ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11462896

EXPLORING HOW AUDIO EFFECTS ALTER EMOTION WITH FOUNDATION MODELS

Stelios Katsis, Vassilis Lyberatos, Spyridon Kantarelis, Edmund Dervakos, and Giorgos Stamou

National Technical University of Athens, Athens, Greece
stelioskatsis12@gmail.com,{vaslyb, spyroskanta}@ails.ece.ntua.gr,
eddiedervakos@islab.ntua.gr, gstam@cs.ntua.gr

ABSTRACT

Audio effects (FX) such as reverberation, distortion, modulation, and dynamic range processing play a pivotal role in shaping emotional responses during music listening. While prior studies have examined links between low-level audio features and affective perception, the systematic impact of audio FX on emotion remains underexplored. This work investigates how foundation models—large-scale neural architectures pretrained on multimodal data—can be leveraged to analyze these effects. Such models encode rich associations between musical structure, timbre, and affective meaning, offering a powerful framework for probing the emotional consequences of sound design techniques. By applying various probing methods to embeddings from deep learning models, we examine the complex, nonlinear relationships between audio FX and estimated emotion, uncovering patterns tied to specific effects and evaluating the robustness of foundation audio models. Our findings aim to advance understanding of the perceptual impact of audio production practices, with implications for music cognition, performance, and affective computing.

Index Terms— Foundation Models, Affect Computing, Audio FX, Model Auditing

1. INTRODUCTION

The relationship between sound and human emotion has been extensively studied across multiple domains, particularly within affective computing and music cognition. Prior research has demonstrated how low-level audio features—such as timbre, tempo, pitch, and rhythm—correlate with emotional responses in listeners [1, 2, 3, 4, 5]. Parallel investigations have examined how environmental auditory conditions, including soundscape complexity and the number of simultaneous sound sources, influence affective perception [6]. However, there remains a notable gap in understanding how audio FX—such as reverberation, distortion, modulation, and dynamic range processing—systematically alter emotional responses in music listening contexts.

Unlike fundamental audio features, audio FX are intentional sound design tools commonly applied in music production and performance to modulate aesthetic and emotional impact. Yet, their role in shaping affective perception has been underexplored. For instance, [7] showed that longer reverberation times in recorded music increase ratings of "sublimity," including emotional qualities like awe and nostalgia. Similarly, distortion has been linked to heightened arousal and affective intensity, potentially evoking excitement or even aggression [2]. Higher rates of vibrato are linked to elevated arousal levels [8, 9]. Additionally, an increase in the number of perceived sound sources correlates with increased arousal but may reduce pleasantness [6].

Despite these isolated findings, there is no comprehensive study that systematically investigates how diverse audio FX in music—individually or in combination—impact emotional perception. Given that audio FX are central to contemporary music production, a focused analysis of their affective consequences is both timely and necessary.

At the same time, advances in foundation models—large-scale neural architectures pre-trained on multimodal datasets, have demonstrated an ability to capture high-level semantic and perceptual relationships across domains, including with audio FX [10, 11]. In the music domain, foundation models such as MusicLM [12], MERT [13], Qwen [14], and CLAP [15] have shown promising abilities in generating emotionally expressive audio, learning disentangled latent representations, and enabling cross-modal retrieval tasks [16]. These models implicitly encode rich associations between musical structure, timbre, and affective meaning, making them ideal tools to probe novel hypotheses about music-emotion relationships.

In this work, we leverage foundation models to investigate how specific audio FX influence the way foundation models perceive emotions in music. We analyze performance changes, shifts in predicted emotional labels and dimensions, and alterations in model embeddings across three state-of-the-art foundation models and diverse datasets with both categorical and dimensional annotations, using six common audio FX at different intensity levels, including fixed real-world. By combining controlled manipulations with these realistic production scenarios, we identify correlations between audio FX and estimated emotions and uncover properties of how foundation models represent these affective cues.

2. MATERIAL

In this study, we employed a diverse set of datasets and models, spanning both deep and shallow architectures, along with several commonly used audio FXs with varying parameter settings. The code, analytical details, and complete experimental results are available in our GitHub repository [1].

