PREDICTION OF *PLEASANTNESS* AND *EVENTFULNESS* PERCEPTUAL SOUND QUALITIES IN URBAN SOUNDSCAPES

Amaia Sagasti, *Martín Rocamora*, *Frederic Font*

Music Technology Group
Universitat Pompeu Fabra, Barcelona
name.surname@upf.edu

ABSTRACT

The acoustic environment induces emotions in human listeners. To describe such emotions, ISO-12913 defines *pleasantness* and *eventfulness* as orthogonal properties that characterise urban soundscapes. In this paper, we study different approaches for automatically estimating these two perceptual sound qualities. We emphasize the comparison of three sets of audio features: a first set from the acoustic and psychoacoustic domain, suggested in ISO-12913; a second set of features from the machine listening domain based on traditional signal processing algorithms; and a third set consisting of audio embeddings generated with a pre-trained audio-language deep-learning model. Each feature set is tested on its own and in combination with ground-truth labels about the sound sources present in the recordings to determine if this additional information improves the prediction accuracy. Our findings indicate that the deep-learning representation yields slightly better performance than the other feature sets when predicting pleasantness, but all of them yield similar performance when predicting eventfulness. Nevertheless, deep-learning embeddings present other advantages, such as faster calculation times and greater robustness against changes in sensor calibration, making them more effective for real-time acoustic monitoring. Furthermore, we observe a clear correlation between the sound sources that are present in the urban soundscape and its induced emotions, specially regarding the sensation of pleasantness. Models like the ones proposed in this paper allow for an assessment of the acoustic environment that goes beyond a characterisation solely based on sound pressure level measurements and could be integrated into current acoustic monitoring solutions to enhance the understanding from the perspective of the induced emotions.

Index Terms— Urban soundscapes, acoustic monitoring, emotions, machine-learning, perception

1. INTRODUCTION

Environmental noise regulations are primarily based on sound pressure level (SPL) measurements. For example, the current European Environmental Noise Directive proposes several SPL-based metrics (like *Ld*, *Le*, *Ln* and their combination, *Lden*) to determine permitted noise levels [1]. The limit values depend on factors such as the time of the day and the designated noise sensitivity of the evaluated area. However, other perspectives argue that SPL is insufficient to reliably characterise the acoustic environment [2]. Some psychoacoustic parameters such as loudness and sharpness [3], or episodic memory and visual perception [4], also play a role in shaping the perception of an acoustic environment. In this field of research, the concept of *soundscape* is key, defining the perceptual and emotional construct related to a physical phenomenon (the acoustic environment). The study of soundscapes constitutes a big challenge due to the intrinsic nature of emotions: they are triggered, brief and unconscious [5]. Addressing these difficulties, the ISO-12913 [6, 7, 8] determines a framework to enable international consensus on the definition and conceptual foundation of soundscapes. The standard proposes a model with *pleasantness* and *eventfulness* as main orthogonal axes to characterise soundscape emotional responses, based on the evidence that physiological responses to all types of stimuli can be organized along the dimensions of *valence* and *arousal* [9, 10], or *pleasantness* and *eventfulness* when applied to soundscapes [11].

In this study, we focus on exploring different approaches for automatically estimating the two aforementioned perceptual sound qualities in urban soundscapes. We put emphasis on the comparison of three feature sets for sound representation: the acoustic and psychoacoustic features suggested in ISO-12913, a set of features from the machine listening domain based on traditional signal processing algorithms, and a third set consisting of the audio embeddings generated by a pre-trained language-audio deep-learning model. Each feature set is tested independently and in combination with ground-truth labels about the sound sources present in each recording to determine if this additional information improves the prediction accuracy. Additionally, we examine the models' suitability for real-time acoustic monitoring applications. Our findings indicate that the deep-learning representation yields slightly better performance than the other feature sets when predicting pleasantness, but all of them yield similar results when predicting eventfulness. Nevertheless, deep-learning embeddings present other advantages, such as presumably faster calculation times and greater robustness against changes in sensor calibration, making them more effective for real-time acoustic monitoring. Furthermore, the addition of sound source information improves the prediction accuracy, especially regarding the sensation of pleasantness, indicating a clear correlation between the sound sources present in the urban soundscape and its induced emotions. Models like the ones proposed in this paper allow for an assessment of the acoustic environment that goes beyond a characterisation solely based on SPL measurements and could thereby contribute to the development of more accurate acoustic monitoring techniques, enhancing the understanding of the evaluated environment from an emotional perspective.

