		
Observations	2700		
Marginal R^2/Conditional R^2	0.08/0.64		

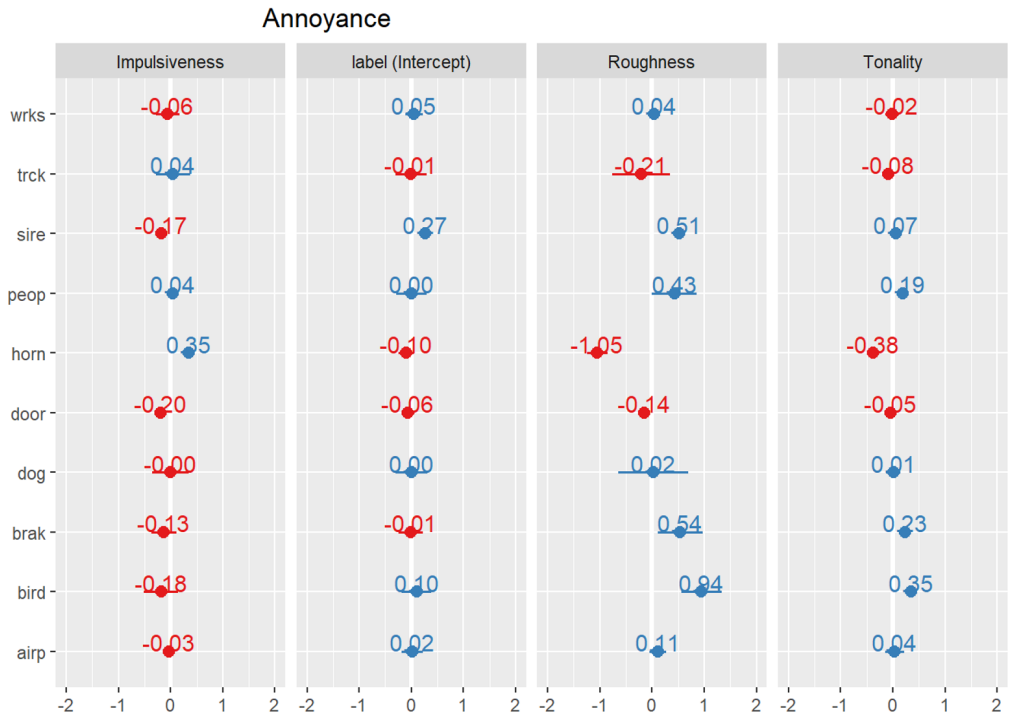

Figure 3. Second-level effects figures representing the regression coefficients by types of sound (label) and for different psychoacoustic parameters.

5. Discussion

Being able to predict noise annoyance from recorded sounds is particularly helpful from a public health perspective. In the context of a smart-city framework, one could imagine a wireless acoustic sensor network (WASN) large enough to cover a whole urban area; having a noise annoyance prediction algorithm at the node position that can return live annoyance scores to a central server from sounds recorded locally by the sensor would make for a useful

application for environmental protection officers and other stakeholders at community or local authority level [52]. A relevant issue to consider from the WASN perspective, is that previous studies conducted in both urban [21] and suburban [20] environments, there is a clear influence of the type of environment around the sensor location on the types of noise detected. Not all the urban or suburban locations for sensors have frequent sirens or horns, it depends on the more common activities (leisure, hospitals, etc.), the type of road (wide, narrow) and even the type of building or house existing in the surroundings, the types of noise detected in the street and their frequency of occurrence varies widely. In the design of a generalist model for quality of life, the number of occurrences, together with the duration and the annoyance caused by all and each noise source should be taken into account, so the former variables in cities and suburban environments is considered.

The fact that no significant differences in annoyance scores were observed between sample groups (i.e., gender or area of residence) is particularly interesting: it is common to assume in soundscape studies that personal and contextual factors play a strong role in how people respond to urban acoustic environments [53]. However, this is probably more relevant when complex sound environments (e.g., multi-source) are being considered and when dealing with relatively longer duration of exposures (e.g., several minutes) as seen in in-situ surveys. For clearly identifiable sources of environmental noise, with signals of short duration (i.e., 1–3 s) like those used for this experiment, it is likely it was easier for the sample to converge on similar annoyance scores, regardless of other demographic factors.

