A REVISIT OF AUDIO EVALUATION THROUGH HUMAN IMPRESSIONS: DEFINING AND MODELING A MULTIDIMENSIONAL PERCEPTUAL TASK

Hiroshi Nishijima[1], *Daisuke Saito*[1], *Nobuaki Minematsu*[1]

[1]The University of Tokyo, Japan. {hiroshi, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

***Abstract*—Current audio evaluation paradigms predominantly rely on technical metrics or single-dimensional subjective scores. These methods inadequately capture the multifaceted nature of human auditory perception. This paper reframes audio evaluation as a multidimensional perceptual task. We formally define subjective impression as a computational problem with measurable dimensions. To this end, we introduce a new dataset of 4,110 environmental sounds from FSD50K. It is annotated with five perceptual dimensions: pleasantness, clarity, brightness, calmness, and immersion. Our analysis reveals both independence and meaningful correlations within this perceptual space. A notable finding is the strong relationship between pleasantness and calmness. Furthermore, we demonstrate the feasibility of automated impression prediction. Our baseline models use fine-tuned BEATs representations and achieve a mean squared error below 0.7. This value corresponds to an average deviation of less than one point on a seven-point scale. This work provides the foundation for a human-centered evaluation of audio generation systems and sound design. It enables assessment based on nuanced perceptual qualities rather than technical fidelity alone.**

***Index Terms*—Subjective impression, Subjective audio evaluation, Multidimensional perception, Semantic differential, Environmental sound**

1. INTRODUCTION

Audio generation systems have been steadily maturing from research prototypes to practical applications. Recent text-to-audio models such as Tango [1], Tango2 [2], Make-An-Audio [3], AudioLDM2 [4], and AudioGen [5] exhibit remarkable technical performance. In addition, ElevenLabs offers a system capable of generating realistic sound effects from textual input, enabling practical use cases [6]. Given these developments, the quantitative evaluation of system performance and generated audio quality is becoming increasingly important.

Current evaluation methods for generated audio comprise technical and subjective assessments. Technical evaluations operate at two levels. Signal-level metrics, such as Signal-to-Noise Ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ), measure acoustic fidelity against a ground-truth signal. In contrast, embedding-level metrics aim to capture higher-level semantic properties. For instance, Fréchet Audio Distance (FAD) [7] measures the distributional similarity between the embeddings of generated audio and those of real-world recordings, while CLAPScore [3] evaluates semantic alignment by computing the similarity between their corresponding audio and text embeddings. On the other hand, subjective evaluations rely on human judgment. The most common method is the Mean Opinion Score (MOS), where listeners rate perceptual quality, often supplemented by tasks that assess the relevance of the audio to a given textual prompt. Crucially, a common thread unites all these approaches: they invariably require a reference for comparison, whether it is a ground-truth signal, a distribution of existing audio, an input text, or a human listener's internal standard of quality and relevance.

However, humans possess the remarkable ability to evaluate sounds in an absolute, non-referential manner, even when a ground-truth signal or a real-world counterpart does not exist. For instance, an abstract, synthesized sound can consistently evoke rich semantic impressions such as "brightness," "clarity," or "tension" without any direct comparison. Furthermore, this perceptual experience is often multifaceted, allowing for multiple, coexisting interpretations of a single audio clip. This capability stands in stark contrast to existing metrics, which are confined to measuring fidelity or semantic alignment against a predefined reference. Consequently, they are ill-equipped to quantify the intrinsic, absolute qualities of generated audio, offering only a limited perspective on a system's true capabilities.

Therefore, integrating this human-like, non-referential evaluative capability into assessment frameworks of audio is a crucial next step. Such an approach promises significant benefits. For one, it would enable the meaningful evaluation of novel or abstract sounds for which no ground truth exists. Moreover, it would allow us to move beyond mere fidelity and begin to quantify more elusive yet critical qualities like the "creativity" or "expressiveness" of generative models. Developing a system capable of this absolute quality assessment is thus essential for a more holistic and meaningful evaluation of modern audio generation technologies.

To achieve this, we propose a framework that reframes audio evaluation by defining it as a human-centered, multidimensional perceptual task. Our primary contributions are as follows:

1) A formal specification of subjective impression as a computational problem with well-defined, measurable dimensions.
2) A dataset of 4,110 environmental sounds, systematically annotated across five key perceptual dimensions using the semantic differential (SD) method.
3) An empirical analysis of the dataset that reveals both the independence and the meaningful correlations within this perceptual space.
4) Baseline computational models that demonstrate the feasibility of automatically predicting these perceptual impressions directly from audio signals.