We utilized three **datasets** capturing both categorical and dimensional emotional annotations: *EMOPIA* [17], *DEAM* [18], and *witheFlow* [5]. The *EMOPIA* dataset is a multimodal (audio and MIDI) resource focused on perceived emotion in pop piano music, containing 1,087 clips from 387 songs. Each clip is annotated with

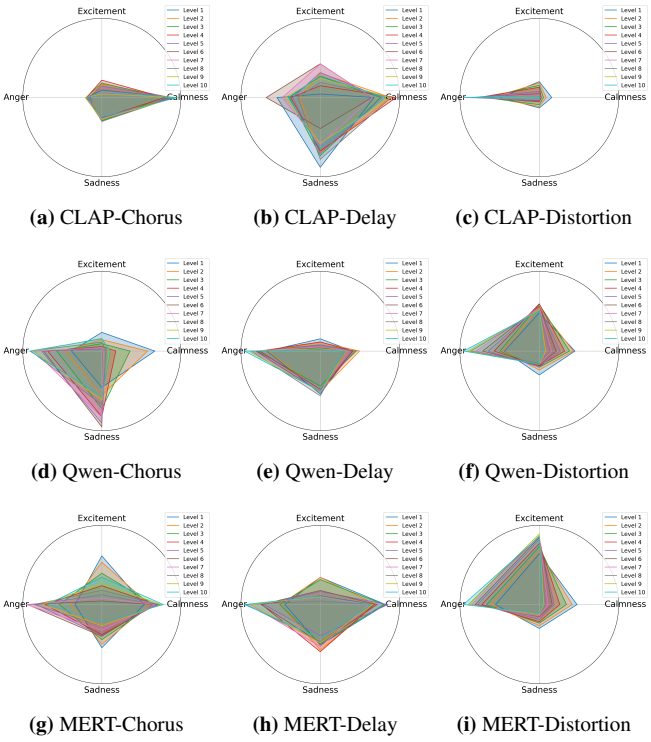
[1] https://github.com/stelioskt/audioFX

Authorized licensed use limited to: National Technical University of Athens (NTUA). Downloaded on May 14,2026 at 06:17:57 UTC from IEEE Xplore. Restrictions apply.

Table 1: Audio FX Δ (Level_{10}-Level_1) across MSE Arousal, MSE Valence, and F1 score. Bold values indicate the largest absolute difference among audio FXs for a specific model and dataset, while underlined values indicate the largest absolute difference among models for a specific audio FX and dataset.

Audio FX	MSE Arousal						MSE Valence						F1 Score Categorical					
	DEAM			witheFlow			DEAM			witheFlow			witheFlow			EMOPIA		
	Qwen	MERT	CLAP	Qwen	MERT	CLAP	Qwen	MERT	CLAP	Qwen	MERT	CLAP	Qwen	MERT	CLAP	Qwen	MERT	CLAP
Chorus	+0.003	+0.008	-0.001	+0.060	+0.110	+0.029	+0.007	+0.004	-0.001	+0.120	+0.120	+0.070	-0.160	-0.190	-0.090	-0.310	-0.110	-0.070
Distortion	+0.012	+0.023	**+0.028**	+0.160	+0.122	-0.027	+0.009	**+0.017**	**+0.027**	**+0.240**	**+0.210**	**+0.110**	**-0.488**	**-0.433**	**-0.149**	**-0.370**	**-0.390**	**-0.350**
Phaser	+0.006	+0.005	+0.002	**+0.171**	+0.129	-0.040	+0.006	+0.006	+0.001	+0.075	+0.168	+0.030	-0.266	-0.230	-0.041	-0.180	-0.100	-0.020
EQ	+0.007	+0.010	+0.014	+0.051	+0.060	-0.033	+0.005	+0.007	+0.011	+0.050	+0.140	-0.050	-0.103	-0.186	+0.041	+0.040	-0.060	+0.030
Delay	+0.003	+0.004	-0.001	+0.016	+0.113	+0.008	+0.005	+0.005	-0.002	+0.060	+0.060	+0.040	-0.091	-0.147	+0.000	-0.200	-0.090	+0.100
Reverb	**+0.021**	**+0.012**	+0.010	-0.012	**+0.141**	**+0.106**	**+0.020**	+0.015	+0.006	+0.030	+0.060	+0.055	-0.022	-0.150	-0.116	-0.160	-0.230	-0.170