Regarding the noise annoyance scores, sharpness came up as an important predictor in the first level of the modelling stage (explaining up to 8% of the variance alone). It is important to highlight that the sharpness calculation method used in this study did not include any loudness correction; nor any loudness-related parameter was selected by the feature selection algorithm. To some extent, this is possibly due to fact that, being an online experiment, it was not possible for the research team to actually calibrate the loudness playback level accurately for the remote participants. On the other hand, considering this aspect from the WASN implementation perspective, this could be seen as an encouraging finding, since calibrating a diffuse acoustic monitoring network may not be practical in real-world scenarios, so it is good to have models that can achieve up to 64% of variance explained regardless of actual levels. Furthermore, in complex acoustic environments, loudness would likely vary over time depending on the relative positions between sound sources and (human) listeners in ways in which the other psychoacoustic parameters such as sharpness and tonality are less likely to. This is something that is impossible for fixed sensors to take into account, so once again it is preferable not to rely on loudness as a predictor.

6. Conclusions

In this study, an online listening experiment was conducted with 100 participants to assess the noise annoyance induced by short recordings of individual environmental noise sources gathered via a wireless acoustic sensors network in Milan. The main conclusions of this study are:

- the acoustic samples gathered from selected sensors in Milan WASN of the DYNAMAP project led us to a structured MUSHRA test to evaluate the annoyance in an off-line perceptual test;
- when considering short recordings of single-source environmental sounds, no significant differences in noise annoyance were observed as a function of demographic factors, such as gender and self-reported area of residence (i.e., from very quiet to very noisy);
- the multi-level linear regression model derived from this case study achieved an overall $R^2 = 0.64$, using sharpness as a fixed effect (the first level), and impulsiveness, roughness, tonality as random effects allowed to vary according to the type of sound (the second level) as predictors for perceived noise annoyance.

Taken together, the results of this study encourage us to continue our research work at the all the stages described in this paper. The improvement of the real-time algorithms to automatically detect the predefined sound sources under study is the first stage to gathering the most relevant samples in all and each of the sensors of a WASN. The application of the annoyance modelling can give the WASN a dimension without precedent; the availability of the objective acoustic measurements conducted by the sensors, and the estimated of annoyance in a real-time evaluation by means of the model. We can start to think about a dynamic annoyance map, which could be more far-reaching than a dynamic noise map.

Author Contributions: Conceptualization, F.A., R.M.A.-P., M.F. (Maria Forster); methodology, F.A., A.M., R.M.A.-P., M.F. (Maria Forster); software, F.O., M.F. (Marc Freixes), A.M.; investigation, M.F. (Marc Freixes), F.O., A.M., R.M.A.-P., F.A., M.F. (Maria Forster); data curation, F.O., A.M., F.A.; writing–review and editing, M.F. (Marc Freixes), F.O., A.M., R.M.A.-P., F.A., M.F. (Maria Forster); visualization, A.M., M.F. (Marc Freixes); supervision, M.F. (Maria Forster), F.A., R.M.A.-P.; project administration, F.A., R.M.A.-P.; funding acquisition, F.A., R.M.A.-P. All authors have read and agreed to the published version of the manuscript.

Funding: The UCL authors are funded by the European Research Council (ERC) under the European Union's Horizon 2020 research programme (grant agreement No. 740696). La Salle authors would like to thank Secretaria d'Universitats i Recerca from the Departament d'Empresa i Coneixement (Generalitat de Catalunya) and Universitat Ramon Llull, under the grant 2020-URL-Proj-054 (Rosa Ma Alsina-Pagès).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is available here: Ferran Orga, Marc Freixes, Rosa Ma. Alsina-Pagès, Alexandra Labairu-Trenchs. (2021). Audio dataset for perceptive studies in DYNAMAP project [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4775328, accessed on 22 April 2021.

Acknowledgments: The authors would like to thank to all the participants who volunteered to conduct the listening experiment.

