DOI: 10.61782/fa.2025.0235

FORUM ACUSTICUM
EURONOISE 2025

Comparison of Deep Learning and Psychoacoustic Models to Predict UAV Noise Impact in Soundscapes

Max W. Ellis[1*] **Marc C. Green**[1] **Michael J. B. Lotinga**[1]
Antonio J. Torija Martínez[1]

[1] Acoustics Research Centre, University of Salford, England (UK)

ABSTRACT

The increasing prevalence of Unmanned Aircraft Systems in urban environments necessitates a deeper understanding of their impact on the experience of urban soundscapes. This study presents Machine Learning models aimed at predicting perceived annoyance of UAS noise. Deep learning models were generated using convolutional recurrent neural networks, trained on a dataset incorporating data from multiple listening experiment. The model predictions are compared with various existing nonlinear models for Psychoacoustic Annoyance. Our expanded dataset includes recent field studies across England and Greece, enhancing the robustness and generalisability of our models. The broader aim of this research is development of a comprehensive soundscape model for UAS noise, which could be incorporated into future 'next generation' smart sound level meters and be used to inform urban planning decisions.

Keywords: *Artificial Intelligence, Machine Listening, Psychoacoustic Annoyance, Soundscape, UAS*

1. INTRODUCTION

Noise signatures from emergent technologies in both civil and commercial domains have prompted growing concern regarding their acoustic impact on the perception of urban soundscapes. As frameworks develop for UAS, commonly referred to as *drones,* to integrate into transport infrastructure, small drones are already employed in agriculture for crop sowing, health monitoring, and logistics for both NHS and Royal Mail deliveries. These drones elicit complex perceptual responses by introducing novel, often highly tonal noise signatures that deviate from broadband noise typical of road or rail transport. Conventional noise assessment methods – historically reliant on A-weighted scales – are insufficient in capturing these effects.

Recent years have seen increased emphasis on applying the soundscape approach [1], acknowledging human perception of environmental sound is shaped not only by acoustic levels but also by context and emotional affect. This shift has encouraged the development of methods that predict perceived affect directly from audio features, whether through psychoacoustic modelling or data-driven approaches such as deep learning. However, while psychoacoustic metrics have been widely used in industrial noise prediction, their application to UAS noise, particularly in complex or multi-source sound environments, remains under investigation.

Our previous work using convolutional neural networks (CNNs) to predict perceived annoyance ratings from mel-spectrograms indicated that deep learning may offer enhanced predictive capabilities for UAS noise perceived annoyance [2]. This study introduces a new convolutional recurrent neural network (CRNN) architecture and three new soundwalk datasets (on top of the previous CNN model and 5 datasets). The objective is to determine which model architectures best predict subjective UAS noise annoyance ratings.

2. BACKGROUND

2.1 Soundscapes and Contextual Perception

The concept of soundscapes, introduced by R. M. Schafer [3], describes how sound interacts with environments and shapes human perception. ISO-12913 [1] formalised this

[*] ***Corresponding author***: *M.Ellis6@edu.salford.ac.uk*

framework, defining soundscapes as "the acoustic environment as perceived or experienced and/or understood by a person or people, in context". A significant advancement in soundscape research is the Circumplex Model [4], which organises soundscapes along dimensions of pleasantness and eventfulness, varying combinations of which can yield soundscapes described as calm, exciting, monotonous, and chaotic. Despite these advances, the influence of *auditory salience* – the prominence of a sound object – remains understudied [5], particularly in relation to electric vehicles (EV's) and UAVs. As next-generation technologies introduce novel noise sources, often containing high-frequency tonal components, their interaction with existing soundscapes can vary dramatically depending on listener expectations, use case, emotional factors and context [6].