This framework represents a transition from metric-based quality assessment to human-centered perceptual task modeling.

2. RELATED WORK

Recent studies have explored various aspects of environmental sound perception, including emotional responses and contextual influences. Research on emotional responses has shown that different acoustic environments evoke feelings such as pleasantness and arousal, with human emotions tracking changes in acoustic features like frequency, intensity, and speed [8]. Furthermore, the role of context, which encompasses spatial, temporal, and functional factors, has been shown to significantly shape how individuals perceive and evaluate soundscapes, highlighting that the same sound can be perceived differently depending on its surrounding environment [9], [10]. To investigate these aspects, methods such as the Semantic Differential (SD) scale [11] or Russell's Circumplex Model of Affect [12] are adopted to quantify these experiences [13]–[15].

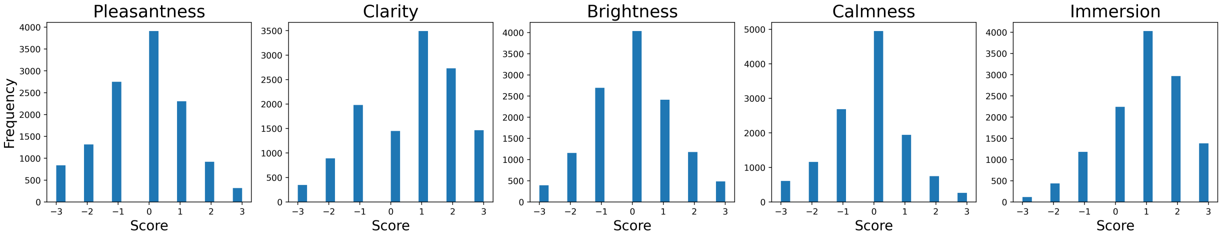

Fig. 1: Histograms of the subjective ratings for each of the five evaluation axes.

Several studies have applied the SD method to examine the impressions of sounds formed by listeners. A line of works conducted listening experiments with 38 pairs of bipolar adjectives grouped into timbral, informational, and affective categories [16], [17]. Another line investigated the prevalence of 12 pairs of representative adjectives in the prior literature [18]. There is another study that analyzed whether addition of visual imagery alters auditory impression [19]. The results of the factor analysis obtained in these investigations converge on three fundamental perceptual dimensions of sound: an aesthetic factor, a brightness factor, and a quantitative factor.

3. DATASET CREATION

3.1. Dataset selection

A dataset was constructed by extending a subset of the FSD50K [20], a large-scale collection of environmental sounds with event-based categorical annotations. To ensure balanced representation across sound categories, 4,110 clips were extracted from both the development set (40,966 clips) and the evaluation set (10,231 clips). This was achieved with a random selection process, by which approximately 8 % of the clips from each class were selected. We chose FSD50K as the basis because its clips typically contain a single dominant sound event. This property makes it suitable for investigating relationships between acoustic events and human impressions.

3.2. Five-dimensional impression space

Based on extensive preliminary pilot studies on environmental sound perception, we define the following five evaluation axes that capture distinct aspects of subjective auditory experiences: **Pleasantness** evaluates the hedonic quality of sounds, ranging from pleasant to unpleasant. This dimension captures the fundamental affective response to auditory stimuli. **Clarity** characterizes the perceived sharpness and articulateness of sounds, ranging from clear to indistinct or vague. It may reflect the listener's ability to resolve and differentiate sound sources. **Brightness** measures the spectral character impression, ranging from bright to dark. This perceptual dimension often encompasses subjective brightness perception beyond simple spectral analysis. **Calmness** evaluates the emotional impacts related to stress and relaxation, ranging from calming to irritating. It captures the potential influence of sounds on listener stress levels and emotional state. **Immersion** assesses the spatial and engaging quality of sounds, ranging from immersive to non-immersive. This dimension may reflect the sound's ability to create a sense of spatial presence and engagement.

In selecting the above perceptual axes, we first chose one bipolar adjective pair from each of the three groups identified in prior work: timbral, informational, and affective. As a result, we obtained Brightness, Calmness, and Immersion. We made this choice in a deterministic manner to avoid redundancy among axes. After that, we added two further factors, Pleasantness and Clarity, which showed the highest contribution ratios reported in [19]. This procedure yielded a five-factor framework.