Conflicts of Interest: The authors declare no conflict of interest. The Funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ANE	Anomalous Noise Event
ANOVA	Analysis of variance
ICBEN	International Committee for the Biological Effects of Noise
L_{eq}	Equivalent Level
MLM	Multi-level Linear regression Modelling
MUSHRA	MUlti Stimulus test with Hidden Reference and Anchor
RTN	Road Traffic Noise
VIF	Variance Inflation Factor
WASN	Wireless Acoustic Sensor Network

References

1. Goines, L.; Hagler, L. Noise Pollution: A Modern Plague. *South. Med. J. Birm. Ala.* **2007**, *100*, 287–294. [CrossRef]
2. World Health Organization. *Environmental Noise Guidelines for the European Region*; Technical Report; World Health Organization: Geneva, Swizterland, 2018.
3. Blanes, N.; Fons, J.; Houthuijs, D.; Swart, W.; de la Maza, M.; Ramos, M.; Castell, N.; van Kempen, E. Noise in Europe 2017: Updated Assessment. In *European Topic Centre on Air Pollution and Climate Change Mitigation (ETC/ACM)*; European Environment Agency: Bilthoven, The Netherlands, 2017.
4. Ndrepepa, A.; Twardella, D. Relationship between noise annoyance from road traffic noise and cardiovascular diseases: A meta-analysis. *Noise Health* **2011**, *13*, 251. [PubMed]
5. Ouis, D. Annoyance from road traffic noise: A review. *J. Environ. Psychol.* **2001**, *21*, 101–120. [CrossRef]

