Improving Perceptual Audio Aesthetic Assessment via Triplet Loss and Self-Supervised Embeddings

Dyah A. M. G. Wisnu[1,2,3], Ryandhimas E. Zezario[1], Stefano Rini[4], Hsin-Min Wang[3], and Yu Tsao[1]

[1]Research Center for Information Technology Innovation, Academia Sinica
[2]College of Informatics, National Chengchi University
[3]Institute of Information Science, Academia Sinica
[4]Institute of Communications Engineering, National Yang Ming Chiao Tung University

***Abstract*—We present a system for automatic multi-axis perceptual quality prediction of generative audio, developed for Track 2 of the AudioMOS Challenge 2025. The task is to predict four Audio Aesthetic Scores—Production Quality, Production Complexity, Content Enjoyment, and Content Usefulness—for audio generated by text-to-speech (TTS), text-to-audio (TTA), and text-to-music (TTM) systems. A main challenge is the domain shift between natural training data and synthetic evaluation data. To address this, we combine BEATs, a pretrained transformer-based audio representation model, with a multi-branch long short-term memory (LSTM) predictor and use a triplet loss with buffer-based sampling to structure the embedding space by perceptual similarity. Our results show that this improves embedding discriminability and generalization, enabling domain-robust audio quality assessment without synthetic training data.**

***Index Terms*—audio-aesthetic, audio-assessment, BEATs, triplet-loss, synthetic-data.**

I. Introduction

The AudioMOS Challenge 2025 advances automatic evaluation for audio generation, with Track 2 targeting Audiobox Aesthetic Scores across four perceptual axes: Production Quality, Production Complexity, Content Enjoyment, and Content Usefulness. Participants develop models to predict these scores for TTS, TTA, and TTM systems. Building on the VoiceMOS Challenges [1]–[3], this edition extends beyond speech to broader audio content. Recent deep-learning-based non-intrusive evaluation methods [4]–[12] increasingly leverage large-scale audio pre-trained models—such as self-supervised learning (SSL) models [13], [14], Whisper [6], and speech language models [15]—to extract robust acoustic representations for improved generalization.

The challenge 2025 edition expands the scope to include music and general audio, while introducing greater domain variability. A key challenge in this track lies in the domain shift: the training and development sets consist of natural audio, while the evaluation set is composed entirely of synthetic samples. To address this mismatch, we propose a system that leverages BEATs [16], a transformer-based audio representation model, in combination with a multi-branch architecture and a triplet loss with buffer-based sampling strategy. The triplet loss enforces a perceptually meaningful structure in the embedding space by pulling together audio samples with similar aesthetic scores and pushing apart those with dissimilar scores, improving robustness to unseen generative content.

II. System Description

A. Dataset

We used the AES-natural dataset provided by the AudioMOS Challenge 2025, which includes 2,950 natural audio samples from three domains:

- Speech (950 samples from EARS, LibriTTS, and CommonVoice 13.0),
- Music (1,000 samples from MUSDB18-HQ and MusicCaps),
- General audio (1,000 samples from a subset of AudioSet).

Each audio sample is rated by 10 expert listeners across four perceptual axes: PQ, PC, CE, and CU. We use the mean of the listener scores, normalized to zero mean and unit variance, as training targets. While modeling individual listener tendencies (e.g., as in MBNet [17] or LDNet [18]) has been shown to improve performance, we focus on the mean score for simplicity and to align with the challenge's evaluation protocol. No listener-specific information or demographics were used. The data was split into 2,700 training and 250 validation samples. The evaluation set consists of 3,000 unseen synthetic samples from TTS, TTA, and TTM systems, which are not accessible during training.

B. Preprocessing and Feature Extraction

All audio waveforms were resampled to 16 kHz. No additional preprocessing such as volume normalization or silence trimming was applied to preserve the original audio content. We use BEATs [16] (Bidirectional Encoder representation from Audio Transformers) as our self-supervised feature extractor to maintain a unified representation for both speech and non-speech audio in the evaluation data. BEATs was pretrained on AudioSet, a large-scale dataset of natural audio events, and outputs high-level representations from multiple transformer layers. Although speech-specific SSL models such as WavLM may offer more efficient representations for speech segments, integrating them with BEATs features is left as future work.