2.2 Psychoacoustic Modelling and Annoyance Frameworks

Psychoacoustic annoyance refers to a modelled estimate derived from quantifiable sound quality metrics such as loudness, sharpness, fluctuation strength, and tonality. In contrast, perceived annoyance (PA) is a subjective judgement made by listeners, typically gathered through listening tests, and reflects the complex interplay of acoustic, contextual, and individual factors not fully captured by existing models. Widmann's annoyance model, originally developed for traditional noise sources [7], has required significant adaptations for UAS noise profiles. Di *et al.* [8] enhanced the model by integrating tonality penalties that account for an additional 0.5 – 1.5 units of annoyance depending on tonal strength and frequency, and an adjusted sharpness rating that is approximately 20% higher than ISO and DIN standards [9]. This adaptation improved prediction accuracy with a reduction in Root Mean Squared Error (RMSE) from 1.27 to 0.89, though limitations remain when analysing complex environments with non-tonal noise characteristics. Ramos-Romero *et al.* (2024) [10] developed a comprehensive taxonomy for assessing UAS noise, examining relationships between design parameters, operational conditions, and psychoacoustic metrics. Their measurements revealed that multirotor UAS produced noise with higher sharpness and tonal components compared to fixed-wing designs. UAS noise typically ranges between 500 Hz – 5 kHz, with tonal peaks in the 1 kHz – 2 kHz range contributing significantly to perceived annoyance. Contextual variations were notable, with sharpness levels measuring 1.9 acums in urban settings versus 1.5 acums in rural areas. Lotinga *et al.* (2023) [11] investigated psychoacoustic metrics for potential integration into UAS noise regulations. Smaller UAVs (e.g. DJI Matrice 300) exhibited higher levels of sharpness and roughness, while larger aircraft produced greater tonal fluctuation strength. Tonal components between 200 Hz – 800 Hz were found to significantly increase annoyance, with sharpness levels exceeding 1.5 acum, correlating with a 25% increase in reported annoyance even when overall sound levels remained constant. Subsequent research from Lotinga [12] found that the adapted Torija et al. [13] PA model demonstrated superior performance to Widmann, More [14] and Di et al. The Sottek Hearing Model [15] offers an advanced framework that simulates human auditory perception by modeling the basilar membrane's response to sound. This physiological approach accounts for nonlinear auditory filtering and is particularly effective for analysing drone noise in varying flight conditions, due to its increased precision in quantifying slow amplitude modulation. While psychoacoustic annoyance models offer interpretable predictions based on engineered features, they are limited by their reliance on fixed-weight formulations and assumptions derived from traditional noise sources. These models often fail to account for contextual variation, spectral complexity, and perceptual non-linearity present in UAV noise. A deep learning approach could learn context-sensitive acoustic representations directly from data, across varying environments and sound source types, perhaps resulting in more accurate predictions.

2.3 Deep Learning Models for Noise Analysis and Prediction

Green and Torija [2] demonstrated that CNNs, an architecture ideal for processing gridlike data such as audio spectrograms, achieves 85% accuracy in predicting perceived annoyance from UAV noise using mel-spectrograms, substantially outperforming traditional regression models that reach only 70% accuracy. The optimal time-frequency resolution parameters were found to be a Fast Fourier-Transform (FFT) length of 256 and hop length of 64 samples, yielding a Mean Absolute Error (MAE) of 0.58 and an R^2 value of 0.72. This performance exceeded traditional PA metrics that predominantly emphasise loudness, suggesting CNNs can identify additional perceptually relevant features such as fluctuation strength and roughness. However, challenges remain in generalising these models across varied environments and UAS types due to differences in noise profiles based on altitude, speed, and environmental conditions. For sequential audio analysis, Recurrent Neural Networks (RNNs) have proven effective in capturing temporal patterns. Grekow's

research on music emotion recognition demonstrated RNNs achieving 83% accuracy in predicting emotional responses along arousal and valence dimensions [16], analogous to the affective eventfulness and pleasantness dimensions specified in ISO 12913. The model showed stronger performance in predicting arousal (correlation coefficient of 0.72) than valence (correlation coefficient of 0.55), using features including tempo, loudness, pitch, and timbre across a dataset of 5,000 music tracks. Addressing the complexity of polyphonic soundscapes, Çakır *et al.* [17] proposed a CRNN utilising convolutional layers to extract spectral features while recurrent layers model temporal dependencies. This resulted in a 10 – 15% improvement in F1 scores compared to standard CNN and RNN models when applied to overlapping urban sounds. This architecture is particularly suitable for UAV noise detection within complex soundscapes, though computational demands may challenge realtime deployment. Casabianca and Zhang [18] introduced a late fusion ensemble approach for acoustic UAV detection, combining multiple deep neural networks through both hard and weighted soft voting strategies. Their ensemble, comprising CNNs trained on mel-spectrograms, achieved up to 94.7% accuracy on unseen augmented datasets, significantly outperforming individual models. The study also highlighted the role of data augmentation in improving scalability across drone types and recording conditions.