one of four categorical emotions—*Excitement, Anger, Sadness*, or *Calmness*—supporting both clip-level classification and song-level emotion analysis. The *DEAM* dataset comprises 1,802 excerpts and full songs annotated with valence and arousal as continuous, time-varying values per second, as well as aggregated ratings, along with metadata on duration, genre, and folksonomy tags, enabling regression-based modeling of emotion dynamics. The part of the *witheFlow* dataset we used includes 235 solo-instrument recordings, annotated by 20 listeners with continuous valence-arousal values per second following the *Circumplex Model of Affect* [19], as well as multi-label categorical tags based on the *Geneva Emotional Music Scale (GEMS-9)* [20], making it suitable for both dimensional and categorical emotion modeling in music.



(a) CLAP-Chorus **(b)** CLAP-Delay **(c)** CLAP-Distortion

(d) Qwen-Chorus **(e)** Qwen-Delay **(f)** Qwen-Distortion

(g) MERT-Chorus **(h)** MERT-Delay **(i)** MERT-Distortion

Fig. 1: Radar plots of emotion predictions for CLAP, Qwen, and MERT across three audio effects for the *EMOPIA* dataset. Each level in the plots depicts the distribution of emotions, normalised based on the greatest value in a given plot.

To investigate and interpret musical emotion, we employed three state-of-the-art **foundation models**, each offering unique capabilities. First, *MERT-v1-330M* [13] is a 330M-parameter model (24 layers with 1024 hidden units) trained on 160,000 hours of music audio using masked language modeling with RVQ-VAE and CQT teacher signals. MERT operates on 24 kHz audio at 75 frames per second and is particularly suited to capturing fine-grained musical and acoustic emotion cues. Second, *CLAP* [21] is a dual-encoder model with roughly 630M parameters, combining a Transformer-based (HTS-AT) audio encoder with a *RoBERTa*-based text encoder. CLAP was trained on 128K–630K audio–text pairs via contrastive learning, using log-Mel spectrogram inputs at 48 kHz to learn a shared audio–text representation space with strong zero-shot generalization. Finally, *Qwen2-Audio-7B* [22] is a 7B-parameter model featuring a *Whisper*-style audio encoder and *Qwen*-style decoder. Qwen2-Audio-7B was trained on over 300K hours of speech, music, and environmental audio to support broad audio-language tasks, including automatic speech recognition, classification, question answering, and conversational interaction.

For the head classifiers of our foundation models, we employed shallow **interpretable models** tailored to each task and dataset, training them while keeping the foundation model frozen. For regression on *DEAM* and *witheFlow* valence-arousal annotations, we used an *XGBRegressor* [23] to predict continuous emotional dimensions, evaluating performance with MAE, MSE, and R^2. For single-label classification on *EMOPIA*, we applied an *XGBClassifier*, assessing weighted accuracy, precision, recall, and F1-score. For multi-label classification on *witheFlow* categorical tags, we implemented a *OneVsRest* strategy with *XGBClassifier*, measuring F1-micro and F1-macro scores with a threshold of 0.5. This configuration allowed us to effectively adapt gradient-boosted tree models to both dimensional and categorical emotion prediction tasks.

To apply controlled alterations to our input and observe the behavior of our chosen foundation model, we used six **audio FX** from the pedalboard library, each scaled in intensity from 1 to 10. Reverb simulates acoustic space by adjusting *room_size*, creating the perception of different environments from small rooms to large halls, which can influence temporal and spectral characteristics of the signal. Delay introduces echoes through *delay_seconds* and *feedback*, affecting rhythmic perception and reinforcing certain frequencies. Distortion adds harmonic saturation via *drive_db*, enhancing harmonic content and introducing nonlinearities that challenge the model's robustness. EQ shapes the frequency spectrum with *low_cutoff* and *high_cutoff* filters, emphasizing or attenuating spectral bands and altering timbral balance. Chorus thickens the sound by duplicating and slightly

Authorized licensed use limited to: National Technical University of Athens (NTUA). Downloaded on May 14,2026 at 06:17:57 UTC from IEEE Xplore. Restrictions apply.