We obtain the hidden states from all BEATs layers and apply a learnable weighted sum, where each layer is scaled by a learnable scalar and the scalars are normalized via a softmax

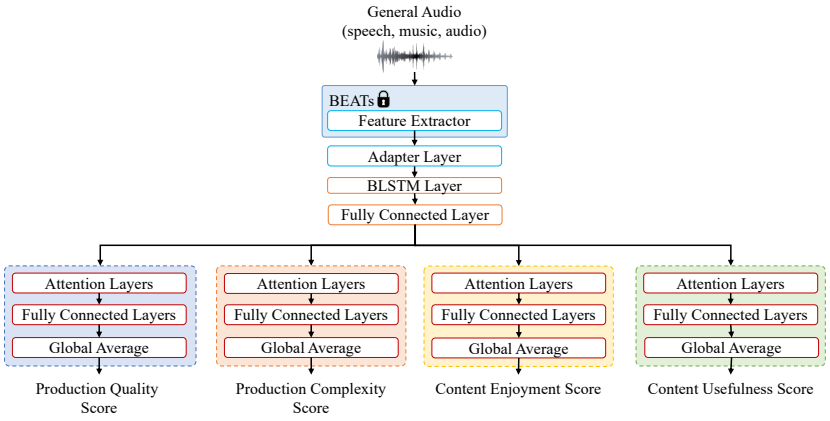


Fig. 1. System Architecture of AESA-Net

so that they sum to one. This weighted combination forms the input feature sequence for downstream modeling.

C. Model Architecture

We propose a unified multi-task architecture called **AESA-Net** (Audio Aesthetics Assessment Network), designed to jointly predict all four perceptual axes. The overall structure, inspired by HAAQI-Net [19], is illustrated in Figure 1. The model consists of a shared backbone for feature extraction and encoding, followed by task-specific heads for each aesthetic axis.

- Adapter Layer: A linear layer projects BEATs embeddings to a lower-dimensional space (linear output).
- BLSTM Layer: A two-layer bidirectional LSTM captures temporal dependencies from the projected features.
- Shared Linear Layer: Outputs from BLSTM are transformed through a shared fully connected layer followed by dropout and activation.

We then branch into four task-specific modules:

- Perceptual Axes (PQ, PC, CE, CU).
  Each axis has its own:
  - Multi-head self-attention layer (to capture content-relevant patterns),
  - Frame-level scoring layer (linear + sigmoid),
  - Adaptive average pooling (to aggregate clip-level scores).

The final outputs are four values in the [0, 1] range, representing the predicted scores for each axis. In addition, frame-level scores are available for auxiliary supervision.

D. Loss Function

We optimize a combination of mean squared error (MSE) loss and triplet loss with buffer sampling.

III. Triplet Loss with Buffer Sampling Strategy

To improve the generalization of the model to unseen (synthetic) audio samples, we apply *triplet loss* on the intermediate embeddings using a *memory buffer*. This encourages the model to learn a perceptually structured latent space, where audio samples with similar perceptual scores are close together, and dissimilar ones are pushed farther apart.

Let:

- z_a: anchor embedding from the current sample
- z_p: positive embedding (similar perceptual score)
- z_n: negative embedding (dissimilar perceptual score)

We maintain a buffer $\mathcal{B}$ that stores recent embedding-score pairs (z_i, y_i), where y_i is the normalized ground-truth perceptual score (e.g., for PQ). For buffer-based sampling, we adopt the strategy proposed in [20].

The training process proceeds as follows:

1) **Buffer Update:**
   After processing a training sample, its embedding z and corresponding target y are added to the buffer $\mathcal{B}$. The buffer has a fixed capacity and behaves as a queue (FIFO), where older entries are replaced when full.
2) **Triplet Sampling:**
   We follow [21], given an anchor (z_a, y_a):
   - A positive example (z_p, y_p) is sampled from $\mathcal{B}$ such that $|y_a - y_p| < \epsilon$
   - A negative example (z_n, y_n) is sampled such that $|y_a - y_n| > \epsilon$

   If both are found, we compute the triplet loss:

$$\mathcal{L}_{\text{Triplet}} = \max\left(\|z_a - z_p\|_2^2 - \|z_a - z_n\|_2^2 + \text{margin},\ 0\right) \tag{1}$$

3) **Prediction Loss (MSE):**
   The model also minimizes the mean squared error (MSE) between the predicted and ground truth perceptual scores:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{2}$$

The total loss used for backpropagation is a combination of the MSE loss and triplet loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \alpha \cdot \mathcal{L}_{\text{Triplet}} \tag{3}$$

where α is a hyperparameter (e.g., $\alpha = 0.2$) controlling the contribution of the triplet loss to the total loss.