3. METHODS

3.1 Datasets

This study retains the five datasets introduced in Green and Torija [2], comprising a total of 587 audio clips with associated annoyance ratings. These were recorded across various controlled listening studies involving both conventional and UAV aircraft, with annotations collected using continuous ratings (KTH dataset) rescaled to match. To extend the data and explore the applicability of existing models, in this work, three new datasets are introduced, collected in: (i) Crescent Meadow, Salford, UK (23 clips) (Green and Torija, 2024) [19]; (ii) Athens, Greece (56 clips) (Green *et al.*, 2025) [20]; (iii) the Isles of Scilly, UK (27 clips) (Green and Torija, 2025) [21], amounting to an increased total of 693 clips. These datasets are distinct from those included in the previous study in that they were collected in field-based soundwalk studies, rather than in a lab-based setting. Participants were exposed to take-off, flyover, and landing UAS maneuvers, conducted at a range of altitudes, and responses were gathered along soundscape dimensions specified in ISO 12913-2, as well as annoyance based on ISO 15666 [22]. The latter were used as target ratings in the present study. Baseline annoyance ratings were gathered from soundwalk locations prior to the commencement of drone operations, and the average annoyance of all respondents at each stop/location were assigned to audio clip names from all eight datasets. Audio recordings were made at each stop using the Zoom H3VR portable Ambisonic recorder[1], with the omnidirectional channel used to derive monaural stimuli for the present study. These soundwalks were conducted as part of a series with the specific aim of investigating the perception of UAS noise introduced within a variety of existing environmental contexts, including a busy city street, quieter park and meadow, and a remote rural island. The extracted clips include both ambience and UAS noise. In line with the original methodology, all audio clips were resampled and trimmed to the loudest six-second segment based on total spectrogram amplitude. Mel-spectrograms were then generated for each clip up to a frequency of 8 kHz, represented by 96 mel-spaced bins with a total dynamic range of 60 dB.

Table 1. Summary of included datasets

Dataset	Participants	Clips	Response Type
MJL	42	80	Single
NG23	41	71	Single
NG22	30	51	Single
RN	50	120	Single
AJ	25	9	Continuous
Meadow	16	24	Single
Scilly	22	27	Single
Athens	110	56	Single

[1] https://www.zoom-europe.com/en/handy-recorders/zoom-h3-vr

3.2 Psychoacoustic Annoyance Models

Sound quality metrics required for Widmann's psychoacoustic annoyance model were derived from the raw audio files using the SQAT library in MATLAB [19]. Pearsons *R* was used to quantify correlations between the Mean Perceived Annoyance's (MPA) and the sound quality metrics. Scatter plots showing correlations between perceived and psychoacoustic annoyance, as well as those with individual SQMs, are shown in Figure 1.

3.3 Deep Learning Models

Two deep neural network architectures were developed to predict perceived annoyance directly from audio spectrograms: a CNN and a CRNN. Similarly to the previous study [2], input spectrograms were generated using the *torchaudio* library [20], with particular attention to time-frequency resolution parameters which were previously found to significantly impact model performance. The CNN architecture consists of two convolutional layers with 3×3 kernels outputting 96 and 32 channels respectively, each followed by 2×8 max pooling and batch normalisation. These feed into three fully-connected layers of 1000, 100, and 1 output units with ReLU activation functions and dropout regularisation (p=0.2) between all layers except the final output. The CRNN extends this architecture by incorporating recurrent processing after the convolutional feature extraction. It utilises the same initial convolutional layers but then reshapes the output for sequential processing through two bidirectional LSTM layers with 128 and 64 hidden units respectively. An attention mechanism focuses on the most relevant temporal features before passing through fully-connected layers of 512 and 128 units, then finally one single unit. The CRNN employs higher dropout (p=0.3) and gradient clipping at a maximum norm of 5.0 to stabilise training. Through systematic grid search of FFT lengths {512, 256, 128, 64} and hop lengths {256, 128, 64, 32}.

3.4 Training & Testing

The models were implemented using PyTorch [21] and trained with MSE loss and Adam optimiser (learning rate 1×10^{-3} for CNN, 5×10^{-4} for CRNN). Early stopping with 40 epochs patience prevented overfitting across a maximum of 200 epochs with batch size 16. The dataset was randomly split into training (75%, n=520), validation (13%, n=91), and test (12%, n=83) sets, with shuffling to ensure even distribution across data sources. Performance was evaluated using MAE and coefficient of determination (R^2).