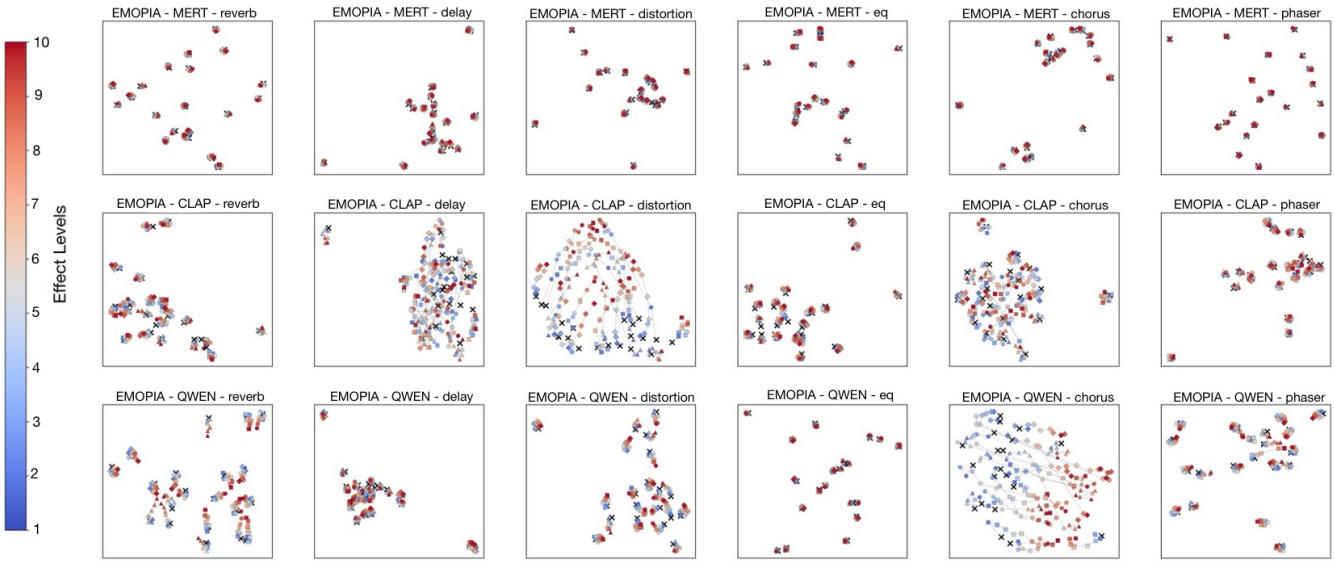


Fig. 2: UMAP visualization of relative distance in the most emotion-discriminative foundation model embeddings for the *EMOPIA* dataset, showing trajectories generated for each input identity after applying audio FX with the intensity ranging from 1 to 10.

detuning the signal with *rate*, *depth*, and *feedback*, producing subtle modulation that can affect perceived pitch and texture. Phaser applies sweeping frequency modulation via *rate*, *depth*, and *feedback*, generating moving notches that alter harmonic relationships over time.

3. EXPERIMENTS

Using the above material, we conducted four experiments to investigate the relationships between audio effects and estimated emotions. First, we analyzed performance changes to assess how audio manipulations impact model accuracy. Second, we examined shifts in emotion predictions to identify changes in predicted emotional labels or dimensions. Third, we explored embedding space alterations to understand how audio effects influence the internal representations learned by the models. Finally, we investigated how real-world audio FX settings affect emotion estimation.