The rest of the paper is structured as follows: Section 2 introduces the related work. Section 3 describes the methods used, detailing the dataset and the features employed. Section 4 describes the evaluation process and Section 5 presents the results of our analysis. Finally, Section 6 consists of a discussion of the findings and their implications, followed by a conclusion in Section 7.

2. RELATED WORK

In recent years, many studies have focused on the two-dimensional model for soundscape emotion assessment, resulting in the creation of datasets and the experimenting of algorithms on them. Fan et al.[12] present diverse valence/arousal classifications using their own dataset EMO-SOUNDSCAPES [13, 14]. In an analogous way, ATHUS (Athens Urban Soundscape) [15], created by the authors of [16], is a dataset for urban soundscape quality recognition which includes pleasantness and unpleasantness annotations. Similarly, the ARAUS (Affective Responses to Augmented Urban Soundscapes) dataset [17], combines real urban soundscape recordings with different *audio maskers* including *traffic*, *construction*, *water*, *wind*, *bird*, and *silence*, creating a large-scale dataset of *augmented soundscapes* labelled with pleasantness and eventfulness scores obtained from listening tests developed according to the ISO-12913 [18]. Using psychoacoustic features, the authors run preliminary experiments for the estimation of pleasantness.

Existing research on automatic sound classification provides insights which are also useful for addressing soundscape quality assessment. As an example of early work, Salamon et al. [19] present a set of classification experiments using traditional machine-learning algorithms applied to their own developed urban soundscape datasets [20, 21]. Later sound classification works adopted deep neural networks to address more complex classification problems (e.g., [22]). However, the most recent approaches involve the use of large pre-trained models to extract audio embeddings (i.e. representations) that can be used to address different classification problems and other sound-related tasks such as sound similarity [23]. In particular, Contrastive Language-Audio Pretraining (CLAP) models [24, 25, 26, 27] use contrastive learning to bring audio and text descriptions into a joint multimodal space, and generate sound representations that capture semantically representative information from the audio.

The studies above provide a good framework for research on urban soundscape characterisation. Nevertheless, two important aspects remain unexplored. Firstly, despite existing research showing that the sound sources present in an acoustic environment contribute to its perceived qualities (e.g. natural sounds contribute positively to the pleasantness of an acoustic environment while construction or traffic noise contributes negatively [11, 18]), there is a lack of experiments incorporating such information as an input for automatically characterising soundscapes. Secondly, none of the studies validates the suitability and robustness of the models in real-time contexts, which is essential for the eventual incorporation of the emotional dimension into acoustic monitoring techniques.

3. METHODS

The core methodology for studying different approaches for predicting the perceptual qualities of pleasantness and eventfulness in urban soundscapes involves data selection, feature extraction, and model training. Our main objective is to evaluate the performance of three different feature sets, and determine which one delivers the best results in terms of accuracy and suitability for real-time applications.

3.1. Dataset

We choose the ARAUS dataset for our experiments because it is the most comprehensive available dataset with pleasantness and eventfulness annotations. ARAUS consists of a set of 25,440 unique and 30s-length augmented audios, created by digitally adding audio maskers (see Section 2) to real urban soundscape recordings. They are organised in a five-fold cross-validation set and an independent test set. Based on the soundscape study methodology suggested in the ISO-12913, the audio clips are individually labelled with 1-5 ratings on how *pleasant*, *annoying*, *eventful*, *uneventful*, *vibrant*, *monotonous*, *chaotic* and *calm* they are according to the participants of a listening test. From these ratings, a global value of pleasantness and eventfulness per recording can be calculated as defined in the standard. These values range from -1 to 1, where negative values indicate unpleasantness or uneventfulness, respectively. Additionally, the ARAUS dataset includes, for each augmented soundscape, pre-calculated acoustic and psychoacoustic features recommended by the ISO-12913. These features are calculated with ArtemiS SUITE [1], which is a proprietary software not easily available to researchers. As part of our work, we provide an open-source Python implementation of such features facilitating the reproducibility of the experiments[2].

3.2. Features

What follows is a description of the three aforementioned feature sets that we consider for our experiments.