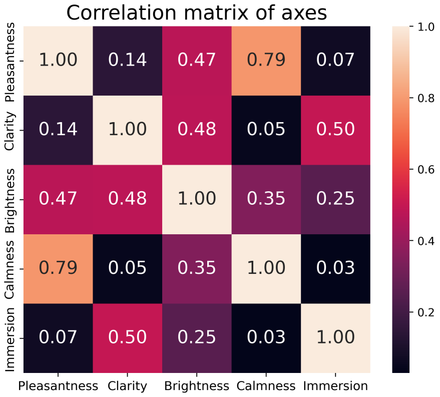

Fig. 2: Heatmap of the correlation matrix between the five evaluation axes.

3.3. Annotation methodology

The SD method was adopted to collect subjective impression annotations. Each audio clip was independently evaluated three times, each by a different annotator who used a seven-point bipolar scale ranging from -3 to $+3$. The midpoint of this scale, 0, was interpreted as neutral. The annotation process was conducted through a web-based interface that presented each audio clip with five evaluation axes simultaneously.

3.4. Analysis of dataset

Figure 1 depicts the distribution characteristics of our dataset. The majority of dimensions manifest approximately normal distributions, with Clarity and Immersion exhibiting positive skews as a consequence of the superior quality of FSD50K clips.

Figure 2 is a correlation matrix of the five axes. It reveals that there is a strong positive correlation between Pleasantness and Calmness ($r = 0.79$), while pairs of the other three dimensions, such as Pleasantness and Immersion, Clarity and Calmness, Calmness and Immersion, show weak correlations ($|r| < 0.1$). The results suggest the complementary nature of our evaluation framework.

4. PREDICTION MODEL

In order to evaluate the characteristics and effectiveness of our collected dataset, and to demonstrate the feasibility of automated impression prediction, predictive models were constructed that estimate scores for each perceptual dimension by extracted acoustic features. The following section first details the methodologies to transform raw audio data into meaningful representations for machine learning, followed by the architectures of the prediction models.

4.1. Feature extraction

In this study, we extracted two sets of features for different analytical purposes. Conventional acoustic features were used to investigate which specific signal characteristics contribute to each perceptual dimension, with Support Vector Regressor (SVR). In parallel, audio representations from a self-supervised learning (SSL) model were utilized to establish a strong performance benchmark for the regression task.

Conventional acoustic features: Based on previous studies that aim to score musical impressions [21], we extracted a set of conventional audio features from each clip, covering signal energy, spectral shape, rhythm, and timbre. The extracted features are:

- Root Mean Square (RMS) and Zero Crossing Rate (ZCR): mean, standard deviation, and frame-to-frame deviation.
- Spectral centroid, rolloff, and bandwidth: first to fourth moments (mean, standard deviation, skewness, kurtosis).
- Spectral contrast: computed over 7 frequency bands with average and variability.
- Mel-frequency cepstral coefficients (MFCCs): 13 coefficients with their delta and delta-delta derivatives.
- Tempo and beat statistics: beat count, mean interval, and its variation.
- Chroma features: harmonic structure over 12 pitch classes.

All features were normalized to have zero mean and unit variance across samples.

Audio representations from SSL models: The efficacy of SSL models in the domain of audio representation learning has been demonstrated across a range of downstream tasks. In this study, we utilized representations extracted from the pre-trained BEATs model [22] as a feature extractor.

4.2. Models

All models in this study are designed as regression models that output continuous values rather than discrete scores. This design choice enables the models to capture the nuanced gradations in human perception that exist between discrete rating points on the 7-point scale.

Mel-spectrogram Regressor (Baseline): As the primary baseline, we implemented a convolutional regressor that takes mel-spectrograms as input features. To ensure a fair comparison with other models, we scaled its architecture to approximately 90 million parameters. This scale matches that of the model created by fine-tuning the BEATs iter3+ pre-trained model.

Support Vector Regressor (SVR): As another baseline, we adopted an SVR with a radial basis function (RBF) kernel, which is well-suited for modeling non-linear relationships.