3.1. Impact on Model Performance

Inspired by [10], that studies the impact of audio effects on foundation models, and [24], that examines performance changes in speech emotion recognition models induced by codec processing, we aimed to explore how incorporating emotional information affects performance in emotion recognition tasks. To investigate this, we trained interpretable machine learning classifiers using embeddings extracted from foundation models across all datasets. This approach allowed us to develop models for multi-class classification, multi-label classification, and regression tasks.

After training our interpretable classifiers, we applied various audio effects at different intensity levels and analyzed their impact on model performance. All foundation models were consistently used as backbones across tasks and datasets. We hypothesized that any observed performance degradation would indicate a potential correlation between the audio FX and the estimated emotion.

Overall, we observed a general decrease in performance, indicating that audio effects may influence emotion recognition. As summarized in Table 1, phaser and distortion in particular evoked the largest performance declines across multiple intensity levels. All three models exhibited similar trends, with the decrease becoming more pronounced as the intensity of the audio effects increased.

3.2. Shifts in Emotion Predictions

Previously trained pipelines were used to run inference on audio samples that had been modified with six common effects applied at varying intensities. Transparent classifiers built on different foundation model backbones generated emotion predictions from these altered inputs (see Section 2). The resulting predictions were compared with those from the original samples, and the observed changes were visualized to reveal how the effects shifted predicted emotions. This approach enabled the examination of potential correlations between specific audio manipulations and estimated emotional responses.

In Fig. 1, we examine how emotional predictions in the *EMOPIA* dataset are affected by different audio effects. Notably, high levels of distortion consistently increase *Anger* predictions while decreasing *Calmness* across all models [25]. In contrast, delay and chorus effects introduce greater variability, suggesting that these manipulations create ambiguity in the models' emotional interpretations [26]. Interestingly, increasing the chorus effect tends to boost *Calmness* predictions in the CLAP and MERT models, whereas increasing the delay effect raises *Anger* predictions in the CLAP and Qwen models.

3.3. Embedding Space Alterations

In this section, we present an experiment investigating how the embeddings of a foundation model—representations that encode emotional information—are altered as common audio effects are gradually applied. Inspired by the visualization approach of Deng et

Authorized licensed use limited to: National Technical University of Athens (NTUA). Downloaded on May 14,2026 at 06:17:57 UTC from IEEE Xplore. Restrictions apply.

al. [10], we identified the most important features from our shallow emotion classification model. Using the top 25 features, we performed UMAP [27] dimensionality reduction and visualized the foundation model's embeddings for inputs modified at different effect intensities. This approach allowed us to observe how audio effects reshape the structure of the embeddings and to identify which models are most sensitive to these changes.

To ensure stability and comparability, embeddings were first standardized per (dataset, model) using zero-mean/unit-variance scaling, then cleaned via (i) variance thresholding (removing features with variance $< 1 \times 10^{-6}$) and (ii) iterative correlation pruning (dropping features with $|r| > 0.95$). Feature selection was then performed: for regression-style datasets (*DEAM*, *witheFlow*) we used *ElasticNetCV* with $l1_ratio \in \{0.5, 0.8\}$, α values log-spaced from 10^{-3} to 10^{2} (50 values), 5-fold cross-validation, $\max_iter = 60000$, and $tol = 2 \times 10^{-3}$, while for classification-style datasets (*EMOPIA*, *witheFlow*) we used logistic regression with an elastic net penalty ($l1_ratio = 0.5$), the SAGA solver, $C = 0.5$, $max_iter = 6000$, and $tol = 2 \times 10^{-3}$. For visualization, we applied UMAP using a cosine metric and spectral initialization; in this experiment, these shallow models were chosen for their linearity and interpretability, with $K = 25$, $n_neighbors = 30$, and $min_dist = 0.5$. To avoid crowding and label imbalance, we sampled 20 tracks per effect for regression datasets, 5 tracks per label for classification datasets, and 3 tracks per label for the multi-label dataset (witheFlow), using a fixed random seed (42), covering the effects reverb, delay, distortion, EQ, chorus, and phaser.