TABLE I
SYSTEM-LEVEL RESULTS ON THE OFFICIAL EVALUATION SET ACROSS FOUR PERCEPTUAL AXES AND THREE DOMAINS.

Domain	Axis	MSE	LCC	SRCC	KTAU
Speech	PQ	0.5138	0.6950	0.7182	0.4909
	PC	0.0060	0.7784	0.5545	0.4182
	CU	0.9539	0.6520	0.7273	0.5273
	CE	1.4739	0.6726	0.7182	0.4909
Music	PQ	0.7242	0.8775	0.6703	0.4872
	PC	0.4843	0.8934	0.8956	0.7436
	CU	1.1404	0.8736	0.7198	0.5385
	CE	2.3145	0.8405	0.8693	0.6839
Audio	PQ	0.6514	0.8662	0.8811	0.7576
	PC	0.0635	0.8264	0.6154	0.4545
	CU	1.9407	0.8380	0.8322	0.6970
	CE	1.8942	0.8133	0.7483	0.5455

TABLE II
COMPARISON WITH THE OFFICIAL BASELINE (SYSTEM-LEVEL)

Axis	Metric	Baseline	Ours	Best
PQ	MSE	0.632	**0.635**	↓
	LCC	**0.903**	0.897	
	SRCC	0.866	**0.896**	↑
	KTAU	0.683	**0.737**	↑
PC	MSE	0.226	**0.198**	↓
	LCC	0.963	**0.984**	↑
	SRCC	**0.934**	0.928	
	KTAU	**0.800**	0.781	
CE	MSE	**1.142**	3.991	
	LCC	0.921	**0.950**	↑
	SRCC	0.841	**0.904**	↑
	KTAU	0.655	**0.741**	↑
CU	MSE	1.478	**0.533**	↓
	LCC	0.879	**0.897**	↑
	SRCC	0.810	**0.894**	↑
	KTAU	0.629	**0.724**	↑

IV. EXPERIMENTS AND RESULTS

A. Training Setup

The model was optimized using the Adam optimizer with an initial learning rate of 1e-4, and trained for up to 100 epochs with early stopping based on validation MSE. The batch size was set to 1 due to variable-length audio input and GPU memory constraints.

B. Evaluation Metrics

We evaluated the model using the official challenge metrics:

- **MSE (Mean Squared Error)**: Measures the squared difference between predicted and ground truth scores for each perceptual axis.
- **SRCC (Spearman Rank Correlation Coefficient)**: Evaluates rank-based monotonic correlation between predicted and reference scores.
- **PCC (Pearson Correlation Coefficient)**: Assesses linear correlation, also referred to as LCC in some literature.
- **KTAU (Kendall's Tau)**: Measures ordinal association between predicted and ground truth rankings, offering a more robust alternative to SRCC in noisy scenarios.

These were computed at the **system-level** by averaging predictions and scores per sample

C. Results

The proposed model was evaluated on the official AudioMOS Challenge 2025 evaluation set, which consists of synthetic audio samples generated from TTS, TTA, and TTM systems. Despite being trained solely on natural audio, the model was tested on three synthetic domains: speech, music, and general audio. We report the system-level performance on all four Audiobox Aesthetic Score axes—Production Quality (PQ), Production Complexity (PC), Content Usefulness (CU), and Content Enjoyment (CE)—using four standard metrics: Mean Squared Error (MSE), Linear Correlation Coefficient (LCC), Spearman Rank Correlation Coefficient (SRCC), and Kendall's Tau (KTAU). **The complete evaluation results are summarized in Table I.**

The results indicate that the model generalizes well to synthetic audio, particularly in the music and general audio domains. Notably, high SRCC and KTAU values are observed for PC, suggesting strong rank-based consistency. We attribute part of this robustness to the use of triplet loss, which encourages a semantically structured embedding space and mitigates the domain gap between natural and synthetic data. Similarly, CU and CE also exhibit competitive correlation scores across domains, further demonstrating the model's ability to maintain reliable perceptual predictions across a range of content types.