Multi-Layer Perceptron (MLP) regressor: In order to assess the impact of various input representations and model configurations on prediction performance, several Multi-Layer Perceptron (MLP) regressors were investigated. All MLP models were designed to have a total parameter count of approximately 90 million. The configurations under investigation are listed as follows:

- **Audio:** Uses frozen BEATs audio representations as input.
- **Audio + Label:** Concatenates frozen BEATs audio representations with the 200 FSD50K sound classes.
- **Audio (Fine-tuned):** Fine-tunes the pre-trained BEATs model end-to-end with the MLP regressor, adapting audio representations to the task.
- **Audio + Label (Fine-tuned):** Fine-tunes the pre-trained BEATs model along with the MLP, incorporating concatenated sound classes.
- **Label Only:** Utilizes only learned embeddings of the 200 FSD50K sound classes as input.

Table 1: Comparison of prediction model performance on the test set (510 samples). All models except SVR are trained on five dimensions simultaneously.

Model	MSE↓	MAE↓	LCC↑	SRCC↑	KTAU↑
Baseline	0.879	0.741	0.519	0.588	0.427
SVR (Average)	1.571	0.988	0.280	0.374	0.281
Audio (Freeze)	0.693	0.665	0.667	0.693	0.517
+ Label	0.723	0.679	0.625	0.675	0.502
Audio (Fine-tuned)	0.691	**0.652**	**0.696**	**0.705**	**0.528**
+ Label	**0.691**	0.652	0.695	0.704	0.528
Label Only	1.209	0.871	0.327	0.379	0.265

4.3. Training details

The training dataset comprised 3,216 clips (78.2 %) for training, 384 clips (9.4 %) for validation, and 510 clips (12.4 %) for testing. A key difference in our experimental setup was the handling of the five perceptual dimensions. For the SVR baseline, which was trained on the conventional acoustic features, we trained five independent regressors, each of which dedicated to predicting a corresponding single dimension. In contrast, other models were trained to predict all five dimensions jointly from a single model. This multi-output design allows the models to potentially leverage inter-dimensional dependencies during learning.

The baseline and all MLP models and were optimized using the Mean Squared Error (MSE) loss. Training adopted the AdamW optimizer [23] with a learning rate of 1×10^{-4} and weight decay of 0.01. Early stopping based on validation loss was applied to mitigate overfitting.

5. EXPERIMENTS AND RESULTS

5.1. Model prediction results

The performance of the constructed models was evaluated by multiple metrics; Mean Squared Error (MSE), Mean Absolute Error (MAE), Spearman's Rank Correlation Coefficient (SRCC), and Kendall's Tau (KTAU).

Table 1 summarizes the performance of various models in the test set. The fine-tuned MLP models achieve MSE values around 0.69. Given our 7-point scale ranging from -3 to $+3$, this value indicates that 68 % of predictions fall within ±0.83 scale points of ground truth (assuming normal error distribution). The result demonstrates the practical utility of the model for most applications.

The performance difference between models with and without class labels is marginal (MSE: 0.6909 vs. 0.6914). It suggests that fine-tuned BEATs features already encode substantial semantic information. The label-only model (MSE: 1.2089) performs considerably worse than the audio-based models. This degradation demonstrates that single use of class labels without audio information is insufficient to capture perceptual impressions. Table 2 presents performance across individual impression axes using the best performing model (Audio Fine-tuned). Performance varies considerably between dimensions, with Immersion showing the poorest prediction accuracy (LCC: 0.465) compared to other dimensions. This performance disparity can be partially explained by the distributional characteristics and inter-dimensional relationships observed in our dataset. Pleasantness and Calmness, which share a strong positive correlation, both achieve relatively robust prediction performance with similar error patterns. This suggests that their shared variance may provide mutually reinforcing signals during multi-dimensional training. The distribution characteristics also play a crucial role: Clarity and Immersion both exhibit positive skews in our

Table 2: Performance evaluation across perceptual dimensions using the Audio (Fine-tuned) model. Std. represents the standard deviation of absolute prediction errors.