In Fig. 2, we show how embeddings from each foundation model respond to audio effects on the *EMOPIA* dataset. Our analysis reveals systematic differences in how foundation models encode the impact of audio effects. CLAP embeddings exhibit large, structured displacements as FX levels increase—particularly for delay, chorus, and distortion—indicating strong sensitivity to timbral manipulations. Qwen also shows noticeable shifts, though with less consistent trajectories, whereas MERT remains relatively stable across all effect levels, suggesting robustness to such manipulations, likely due to its training on music-specific tasks. Overall, trajectory length and variance in embedding space can serve as metrics for quantifying the influence of audio FX on emotion, with the magnitude of this impact depending strongly on the choice of foundation model.

3.4. Real World Scenario

We tested real-world effect chains designed to mimic iconic sounds from Pink Floyd, U2 (emphasizing reverb and delay to create atmospheric textures), and Rage Against the Machine (heavy distortion with moderate chorus) [2] [3]. The goal of this experiment was to examine whether, in real-world scenarios, carefully crafted audio FX chains can produce stronger shifts in emotion predictions due to their intentional artistic design.

As shown in Fig. 3, moving from isolated effects to complete FX chains reveals how cumulative production choices shift embeddings in latent space. MERT and Qwen produce similar trajectories, both showing structured, directional shifts that reflect the cumulative impact of effects. The distortion-heavy chain of Rage Against the Machine yields almost unidirectional trajectories, consistent with strong spectral shaping imposing a uniform transformation. Reverb- and delay-heavy chains of U2 and Pink Floyd also follow a generally consistent pattern, though some divergent paths reflect the more complex ways spatial and temporal effects influence latent representations. CLAP, in contrast, shows shorter, scattered movements, suggesting that its sensitivity is dampened when effects are combined. Overall, real-world effect chains generate larger, more coherent shifts, highlighting that artists naturally select combinations to evoke stronger emotional responses. The nature of these trajectories—unidirectional versus multidirectional—thus serves as a proxy for how different categories of audio effects shape emotional perception in foundation models.

[2] https://www.kitrae.net/music/David_Gilmour_Effects_And_Gear.html

[3] https://www.guitarchalk.com/guitar-amp-settings-guide/#t-1649265629386

4. CONCLUSIONS AND FUTURE WORK

Our study demonstrates that audio effects noticeably alter estimated emotion in music. Distortion, in particular, strongly increase *Anger* while reducing *Calmness*, whereas chorus and delay introduce more variability in predictions. Analysis of embedding-space trajectories shows that the magnitude and structure of latent-space shifts reflect the emotional impact of FX, with real-world effect chains producing the most pronounced and coherent changes, likely because they are intentionally designed by artists to evoke strong emotional responses. Among the models, MERT exhibited relative robustness, while CLAP and Qwen were more sensitive to manipulations.

For future work, we plan to investigate the simultaneous application of multiple audio effects, extend the study to additional datasets and foundation models, explore alternative probing techniques to gain deeper insight into how complex audio transformations influence emotional perception in music, and include comparisons to classical feature-based systems as well as analyses of human ground truth after effects.

Fig. 3: UMAP visualization of relative distance in the most emotion-discriminative foundation model embeddings for the *witheFlow* dataset, showing trajectories generated for each input after applying real-world scenario audio FX.

Authorized licensed use limited to: National Technical University of Athens (NTUA). Downloaded on May 14,2026 at 06:17:57 UTC from IEEE Xplore. Restrictions apply.

5. ACKNOWLEDGMENTS

This work was supported by AWS resources provided by the National Infrastructures for Research and Technology (GRNET), funded by the EU Recovery and Resilience Facility. The research project was implemented within the framework of the H.F.R.I. call "Basic Research Financing (Horizontal Support of All Sciences)" under the National Recovery and Resilience Plan "Greece 2.0," funded by the European Union – NextGenerationEU (H.F.R.I. Project No. 15111: Emotional Artificial Intelligence in Music Expression).

6. REFERENCES

[1] Cyril Laurier, Olivier Lartillot, Tuomas Eerola, and Petri Toiviainen, "Exploring relationships between audio features and emotion in music," in *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music*. Citeseer, 2009, pp. 260–264.