Psychoacoustic features The standard ISO-12913 suggests a set of acoustic and psychoacoustic features to characterise urban soundscapes: sharpness, loudness, fluctuation strength, roughness, tonality, L_{Aeq} and L_{Ceq}. We compiled existing open source implementations for these features, and wrote custom implementations for the missing ones. For each feature, we use the statistics mean, maximum, and the 5th, 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th, and 95th percentiles calculated over time. Additionally, replicating ARAUS, the band powers summed over third-octave bands (5Hz to 20kHz) are included. This results in a total of 117 features. It should be noted that these features should not be computed directly on the WAV signal, but on the peak-Pascals pressure signal that results after applying a gain correction to the raw waveform. Thus, the waveform represents the SPL at which the signal was recorded, or, in this case, the level at which it was played in the listening tests. The L_{eq} value, provided in the ARAUS dataset, is used to calculate the mentioned calibration factor (designated as *wav gain*). This feature set is referred to as *ARAUS features*.

Signal processing features This set includes features typically used in traditional machine listening systems. *Freesound Extractor* algorithm from the Essentia audio analysis library[3] generates an extensive set of features from which we use: average loudness; loudness EBU-128; dynamic complexity; spectral flatness, roll-off, flux, skewness, spread, kurtosis and centroid; energy per bands (low, middle-low, middle-high, high); 13th first MFCCs; dissonance; zero-crossing rate; temporal centroid, kurtosis, skewness and spread; log attack-time; inharmonicity; and bpm. For each feature, we compute the statistics mean, variance, and the 20th and 80th percentiles over time, resulting in a total of 139 features. Contrary to the set above, these features are directly linked to the raw audio signal. However, a gain adjustment is performed to ensure that the signal

[1] https://www.head-acoustics.com/products
[2] https://github.com/MTG/soundlights
[3] https://essentia.upf.edu

amplitude proportions between different audio clips reflect the volume at which they were played during the listening tests. To achieve this, we apply the corresponding *wav gain* to each audio and then divide by a common normalization factor to prevent clipping. This set is referred to as *Freesound features*.

CLAP embeddings This set of features consists of the 512-length audio embedding generated using LAION-AI's CLAP model [24]. Since this model is trained using audio-text pairs, the resulting vector is expected to capture semantic information from the audio. This is unlike the *ARAUS* and *Freesound* feature sets, which only represent acoustic information from the sounds. The same scaling procedure used for the *Freesound* feature set is applied in this case. This representation is referred to as *CLAP features*.

The above feature sets are tested independently, but also in combination with information about which sound sources are present in the urban soundscape. Ideally, this information should include the predominant sound source. However, no dataset contains realistic urban soundscape sounds with both this source information and the pleasantness/eventfulness annotations. The ARAUS dataset provides the maskers (see Section 2) that were used to generate each augmented soundscape. Even though these sound sources might not always be predominant, it is guaranteed that they are present. Therefore, we use the maskers' information as a proxy for sound source information, and represent it with one-hot vectors. These six features are referred to as *sources features*.

3.3. Models

The emphasis of this work is not on the models to be trained but on the feature sets. Nevertheless, a number of preliminary experiments were carried out in which the performances of some classic machine-learning regression models were compared (like Support Vector Regression, Multi-layer Perceptron Regressor or regression based on K-Nearest-Neighbours). In these experiments, the best results were obtained by an Elastic Net model (as used in [18]), and a Random Forest Regressor. Therefore, these two models are implemented in our experiments using the Scikit-Learn library [4].

4. EVALUATION

To evaluate the predictive performance and robustness of the feature sets and models, we design a multi-faceted evaluation framework which not only includes the use of ARAUS data folds for cross-validation and model testing but it also involves the creation of a new testing set with data not present in the original dataset. Additionally, the analysis of models' robustness against sensor calibration is evaluated by introducing controlled variations in audio signals. Mean Absolute Error (MAE) is used as the main evaluation metric because it allows for a straightforward interpretation that represents the average absolute difference between the predicted and the ground-truth values.

4.1. Data folds

ARAUS includes five folds of augmented soundscapes for cross-validation and one test fold of 48 audios, reported under the labels *Val* and *fold-0* in Table 1, respectively. In addition, we create a complementary testing fold using 25 urban recordings downloaded from Freesound[5]. The selection was carried out manually by the authors and consists of 30-second excerpts of real urban environment recordings that include sources such as traffic, construction, rain, wind, voices, and music. Following ISO-12913, a listening test was carried out where 22 participants rated the 25 audios with 1-5 scales on how *pleasant*, *annoying*, *eventful*, *uneventful*, *vibrant*, *monotonous*, *chaotic* and *calm* the soundscapes were perceived. From those ratings, ground-truth pleasantness and eventfulness metrics were calculated following the same standard. The audios were calibrated and played at appropriate and varied L_{eq} values, regardless of the audio content. We refer to this fold as *fold-Fs*.