Dimension	LCC ↑	SRCC ↑	KTAU ↑	MAE ↓	Std.
Pleasantness	0.6300	0.5989	0.4492	0.6353	0.8248
Clarity	0.6296	0.6444	0.4760	0.7221	0.9034
Brightness	0.6129	0.6120	0.4515	0.6309	0.8047
Calmness	0.5816	0.5874	0.4332	0.6219	0.7848
Immersion	0.4652	0.4805	0.3440	0.6522	0.8225

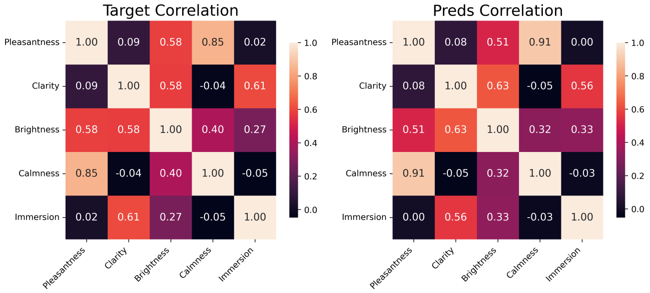

Fig. 3: Comparison of Pearson correlation matrices computed from ground-truth scores (Target, left) and Audio (Fine-tuned) model predictions (Preds, right).

dataset (Figure 1), indicating concentration toward higher ratings. This skewed distribution reduces the diversity of training examples across the full perceptual range, particularly for negative ratings, which may contribute to their relatively higher prediction errors.

Figure 3 is a comparison between Pearson correlation matrices computed from ground-truth scores (left) and those matrices based on predictions from the Audio (Fine-tuned) model (right) on the test set. The model largely captures the strong positive correlation between pleasantness and calmness (ground truth: $r = 0.85$, predictions: $r = 0.91$). In particular, dimension pairs with low correlation in ground truth, such as pleasantness–immersion and calmness–immersion, maintain their low correlations similarly in predictions. This demonstrates that our multidimensional simultaneous training successfully captures inter-dimensional dependencies while avoiding unnecessary coupling between independent perceptual axes.

5.2. Feature importance analysis

To identify the acoustic drivers for each perceptual dimension, we treated the analysis as a classification problem. The continuous outputs from each SVR model were discretized into seven classes, and Linear Discriminant Analysis (LDA) was then applied to determine the most discriminative input features for separating these classes.

The analysis revealed dimension-specific acoustic signatures. Brightness was strongly associated with variability in high-frequency components. These fluctuations in spectral energy at higher bands played a key role in how sounds were perceived as bright. Calmness depended on temporal regularity such as rhythm stability and onset variability. It was also related to reduced low-frequency energy, which reflected the soothing impression of steady and less energetic sounds.

5.3. Error analysis

As shown in Figure 4, the predictions track the diagonal most closely near the frequent scores, namely $+1$ for Clarity and Immersion and 0 for Pleasantness, Brightness, and Calmness. In these regions the ample training data yield small bias and narrow confidence intervals. Toward the scale extremes the mapping contracts: ground-truth values of ± 3 are predicted at about ± 1.5, and the error bars widen. This regression

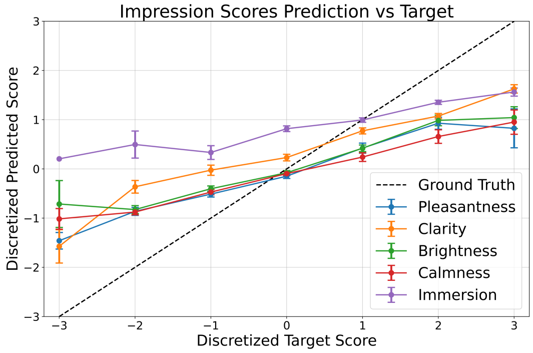

Fig. 4: Mean predicted impression scores for each perceptual axis, averaged over discretized target scores. The dashed line represents the ground truth.

to the mean reflects both the scarcity of extreme samples and the quadratic penalty imposed by the mean-squared-error objective. The results indicate that stronger class balancing and ordinal-aware loss functions are required to maintain calibration across the full rating range.

6. DISCUSSION AND LIMITATIONS

While our results demonstrate feasibility of automated impression prediction, several limitations exist. First, our dataset derives from FSD50K's high-quality recordings, which may limit generalizability to real-world audio that often contains overlapping acoustic events and diverse recording conditions. Notably, our annotations are restricted to clips containing a single dominant sound event, whereas ambient sounds in real-world environments typically involve complex mixtures of multiple concurrent sources. Second, the annotation process may reflect cultural or demographic biases despite following established methodology. The performance gap at extreme ratings highlights the sparse training examples at perceptual boundaries. This suggests that specialized data collection strategies would be necessary.