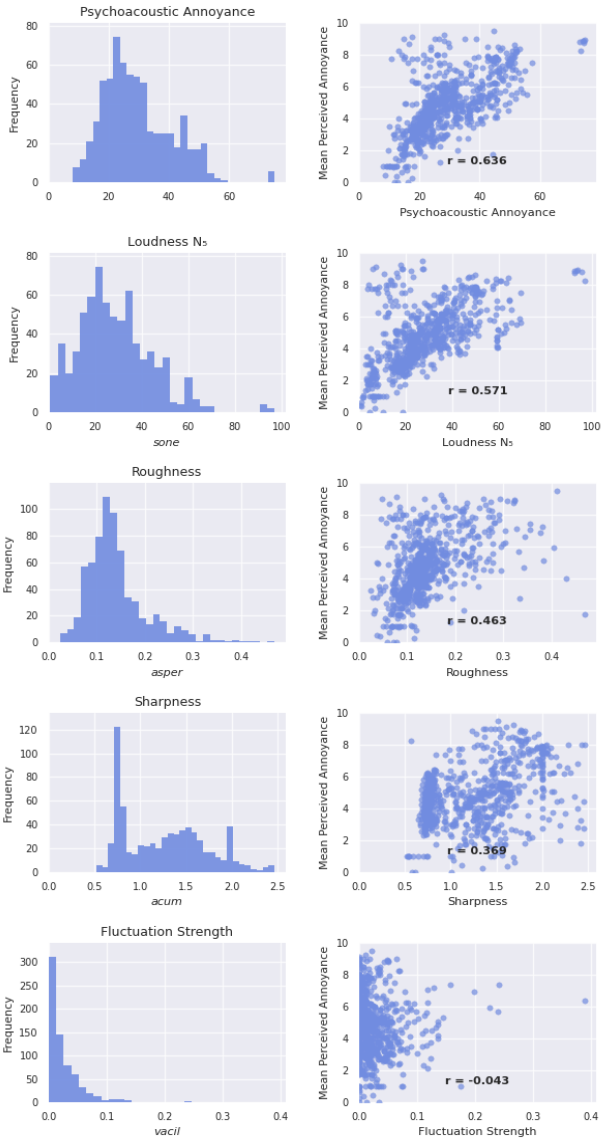

Figure 1. Histograms of psychoacoustic features and correlations of these to perceived annoyance.

Regularisation techniques included weight decay (1×10^{-5}) and gradient clipping at a maximum norm of 5.0 to prevent overfitting and stabilise training. Computation time per epoch ranged from 70 to 120 seconds, with training conducted on a Mac Mini M2 to utilise Metal Performance Shaders (MPS) backend for GPU training acceleration [22]. For comparison, two Support Vector Regression (SVR) baseline models were implemented: one with sigmoid kernel fitted directly to Widmann psychoacoustic annoyance values, and another with polynomial kernel (degree 3)

trained on the raw sound quality metrics.

4. RESULTS AND DISCUSSION

The SVR model trained on PA values using a sigmoid kernel achieved an MAE of 0.87 and an R^2 of 0.49, while the polynomial SVR trained on raw psychoacoustic features yielded a marginally better performance, with an MAE of 0.77 and R^2 of 0.56. As seen in the previous study, this suggests that a model trained directly on the constituent features of PA can outperform the composite metric itself. The sigmoid curve fit, shown in Figure 2, follows a similar logistic shape to that reported in earlier work, with perceived annoyance values beginning to plateau at higher PA levels. A notable concentration of moderate-to-high annoyance ratings (7-10 range) was identified, predominantly originating from the Athens dataset. This clustering suggests potential contextual influences beyond the immediate acoustic stimuli. Qualitative observations during the soundwalk revealed localised environmental factors that may have significantly influenced participant responses. Specifically, the Athens dataset exhibited a marked tendency toward elevated annoyance ratings, potentially attributable to pre-existing environmental disturbances independent of the specific acoustic stimulus. Despite these potential outliers, the SVR model maintained robust predictive performance, demonstrating remarkable resilience to dataset heterogeneity.

