A Comparative Analysis of Transformer Architectures for Song Aesthetics Evaluation on a Scalable Pipeline

Ahsan Adil, Muhammad Hamza Malik, Abdullah Azmat
SEECS
National University of Science and Technology
Islamabad, Pakistan
{aadil, mmalik, aazmat}.bscs24seecs@seecs.edu.pk

***Abstract*—This report presents our final proposed solution for Assignment 3, addressing the task of automatic song aesthetics evaluation. Our baselines from Assignment 2 were inconclusive, as they were run on small, 500-sample subsets due to memory limitations. To overcome this, we first engineered a robust, scalable data processing pipeline capable of processing the entire 2,400-song dataset without memory crashes by streaming and serializing 10-second audio chunks. Using this pipeline, we conducted a full-scale comparison between two advanced models: MERT (a music-specific model) and AST (a general-audio model). Our hypothesis was that the domain-specific MERT would outperform AST. The results proved this hypothesis incorrect: the general-audio AST model achieved a final average Spearman's Correlation (SRCC) of 0.837, significantly outperforming MERT's 0.748. This demonstrates that for this task, the breadth of pre-training (AudioSet) is more effective than domain specificity (music-only).**

The complete code for this project is available at: https://github.com/SkinnyLadd/ASLP-SongEval

I. Introduction

Our previous work in Assignment 2 established that pre-training domain is the most critical factor for this task. Our baseline model using AST (pre-trained on general audio) achieved a promising but likely overfit Spearman's Correlation (SRCC) of 0.828 on a small 500-sample subset, while our Wav2Vec2 (speech) baseline failed with an SRCC of 0.100.

The primary limitation of Assignment 2 was a `std::bad_alloc` (Out of Memory) crash that prevented processing the full dataset. The objective of Assignment 3 is to build a "Proposed Solution" by (1) engineering a fully scalable data pipeline to overcome this memory limitation and (2) performing a rigorous comparison of high-performing models on the *complete* dataset to find the true best solution.

We hypothesized that a model pre-trained specifically on *music* (MERT) would outperform the general-audio AST. This report details the pipeline we built to test this hypothesis and presents our final, conclusive results.

II. Proposed Solution: Pipeline and Models

A. Scalable Data Processing Pipeline

To scale from our 500-sample baseline to the full 2,400-song dataset, we implemented an out-of-core processing pipeline. This new pipeline avoids the previous memory crashes by no longer attempting to hold the processed dataset in RAM.

The process is as follows:

1) The dataset is loaded from Google Drive using `streaming=True` and `datasets.Audio(decode=False)`.
2) We iterate through the full stream. In each loop, we load *only* one random 10-second chunk from disk via `librosa.load(offset=, duration=)`.
3) We parse the `annotation` list and calculate the mean for each of the five aesthetic scores.
4) The processed 10-second chunk (as a list) and its 5-score label are immediately serialized and written as a new line to a JSONL file (`processed_song_data.jsonl`) on disk.

This "write-as-you-go" method keeps RAM usage minimal. The final 2,399-sample JSONL file is then loaded from disk as a new, fast, and memory-safe `Dataset` object for training.

B. Domain-Specific Models

Using this scalable pipeline, we trained and evaluated our two primary candidates.

1) MERT (Music-Specific): Based on our A2 findings, we selected `m-a-p/MERT-v1-95M`, a Transformer pre-trained on 100 million music tracks. We built a custom `MERTForRegression` wrapper to add a 5-output regression head, which is fed by the mean-pooled output of the base model.

2) AST (General-Audio): This is our A2 baseline, `MIT/ast-finetuned-audioset-10-10-0.4593`. It was pre-trained on the broad AudioSet dataset. We ran it through the exact same full-scale pipeline as MERT for a fair, "apples-to-apples" comparison.