Feature set	Sound sources info	Model	Train	Val	Test fold-0	Test fold-Fs	Var. %
PLEASANTNESS - MAE							
ARAUS	no	RFR	0.29	0.30	0.24	0.21	4.31
	yes	RFR	0.29	0.29	0.26	0.18	
Freesound	no	EN	0.29	0.30	0.22	0.19	2.19
	yes	EN	0.29	0.29	0.22	0.19	
CLAP	no	RFR	0.10	0.28	0.22	0.14	0.53
	yes	RFR	0.10	0.28	0.22	0.14	
EVENTFULNESS - MAE							
ARAUS	no	EN	0.30	0.30	0.15	0.20	1.57
	yes	EN	0.30	0.30	0.14	0.20	
Freesound	no	RFR	0.13	0.29	0.16	0.22	0.02
	yes	RFR	0.13	0.29	0.16	0.22	
CLAP	no	RFR	0.10	0.29	0.20	0.18	-0.41
	yes	RFR	0.10	0.29	0.20	0.18	

Table 1: MAE results for the best performing models for the cross-validation folds and the two testing folds. The MAE variation percentage is included in the last column, representing the mean percentage variation in MAE when adding the sound sources information to the feature set (a positive percentage indicates an improvement in prediction). Note: *EN* and *RFR* stand for Elastic Net and Random Forest Regressor, respectively.

4.2. Robustness analysis

As has been mentioned, to evaluate the robustness of the studied models against different input signal calibration conditions, five controlled variations of the testing fold *fold-0* are generated by modifying the audio signals with *wav gain* adjustments of -6dB, +6dB, +12dB and +18dB; and a fifth variation with random *wav gain* within a fixed range [0-20dB].

5. RESULTS

Table 1 presents the MAE scores for the different combinations of models and feature sets evaluated. For pleasantness, the CLAP representation outperforms the other two feature sets in both test folds, reaching an MAE of 0.14 in *fold-Fs*. This indicates that, on a scale of [-1, 1], the predictions deviate by an average of 0.14. In terms of MAE variation resulting from the inclusion of the *source features*, the CLAP representation shows the smallest improvement,

[4] https://scikit-learn.org

[5] https://freesound.org

with just 0.53%, compared to 4.31% for *ARAUS features* and 2.19% for *Freesound features*. Regarding eventfulness, *ARAUS features* outperform the others when taking into account both test folds, but the smallest MAE for *fold-Fs*, 0.18 points, is achieved by *CLAP features*. When examining the MAE percentage variation, the inclusion of sound sources information has a smaller impact, with percentages closer to zero than those observed for pleasantness.

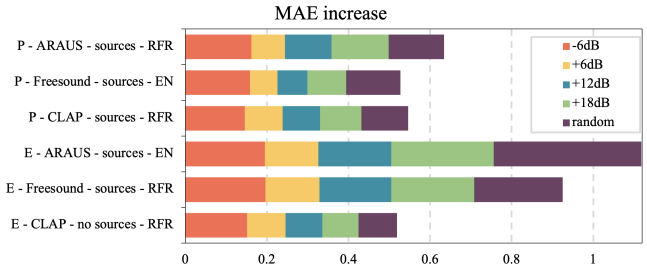

Figure 1: Increase in MAE value provoked by each *fold-0* variation with respect to the original and unvaried *fold-0* MAE.

Furthermore, all controlled calibration variations generated of *fold-0* result in higher MAE values. Figure 1 illustrates the increase of MAE only for the best-performing model for each feature set. The first noticeable observation is that the impact is greater on eventfulness, where the MAE increase is more pronounced. Also, it can be noted that the *ARAUS* feature set is more negatively affected in both cases, whereas the CLAP feature set appears to be the least affected. Among the variations, the 6dB increase in *wav gain* caused the smaller impact.

In terms of calculation time, *CLAP features* is the fastest set, taking 0.5s to calculate the embeddings for a 30s-long stereo audio file (sampled at 48kHz, run in a MacBook Pro M3). *Freesound features* and *ARAUS features* take 8x and 144x longer, respectively. Note that this comparison is limited as these feature sets are implemented in different frameworks and languages.