6. Basner, M.; Samel, A.; Isermann, U. Aircraft noise effects on sleep: Application of the results of a large polysomnographic field study. *J. Acoust. Soc. Am.* **2006**, *52*, 109–123. [CrossRef] [PubMed]
7. Hygge, S.; Evans, G.W.; Bullinger, M. A prospective study of some effects of aircraft noise on cognitive performance in schoolchildren. *Psychol. Sci.* **2002**, *13*, 469–474. [CrossRef] [PubMed]
8. Licitra, G.; Fredianelli, L.; Petri, D.; Vigotti, M.A. Annoyance evaluation due to overall railway noise and vibration in Pisa urban areas. *Sci. Total Environ.* **2016**, *568*, 1315–1325. [CrossRef] [PubMed]
9. Gidlöf-Gunnarsson, A.; Ögren, M.; Jerson, T.; Öhrström, E. Railway noise annoyance and the importance of number of trains, ground vibration, and building situational factors. *Noise Health* **2012**, *14*, 190. [CrossRef] [PubMed]
10. Berglund, B.; Lindvall, T.; Schwela, D.H. *Guidelines for Community Noise*; World Health Organization: Geneva, Switzerland, 1995.
11. Zwicker, E.; Fastl, H. *Psychoacoustics: Facts and Models*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 22.
12. Guski, R.; Schreckenberg, D.; Schuemer, R. WHO Environmental Noise Guidelines for the European Region: A Systematic Review on Environmental Noise and Annoyance. *Int. J. Environ. Res. Public Health* **2017**, *14*, 1539. [CrossRef]
13. Botteldooren, D.; Dekoninck, L.; Gillis, D. The influence of traffic noise on appreciation of the living quality of a neighborhood. *Int. J. Environ. Res. Public Health* **2011**, *8*, 777–798. [CrossRef]
14. Jakovljevic, B.; Paunovic, K.; Belojevic, G. Road-traffic noise and factors influencing noise annoyance in an urban population. *Environ. Int.* **2009**, *35*, 552–556. [CrossRef]
15. De Muer, T.; Botteldooren, D.; De Coensel, B.; Berglund, B.; Nilsson, M.; Lercher, P. A model for noise annoyance based on notice-events. In Proceedings of the 2005 International Congress and Exposition on Noise Control Engineering (Internoise 2005)-Paper 2033, Rio de Janeiro, Brazil, 7–10 August 2005.
16. Bravo-Moncayo, L.; Lucio-Naranjo, J.; Chávez, M.; Pavón-García, I.; Garzón, C. A machine learning approach for traffic-noise annoyance assessment. *Appl. Acoust.* **2019**, *156*, 262–270. [CrossRef]
17. Labairu-Trenchs, A.; Alsina-Pagès, R.M.; Orga, F.; Foraster, M. Noise Annoyance in Urban Life: The Citizen as a Key Point of the Directives. *Proceedings* **2019**, *6*, 1. [CrossRef]
18. Alsina-Pagès, R.M.; Freixes, M.; Orga, F.; Foraster, M.; Labairu-Trenchs, A. Perceptual evaluation of the citizen's acoustic environment from classic noise monitoring. *Cities Health* **2021**, *5*, 145–149. [CrossRef]
19. Sevillano, X.; Socoró, J.C.; Alías, F.; Bellucci, P.; Peruzzi, L.; Radaelli, S.; Coppi, P.; Nencini, L.; Cerniglia, A.; Bisceglie, A.; et al. DYNAMAP—Development of low cost sensors networks for real time noise mapping. *Noise Mapp.* **2016**, *3*, 172–189. [CrossRef]
20. Alías, F.; Orga, F.; Alsina-Pagès, R.M.; Socoró, J.C. Aggregate Impact of Anomalous Noise Events on the WASN-Based Computation of Road Traffic Noise Levels in Urban and Suburban Environments. *Sensors* **2020**, *20*, 609. [CrossRef] [PubMed]
21. Alías, F.; Socoró, J.C.; Alsina-Pagès, R.M. WASN-Based Day–Night Characterization of Urban Anomalous Noise Events in Narrow and Wide Streets. *Sensors* **2020**, *20*, 4760. [CrossRef] [PubMed]
22. Bech, S.; Zacharov, N. *Perceptual Audio Evaluation-Theory, Method and Application*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
23. Recommendation, I. *1534-1: Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*; International Telecommunication Union: Geneva, Switzerland, 2003.
24. Fields, J.; De Jong, R.; Gjestland, T.; Flindell, I.; Job, R.; Kurra, S.; Lercher, P.; Vallet, M.; Yano, T.; Research Team at Ruhr University; et al. Standardized general-purpose noise reaction questions for community noise surveys: Research and a recommendation. *J. Sound Vib.* **2001**, *242*, 641–679. [CrossRef]
25. Björk, J.; Ardö, J.; Stroh, E.; Lövkvist, H.; Östergren, P.O.; Albin, M. Road traffic noise in southern Sweden and its relation to annoyance, disturbance of daily activities and health. *Scand. J. Work. Environ. Health* **2006**, *32*, 392–401. [CrossRef]
26. Lim, C.; Kim, J.; Hong, J.; Lee, S. The relationship between railway noise and community annoyance in Korea. *J. Acoust. Soc. Am.* **2006**, *120*, 2037–2042. [CrossRef]
27. Lercher, P.; van Kamp, I.; von Lindern, E.; Botteldooren, D. Perceived Soundscapes and Health-Related Quality of Life, Context, Restoration, and Personal Characteristics: Case Studies. In *Soundscape and the Built Environment*; CRC Press: Boca Raton, FL, USA, 2016; [CrossRef]
28. Likert, R. A technique for the measurement of attitudes. *Arch. Psychol.* **1932**, *22*, 55.
29. Buchholz, S.; Latorre, J.; Yanagisawa, K. Crowdsourced Assessment of Speech Synthesis. In *Crowdsourcing for Speech Processing*; John Wiley & Sons, Ltd.: Oxford, UK, 2013; pp. 173–216. [CrossRef]
30. Jillings, N.; Moffat, D.; De Man, B.; Reiss, J.D. Web Audio Evaluation Tool: A browser-based listening test environment. In Proceedings of the 12th Sound and Music Computing Conference, Maynooth, Ireland, 26 July–1 August 2015.
31. Di, G.Q.; Chen, X.W.; Song, K.; Zhou, B.; Pei, C.M. Improvement of Zwicker's psychoacoustic annoyance model aiming at tonal noises. *Appl. Acoust.* **2016**, *105*, 164–170. [CrossRef]
32. Lam, K.C.; Chan, P.K.; Chan, T.C.; Au, W.H.; Hui, W.C. Annoyance response to mixed transportation noise in Hong Kong. *Appl. Acoust.* **2009**, *70*, 1–10. [CrossRef]
33. Ascigil-Dincer, M.; Demirkale, S.Y. Model development for traffic noise annoyance prediction. *Appl. Acoust.* **2021**, *177*, 107909. [CrossRef]
34. Alsina-Pagès, R.M.; Alías, F.; Socoró, J.C.; Orga, F. Detection of Anomalous Noise Events on Low-Capacity Acoustic Nodes for Dynamic Road Traffic Noise Mapping within an Hybrid WASN. *Sensors* **2018**, *18*, 1272. [CrossRef] [PubMed]
35. Alías, F.; Socoró, J.C. Description of anomalous noise events for reliable dynamic traffic noise mapping in real-life urban and suburban soundscapes. *Appl. Sci.* **2017**, *7*, 146. [CrossRef]