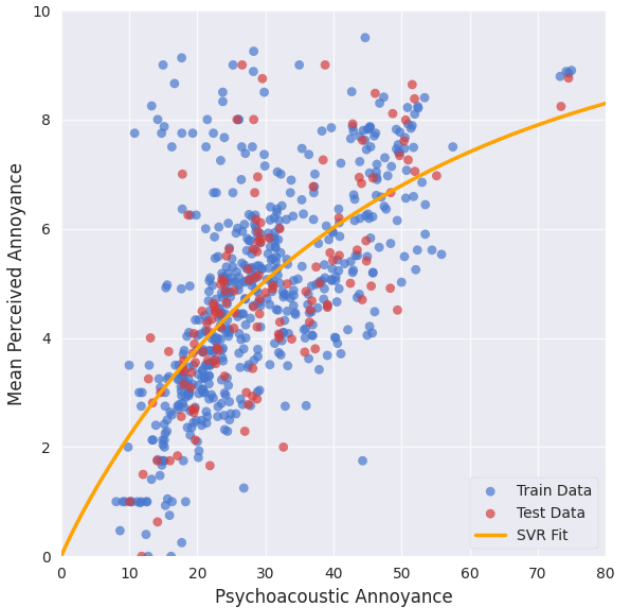

Figure 2. Fit of perceived annoyance to Widmann psychacoustic annoyance values by SVR model.

Table 2. Results for each network architecture and frame/hop length combination.

Frame / Hop	CNN		CRNN	
	MAE	R^2	MAE	R^2
512 / 256	0.77	0.55	0.65	0.66
512 / 128	0.84	0.29	0.62	0.7
512 / 64	**0.56**	**0.75**	0.55	0.76
512 / 32	0.85	0.54	0.7	0.78
256 / 128	0.76	0.62	0.56	0.81
256 / 64	0.64	0.71	**0.49**	**0.86**
256 / 32	0.87	0.39	0.83	0.46
128 / 64	0.72	0.56	0.58	0.79
128 / 32	0.81	0.46	0.71	0.63
64 / 32	0.77	0.5	0.94	0.44

Table 2 shows that CNN models exhibited varied performance across different time-frequency resolutions. The optimal configuration was identified as an FFT length of 512 samples and hop length of 64 samples, achieving MAE 0.56 / R^2 0.75. This configuration demonstrated a robust balance between spectral and temporal feature extraction. Configurations with shorter hop lengths (32 samples) consistently underperformed, with MAE values ranging from 0.85 to 0.87 and lower R^2 values. The 256 / 64 configuration also showed strong performance, with MAE 0.64 / R^2 0.71.

The CRNN models demonstrated superior performance compared to the CNN models. The 256 / 64 configuration in particular emerged as the standout, with an impressive MAE 0.49 / R^2 0.86, indicating a significant improvement in predictive accuracy over the CNN. The recurrent architecture showed more consistent performance across different FFT and hop length configurations. The 512 / 64 configuration also performed well, with MAE 0.55 / R^2 0.76. Configurations with 32-sample hop lengths showed decreased performance, with the 512 / 32 configuration achieving an R^2 of 0.78 but other configurations showing less consistent results.

The lower MAE and higher R^2 values of the CRNN results suggest that the bidirectional LSTM layers and attention mechanism provide additional context for predicting perceived annoyance. All models achieved their best ratings with hop 64 and FFT 256 or 512, aligning with the previous hypothesis about capturing perceptually relevant acoustic fluctuations. This configuration provides a 4ms time resolution, potentially capturing psychoacoustic features analogous to roughness and temporal modulation.

5. CONCLUSIONS AND FUTURE WORK

Several promising avenues for future research emerge from this study. First, investigating transfer learning techniques could potentially enhance model generalisability. The incorporation of additional datasets from various acoustic environments might improve model robustness and predictive capabilities. Expanding the research to other sound sources beyond UAS could establish a more comprehensive framework for sound affect prediction. As tonal, high-frequency and synthetic novel noises from e-mobility proliferate into urban environments, it would be invaluable to expand on the current models to classify sounds in noise clips and apply weightings based on source type, detectability and intermittency to better measure more eventful soundscapes. Finally, exploring objective ratings of annoyance from psychoacoustic models from Di, More, Torija *et al*. and NASA, along with their relative metrics, could help better calculate both subjective target annoyances and objective annoyance-model based ratings. This could, in theory, lead to a framework for open-source sound libraries to become extremely useful and adaptable to noise annoyance prediction, classification, training and learning.

6. ACKNOWLEDGMENTS

The authors would like to sincerely thank Abertay University, Dundee, for the support and supervision of work carried out during Max W. Ellis' postgraduate study, where the research framework, spectrograms, and data processing of this study were developed.

7. REFERENCES

[1] ISO, 2014. *ISO 12913-1:2014. Acoustics – Soundscape – Part 1: Definition and conceptual framework.* Geneva: ISO.ISO, 2014. ISO 12913-1:2014. Acoustics – Soundscape – Part 1: Definition and conceptual framework. Geneva: ISO.