6. DISCUSSION

The experimental results indicate that, for predicting pleasantness, *CLAP features* outperform the other two sets, achieving an MAE of 0.22 and 0.14 for *fold-0* and *fold-Fs*, respectively. These results occur both when *CLAP features* are used alone and when combined with *sources features*, with only a 0.53% difference in performance between the two scenarios. Since CLAP embeddings intrinsically contain semantic information about the audio, additional sound source information is redundant. Conversely, for feature sets that lack this semantic data, including the source information positively impacts the accuracy in the prediction of pleasantness: the performances of *ARAUS* and *Freesound* feature sets improve by 4.31% and 2.19%, respectively. These findings suggest a clear correlation between the sound sources that are present in the urban soundscape and the perceived sensation of pleasantness. In fact, these results coincide with those obtained in the listening test. A quantitative analysis, which can be seen in Figure 2, shows a clear source-class separation on the pleasantness scale depending on the predominant sound source: construction and traffic noises are positioned on the negative side of the axis, while natural sounds are on the positive side.

For predicting eventfulness, all feature sets perform similarly,

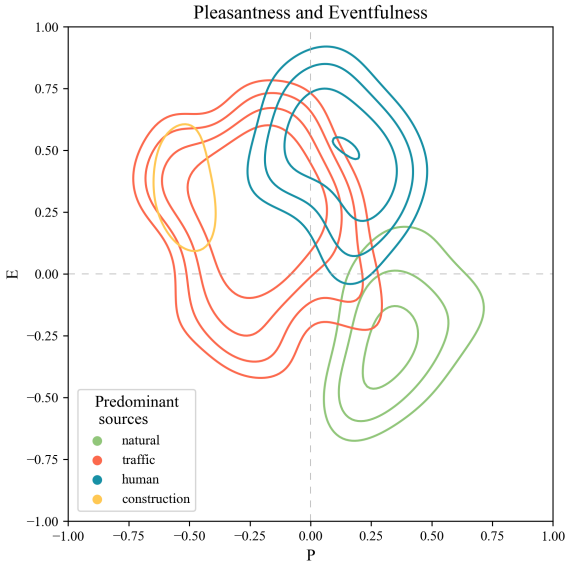

Figure 2: Two-dimensional Kernel Density Estimate plot of the pleasantness(P) and eventfulness(E) values reported from the answers to the listening test.

with *ARAUS features* showing slightly better results when considering the MAE mean of both test sets. Besides, the impact of the inclusion of *source features* is negligible, being smaller than 2% for *ARAUS features*, and close to zero for *Freesound* and *CLAP* feature sets. This indicates a weaker correlation between the sound sources present in the soundscape and the sensation of eventfulness, coinciding again with the data extracted from the listening test, where there is more overlap between class groups when seen from the eventfulness axis (see Figure 2).

In terms of robustness against changes in sensor calibration, none of the trained models demonstrate strong capabilities, as MAE increases notably in every *fold-0* variation case. Nevertheless, predictions of eventfulness are more negatively affected, potentially indicating a correlation between SPL, or loudness, and the perception of eventfulness. Moreover, models trained with *CLAP features* seem to be slightly less impacted by the calibration changes. In addition to this, their rapid generation time suggests that *CLAP features* are adequate for real-time contexts.

7. CONCLUSION

This research shows that CLAP embeddings generated by LAION-AI's CLAP model demonstrate high performance as input to models for predicting *pleasantness* and *eventfulness* perceptual sound qualities. Even though the sound representation does not present strong robustness to variations in sensor calibration, it can be computed rapidly, making it suitable for real-time applications. Moreover, our study indicates a clear correlation between the sound sources present in an urban soundscape and its sensation of pleasantness. Future research directions could include evaluating the developed models in the context of a real-world acoustic sensor network and incorporating sound classification and source separation technologies to improve the models' accuracy and capabilities for meaningful soundscape characterisation and monitoring.

8. ACKNOWLEDGMENTS

This work was supported by the project "Soundlights: Distributed Open Sensors Network and Citizen Science for the Collective Management of the City's Sound Environments" (9382417), funded by *BIT Habitat (Ajuntament de Barcelona)* under the program *La Ciutat Proactiva*; and by the *IA y Música: Cátedra en Inteligencia Artificial y Música* (TSI-100929-2023-1) by the *Secretaría de Estado de Digitalización e Inteligencia Artificial* and *NextGenerationEU* under the program *Cátedras ENIA 2022*. Additionally, we would like to acknowledge *Bitlab Cooperativa Cultural* for the fruitful discussions in the context of the Soundlights project.