36. Orga, F.; Alías, F.; Alsina-Pagès, R.M. On the Impact of Anomalous Noise Events on Road Traffic Noise Mapping in Urban and Suburban Environments. *Int. J. Environ. Res. Public Health* **2017**, *15*, 13. [CrossRef] [PubMed]
37. Aletta, F.; Kang, J.; Axelsson, Ö. Soundscape descriptors and a conceptual framework for developing predictive soundscape models. *Landsc. Urban Plan.* **2016**, *149*, 65–74. [CrossRef]
38. Aletta, F.; Axelsson, Ö.; Kang, J. Dimensions underlying the perceived similarity of acoustic environments. *Front. Psychol.* **2017**, *8*, 1162. [CrossRef] [PubMed]
39. Calculation of Loudness Level and Loudness from the Sound Spectrum—Zwicker Method—Amendment 1: Calculation of the Loudness of Time-Variant Sound. 17.140.01—Acoustic Measurements and Noise Abatement in General. Available online:https://www.iso.org/obp/ui/#!iso:std:63078:en (accessed on 4 April 2021).
40. Sottek, R. Sound quality evaluation of noises with spectro-temporal patterns. In Proceedings of the INTER-NOISE and NOISE-CON Congress and Conference Proceedings, Hong Kong, China, 27–30 August 2017; Volume 255, pp. 3927–3936.
41. *ECMA-74 Measurement of Airborne Noise Emitted by Information Technology and Telecommunications Equipment*, 17th ed.; December 2019. Available online: https://www.ecma-international.org/publications-and-standards/standards/ecma-74/ (accessed on 4 April 2021).
42. Quené, H.; van den Bergh, H. On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Commun.* **2004**, *43*, 103–121. [CrossRef]
43. Volpert-Esmond, H.I.; Page-Gould, E.; Bartholow, B.D. Using multilevel models for the analysis of event-related potentials. *Int. J. Psychophysiol.* **2021**, *162*, 145–156. [CrossRef]
44. Gelman, A. Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics* **2006**, *48*, 432–435. [CrossRef]
45. Frees, E.W.; Kim, J.S. Multilevel Model Prediction. *Psychometrika* **2006**, *71*, 79–104. [CrossRef]
46. Wilkinson, G.N.; Rogers, C.E. Symbolic Description of Factorial Models for Analysis of Variance. *J. R. Stat. Soc. Ser. Appl. Stat.* **1973**, *22*, 392–399. [CrossRef]
47. Harrison, X.A.; Donaldson, L.; Correa-Cano, M.E.; Evans, J.; Fisher, D.N.; Goodwin, C.E.; Robinson, B.S.; Hodgson, D.J.; Inger, R. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* **2018**, *6*, e4794. [CrossRef]
48. Kuznetsova, A.; Brockhoff, P.B.; Christensen, R.H.B. lmerTest Package: Tests in Linear Mixed Effects Models. *J. Stat. Softw.* **2017**, *82*, 1–26. [CrossRef]
49. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
50. Lüdecke, D. sjPlot: Data Visualization for Statistics in Social Science. R Package Version 2.8.7. 2021. Available online: https://strengejacke.github.io/sjPlot/ (accessed on 20 April 2021).
51. Barton, K. MuMIn: Multi-Model Inference. R Package Version 1.43.17. 2020. Available online: https://r-forge.r-project.org/R/?group_id=346 (accessed on 20 April 2021).
52. Kang, J.; Aletta, F.; Margaritis, E.; Yang, M. A model for implementing soundscape maps in smart cities. *Noise Mapp.* **2018**, *5*, 46–59. [CrossRef]
53. Kang, J.; Aletta, F.; Gjestland, T.T.; Brown, L.A.; Botteldooren, D.; Schulte-Fortkamp, B.; Lercher, P.; van Kamp, I.; Genuit, K.; Fiebig, A.; et al. Ten questions on the soundscapes of the built environment. *Build. Environ.* **2016**, *108*, 284–294. [CrossRef]