To further validate our approach, we compare our model against the official baseline system [22] provided by the challenge organizers. As shown in Table II, our model achieves higher rank-based metrics (SRCC and KTAU) across all four perceptual axes, indicating improved ordinal alignment with human ratings. In particular, PQ and CU show marked improvements in SRCC and KTAU, likely benefiting from the triplet loss mechanism. Although the MSE for CE is higher than the baseline, our model yields better correlation scores (LCC, SRCC, KTAU), suggesting more accurate ranking of perceptual quality, even if the absolute values deviate. These results confirm the effectiveness of our model design and its generalization capability to unseen generative content.

V. CONCLUSION

In this work, we proposed a multi-axis perceptual quality prediction model that integrates triplet loss with a buffer-based sampling strategy to improve generalization. The model was trained solely on natural audio data but evaluated on unseen synthetic audio in the AudioMOS 2025 Challenge. Experimental results on the evaluation set show that the inclusion of triplet loss enhances both rank-based and correlation-based performance metrics. The learned embedding space exhibits improved structure and consistency across various audio domains, enabling more robust predictions even in the absence of synthetic training data.

REFERENCES

[1] W.-C. Huang and E. Cooper and Y. Tsao and H.-M. Wang and T. Toda and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4536–4540.

[2] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The voicemos challenge 2023: Zero-shot subjective speech quality prediction for multiple domains," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[3] W.-C. Huang, S.-W. Fu, E. Cooper, R. E. Zezario, T. Toda, H.-M. Wang, J. Yamagishi, and Y. Tsao, "The voicemos challenge 2024: Beyond speech quality prediction," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 803–810.

[4] Z. Yang, W. Zhou, C. Chu, S. Li, R. Dabre, R. Rubino, and Y. Zhao, "Fusion of self-supervised learned models for MOS prediction," in *Proc. Interspeech*, 2022, pp. 5443–5447.

[5] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4521–4525.

[6] R. E. Zezario, Y.-W. Chen, S.-W. Fu, Y. Tsao, H.-M. Wang, and C.-S. Fuh, "A study on incorporating Whisper for robust speech assessment," in *Proc. International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.

[7] P. Manocha, D. Williamson, and A. Finkelstein, "Corn: Co-trained full- and no-reference speech quality assessment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 376–380.

[8] S. Wang, W. Yu, Y. Yang, C. Tang, Y. Li, J. Zhuang, X. Chen, X. Tian, J. Zhang, G. Sun, L. Lu, and C. Zhang, "Enabling auditory large language models for automatic speech quality evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2025, pp. 1–5.

[9] S. Cuervo and R. Marxer, "Speech foundation models on intelligibility prediction for hearing-impaired listeners," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 1421–1425.

[10] R. Mogridge, G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, and A. Ragni, "Non-intrusive speech intelligibility prediction for hearing-impaired users using intermediate ASR features and human memory models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 306–310.

[11] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "A review on subjective and objective evaluation of synthetic speech," *Acoustical Science and Technology*, pp. e24–12, 2024.

[12] S.-W. Fu, K.-H. Hung, Y. Tsao, and Y.-C. F. Wang, "Self-supervised speech quality estimation and enhancement using only clean speech," in *Proc. International Conference on Learning Representations (ICLR)*, 2024.

[13] E. Cooper, W.-H. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022.

[14] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2023.

[15] S. Maiti, Y. Peng, T. Saeki, and S. Watanabe, "Speechlmscore: Evaluating speech generation using speech language model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP) 2023*, 2023, pp. 1–5.

[16] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. International Conference on Machine Learning (ICML)*, Honolulu, Hawaii, USA, 2023, pp. 5178–5193.

[17] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNet: MOS prediction for synthesized speech with mean-bias network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 391–395.

[18] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 896–900.

[19] D. Wisnu, R. E. Zezario, S. Rini, H.-M. Wang, and Y. Tsao, "HAAQI-Net: A non-intrusive neural music audio quality assessment model for hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 1877–1892, 2025.

[20] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[21] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3733–3742.

[22] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov *et al.*, "Meta audiobox aesthetics: Unified automatic quality assessment for speech," arXiv preprint arXiv:2502.05139, 2025.