9. REFERENCES

[1] European Parliament and Council, "Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the assessment and management of environmental noise - Declaration by the Commission in the Conciliation Committee on the Directive relating to the assessment and management of environmental noise," 2002, Official Journal of the European Union, L 189, p. 12–25. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2002/49/oj

[2] M. Raimbault and D. Dubois, "Urban soundscapes: Experiences and knowledge," *Cities*, vol. 22, no. 5, pp. 339–350, 2005.

[3] J. M. Morillas, V. G. Escobar, G. R. Gozalo, R. Vílchez-Gómez, J. A. M. Sierra, J. T. Carmona, C. P. Gajardo, and F. J. C. D. Río, "Sound quality in urban environments and its relationship with acoustic parameters," in *Noise Control and Acoustics Division Conference (NCAD)*, 2013.

[4] B. Truax, "Environmental sound and its relation to human emotion," *Canadian Acoustics*, vol. 44, no. 3, 2016.

[5] A. Fiebig, P. Jordan, and C. C. Moshona, "Assessments of acoustic environments by emotions – the application of emotion theory in soundscape," *Frontiers in Psychology*, vol. 11, 2020.

[6] International Organization for Standardization, "ISO 12913-1. Acoustics-Soundscape-Part 1: Definition and conceptual framework," 2014. [Online]. Available: www.iso.org

[7] ——, "ISO 12913-2. Acoustics-Soundscape-Part 2: Data collection and reporting requirements," 2018. [Online]. Available: www.iso.org

[8] ——, "ISO 12913-3. Acoustics-Soundscape-Part 3: Data analysis," 2019. [Online]. Available: www.iso.org

[9] A. Mehrabian and J. A. Russell, *An approach to environmental psychology.* The MIT Press, 1974.

[10] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[11] Östen Axelsson, M. E. Nilsson, and B. Berglund, "A principal components model of soundscape perception," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 2836–2846, 2010.

[12] J. Fan, F. Tung, W. Li, and P. Pasquier, "Soundscape emotion recognition via deep learning," in *Sound and Music Computing Conference (SMC)*, 2018.

[13] J. Fan, M. Thorogood, and P. Pasquier, "Emo-soundscapes: A dataset for soundscape emotion recognition," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017.

[14] "Emo-Soundscapes — Dataset." [Online]. Available: https://www.metacreation.net/projects/emo-soundscapes/

[15] "ATHUS (Athens Urban Soundscape) - Dataset." [Online]. Available: https://users.iit.demokritos.gr/~tyianak/soundscape/

[16] T. Giannakopoulos, M. Orfanidi, and S. Perantonis, "Athens Urban Soundscape (ATHUS): A Dataset for Urban Soundscape Quality Recognition," in *Multimedia Modelling (MMM)*, 2019.

[17] "ARAUS (Affective Responses to Augmented Urban Soundscapes) - Dataset." [Online]. Available: https://researchdata.ntu.edu.sg/dataset.xhtml?persistentId=doi:10.21979/N9/9OTEVX

[18] K. Ooi, Z. T. Ong, K. N. Watcharasupat, B. Lam, J. Y. Hong, and W. S. Gan, "ARAUS: A Large-Scale Dataset and Baseline Models of Affective Responses to Augmented Urban Soundscapes," *IEEE Transactions on Affective Computing*, vol. 15, pp. 105–120, 2024.

[19] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *ACM Multimedia Conference (MM)*, 2014.

[20] "UrbanSound - Dataset." [Online]. Available: https://urbansounddataset.weebly.com/urbansound.html

[21] "UrbanSound8K - Dataset." [Online]. Available: https://urbansounddataset.weebly.com/urbansound8k.html

[22] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[23] R. O. Araz, D. Bogdanov, P. Alonso-Jiménez, and F. Font, "Evaluation of deep audio representations for semantic sound similarity," in *International Conference on Content-based Multimedia Indexing (CBMI)*, In Press.

[24] "LAION-AI/CLAP Github Repository." [Online]. Available: https://github.com/LAION-AI/CLAP

[25] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[26] "microsoft/CLAP Github Repository." [Online]. Available: https://github.com/microsoft/CLAP

[27] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.