[2] Green, M. C. and Torija, A. J., 2024. *Soundwalking in Salford: A Soundscape Approach to Drone Noise Assessment.* Quiet Drones, Manchester.

[3] Schafer, R. M., 1993. *The Soundscape: Our Sonic Environment and the Tuning of the World.* Rochester, VT: Destiny Books.

[4] Axelsson, Ö., Nilsson, M.E. & Berglund, B. (2010) 'A principal components model of soundscape perception', The Journal of the Acoustical Society of America, 128(5), 2836–2846.

[5] Podwinska, Z., Fazenda, B.M. & Davies, W.J. (2019). Testing spatial aspects of auditory salience. Proc. of the 25th International Conf. on Auditory Display (ICAD).

[6] Fang, X., Aletta, F., Mitchell, A., Oberman, T., and Kang, J., 2024. *Determining factors for the appropriateness of soundscapes: A cross-sectional large-sample study in London (UK). J. Acoust. Soc. Am.*, 156(5), pp.3588–3607.

[7] Zwicker, E. and Fastl, H., 2013. Psychoacoustics: Facts and Models. 3rd ed. Berlin: Springer.

[8] Di, Z., Shi, L., Chen, Y., Li, X. and Zhang, M., 2022. Annoyance prediction of substation noise based on the modified psychoacoustic annoyance model. Appl. Acoust., 188, p.108594.

[9] DIN, 2010. *DIN 45631/A1: Calculation of loudness level and loudness from the sound spectrum – Zwicker method – Amendment 1: Calculation procedure for loudness above 10 kHz.* Berlin: Beuth Verlag.

[10] Ramos-Romero, M. F., Green, N., and Torija, A. J., 2024. *How do flight operations and ambient acoustic environments influence the noticeability and noise annoyance associated with UAS?* Proc. INTER-NOISE 2024.

[11] Lotinga, M. J. B., Torija, A. J., and Green, M. C., 2023. *Dose–response function of community annoyance to urban air mobility (UAM) noise. Appl. Acoust.*, 210, p.109369.

[12] Lotinga, M. J. B., 2024. *Psychoacoustic modelling of UAS noise in diverse acoustic environments.* Proc. INTER-NOISE 2024.

[13] Torija Martínez, A. J., Self, R. H., and Li, Z., 2020. *Effects of a hovering unmanned aerial vehicle on urban soundscapes perception. Transp. Res. Part D: Transp. Environ.*, 78.

[14] More, S., 2010. *Aircraft noise characteristics and metrics.* PhD Thesis. Purdue University.

[15] Sottek, R. and Genuit, K., 2005. *Models of signal processing in human hearing. Int. J. Electron. Commun. (AEÜ)*, 59(3), pp.157–165.

[16] Grekow, J., 2021. *Music emotion recognition using recurrent neural networks with attention. Neural Comput. Appl.*, 33, pp.13227–13236.

[17] Çakır, E., Parascandolo, G., Heittola, T., Huttunen, H. and Virtanen, T., 2017. *Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Trans. Audio Speech Lang. Process.*, 25(6), pp.1291–1303.

[18] Casabianca, T. and Zhang, Y., 2021. *Late Fusion CNN Ensemble for UAV Sound Detection in Complex Environments. Drones*, 5(3), p.54.

[19] Green, M. C. and Torija, A. J. (2024). Soundwalking in Salford: A Soundscape Approach to Drone Noise Assessment. Quiet Drones, Manchester.

[20] Green, M.C., Lotinga, M. J., and Torija, A. J. (2025). Shaping future soundscapes: Affective impact of unmanned aircraft systems noise in urban environments. Submitted for publication, 2025.

[21] Green, M.C., & Torija, A.J. (2025). Soundwalking in Scilly: UAS Noise Impact on Remote Soundscapes. DAS / DAGA, Copenhagen, Denmark, March 2025.

[22] ISO, 2021. *ISO/TS 15666:2021. Acoustics – Assessment of noise annoyance by means of social and socio-acoustic surveys.* Geneva: ISO.

[23] Greco, G. F., Merino-Martínez, R., Osses, A., and Langer, S. C., 2023. *SQAT: A MATLAB-based toolbox for quantitative sound quality analysis.* In: *Proc. INTER-NOISE 2023.*

[24] Yang, Y., et al., 2021. *Torchaudio: Building audio and speech ML tools for PyTorch.* In: *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp.675–679.

[25] Paszke, A., Gross, S., Massa, F., et al., 2019. *PyTorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst.*, 32, pp.8024–8035.