[2] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva, "Audio features for music emotion recognition: a survey," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 68–88, 2020.

[3] Vassilis Lyberatos, Spyridon Kantarelis, Edmund Dervakos, and Giorgos Stamou, "Perceptual musical features for interpretable audio tagging," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 878–882.

[4] Vassilis Lyberatos, Spyridon Kantarelis, Edmund Dervakos, and Giorgos Stamou, "Challenges and perspectives in interpretable music auto-tagging using perceptual features," *IEEE Access*, vol. 13, pp. 60720–60732, 2025.

[5] Vassilis Lyberatos, Spyridon Kantarelis, Ioanna Zioga, Christina Anagnostopoulou, Giorgos Stamou, and Anastasia Georgaki, "Music interpretation and emotion perception: A computational and neurophysiological investigation," 2025.

[6] Zhihui Han, Jian Kang, and Qi Meng, "Effect of sound sequence on soundscape emotions," *Applied Acoustics*, vol. 207, pp. 109371, 2023.

[7] Hannah Wilkie and Peter Harrison, "Reverberation time and musical emotion in recorded music listening," *Music Perception: An Interdisciplinary Journal*, vol. 42, no. 4, pp. 329–344, 2025.

[8] Klaus R Scherer, Johan Sundberg, Lucas Tamarit, and Gláucia L Salomão, "Comparing the acoustic expression of emotion in the speaking and the singing voice," *Computer Speech & Language*, vol. 29, no. 1, pp. 218–235, 2015.

[9] Wenyi Song, Anh-Dung Dinh, and Andrew Brian Horner, "The emotional characteristics of the violin with different pitches, dynamics, and vibrato," in *Proceedings of Meetings on Acoustics*. Acoustical Society of America, 2024, vol. 55, p. 035004.

[10] Victor Deng, Changhong Wang, Gael Richard, and Brian McFee, "Investigating the sensitivity of pre-trained audio embeddings to common effects," in *ICASSP 2025*. IEEE, 2025, pp. 1–5.

[11] Jui Shah, Yaman Kumar Singla, Changyou Chen, and Rajiv Ratn Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[12] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, "Musiclm: Generating music from text," 2023.

[13] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al., "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.

[14] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al., "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[15] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "Clap: Learning audio concepts from natural language supervision," in *Proceedings of the 2022 International Conference on Machine Learning (ICML)*, 2022.

[16] Alican Akman, Qiyang Sun, and Björn W Schuller, "Audio explanation synthesis with generative foundation models," in *ICASSP 2025*. IEEE, 2025, pp. 1–5.

[17] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang, "Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," *arXiv preprint arXiv:2108.01374*, 2021.

[18] Anna Alajanki, Yi-Hsuan Yang, and Mohammad Soleymani, "Benchmarking music emotion recognition systems," *PloS one*, pp. 835–838, 2016.

[19] James A Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.

[20] Peer-Ole Jacobsen, Hannah Strauss, Julia Vigl, Eva Zangerle, and Marcel Zentner, "Assessing aesthetic music-evoked emotions in a minute or less: A comparison of the gems-45 and the gems-9," *Musicae Scientiae*, p. 10298649241256252, 2024.

[21] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023*. IEEE, 2023, pp. 1–5.

[22] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al., "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[23] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[24] Aaron Albin and Elliot Moore, "Objective study of the performance degradation in emotion recognition through the amr-wb+ codec.," in *INTERSPEECH*, 2015, pp. 1319–1323.

[25] Jinxing Gao, Diqun Yan, and Mingyu Dong, "Black-box adversarial attacks through speech distortion for speech emotion recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 20, 2022.

[26] Valerio Cesarini and Giovanni Costantini, "Reverb and noise as real-world effects in speech recognition models: A study and a proposal of a feature set," *Applied Sciences*, vol. 14, no. 23, pp. 11446, 2024.

[27] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

Authorized licensed use limited to: National Technical University of Athens (NTUA). Downloaded on May 14,2026 at 06:17:57 UTC from IEEE Xplore. Restrictions apply.