The strong pleasantness-calmness correlation ($r = 0.79$) suggests certain perceptual qualities naturally co-occur, enabling more efficient annotation protocols. Our continuous regression approach captures subtle perceptual gradations lost in discrete categorization, particularly valuable for audio generation systems requiring fine-grained optimization feedback.

In addition, our framework does not exclude degradations such as noise or artifacts; they can be reflected in impression scores through dimensions like clarity, pleasantness, and calmness. It complements fidelity-based metrics by linking potential degradations with broader perceptual impact.

7. CONCLUSION

We have presented a framework that reframes audio evaluation as a multidimensional perceptual task. Our approach moves beyond referential metrics and captures the absolute characteristics of sound. We have created a dataset of environmental sounds with five-dimensional subjective annotations and developed models that have achieved practical prediction accuracy. This success confirms both that human impressions can be formulated as a computational problem and automated assessment is feasible. Our approach opens new possibilities for evaluating audio based on human-centered criteria, not just technical fidelity. For future work, this framework should be extended to broader acoustic domains and integrated with audio generation systems for human-centered optimization.

REFERENCES

[1] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction guided latent diffusion model," in *Proc. ACM Multimedia*, 2023, pp. 3590–3598.

[2] N. Majumder, C.-Y. Hung, D. Ghosal, W.-N. Hsu, R. Mihalcea, and S. Poria, "Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization," in *Proc. ACM Multimedia*, 2024, pp. 564–572.

[3] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *Proc. ICML*, 2023.

[4] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2871–2883, 2024.

[5] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," in *Proc. ICLR*, 2023.

[6] ElevenLabs, "AI Sound Effects Generator," https://elevenlabs.io/sound-effects, 2025, accessed: 2025-07-11.

[7] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," in *Proc. ICASSP*, 2018, pp. 2350–2354.

[8] W. Ma and W. F. Thompson, "Human emotions track changes in the acoustic environment," *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14 563–14 568, 2015.

[9] X. Zhao, S. Zhang, Q. Meng, and J. Kang, "Influence of contextual factors on soundscape in urban open spaces," *Applied Sciences*, vol. 8, no. 12, 2018.

[10] R. Risley, V. Shafiro, S. Sheft, A. Balser, and B. Gygi, "The role of context in the perception of environmental sounds," *Proceedings of Meetings on Acoustics*, vol. 15, no. 1, p. 060008, 08 2015.

[11] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The measurement of meaning*. University of Illinois press, 1957.

[12] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[13] 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. Volume 11 - 2020, 2020.

[14] D. Västfjäll, "Emotional reactions to tonal and noise components of environmental sounds," *Psychology*, vol. 04, pp. 1051–1058, 01 2013.

[15] Y. Yin, Y. Shao, Y. Hao, and X. Lu, "Perceived soundscape experiences and human emotions in urban green spaces: Application of russell's circumplex model of affect," *Applied Sciences*, vol. 14, no. 13, 2024.

[16] K. Abe, K. Ozawa, Y. Suzuki, and T. Sone, "Evaluation of environmental sounds using adjectives describing sound quality, emotional state, and information carried by sounds." *Journal of the Acoustical Society of Japan*, vol. 54, pp. 343–350, 1998.

[17] ——, "The influence of visual information on perception of environmental sounds," *Journal of the Acoustical Society of Japan*, vol. 56, pp. 793–804, 2000.

[18] M. Miyakawa and S. Aono, "Reexamination of rating scales on the impressions of environmental sounds," *Acoustical science and technology*, vol. 23, no. 3, p. 179, 2002.

[19] M. Miyakawa, T. Nakatsukasa, and S. Aono, "The effects of visual and auditory information on the impression of sound environment," *The Journal of the INCE of Japan*, vol. 26, no. 1, pp. 53–59, 2002.

[20] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, 2021.

[21] Y. Ito, Y. Yamanishi, and S. Kato, "Estimation of music impressions using musical fluctuation features," *The Journal of the Acoustical Society of Japan*, vol. 68, no. 1, pp. 11–18, 2011.

[22] S. Chen, Y. Wang, C. Gong, Z. Liu, E.-H. Song, S. Wang, Y. Zhang, J. Xuan, Y. Wang, J. Wu *et al.*, "Beats: Audio pre-training with acoustic tokenizers," in *Proc. ICML*, 2022.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.