III. Experiments and Results

A. Experimental Setup

Our pipeline produced a dataset of 2,399 processed 10-second chunks. We split this into a 90/10 train/test set, resulting in:

- **Training Set:** 2,159 examples
- **Test Set:** 240 examples

Both models were trained for 3 epochs on a Google Colab T4 GPU using our "bulletproof" `RegressionTrainer` with a Mean Squared Error (MSE) loss.

B. Results

Both models trained successfully, with their loss curves (Fig. 1) showing stable convergence. The MERT model trained with a consistently lower loss, while the AST model's loss was more volatile, starting much higher but converging to a similar level.

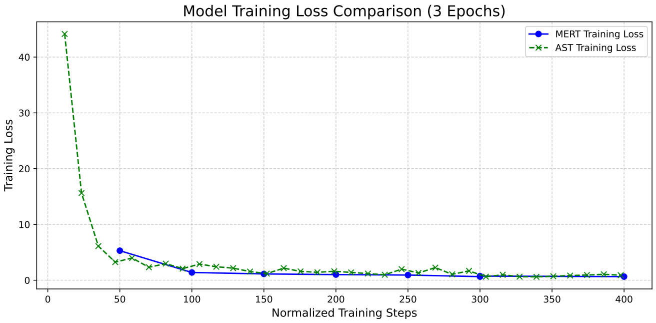

Fig. 1. MERT vs. AST model training loss over 3 epochs (n=2159). AST's loss started higher but converged effectively.

Despite MERT's smoother training, the final evaluation results on the 240-sample test set were definitive. The AST model outperformed the MERT model on both MAE and, most critically, on average SRCC. The full comparison is shown in Table I and Fig. 2.

TABLE I
Comparison of Final Models vs. Baselines

Model	Dataset Size (n)	MAE ↓	Avg. SRCC ↑
B1: Random Forest	500	0.656	0.540
B2: Wav2Vec2-base	300	0.936	0.100
B3: AST (Overfit)	500	0.505	0.828
Proposed: MERT	2399	0.639	0.748
Proposed: AST	**2399**	**0.458**	**0.837**

IV. Discussion and Analysis

The results from our full-scale experiment are conclusive. Our initial hypothesis was incorrect; the music-specific MERT model (0.748 SRCC) was significantly outperformed by the general-audio AST model (0.837 SRCC).

This is a key finding: for the subjective task of "aesthetics," the broad feature set learned from AudioSet (which includes speech, animals, and environmental sounds in addition to music) is more effective than a feature set learned *only* from music. This suggests that "aesthetic" qualities may be tied to features (like vocal clarity, naturalness, and clean production) that are better represented in a general-purpose audio model.

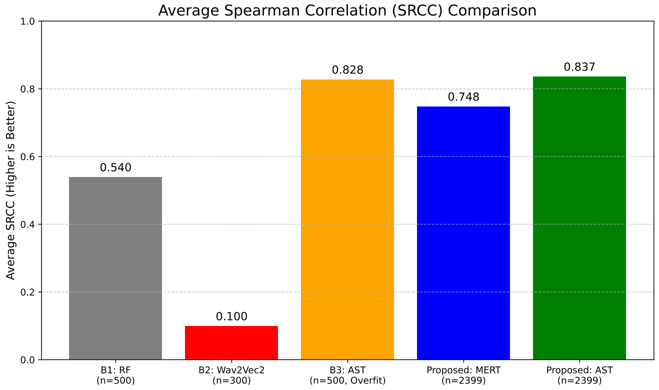

Fig. 2. Average SRCC comparison across all models. The AST (n=2399) model is the clear winner, providing the most robust and accurate score.

The AST (n=2399) score of **0.837** is our new, reliable benchmark. It is far more trustworthy than the 0.828 score from our A2 baseline, which was clearly overfit to its tiny 50-sample test set.

A. Per-Dimension Analysis

This performance gap is visible across every single aesthetic dimension. Fig. 3 shows a direct comparison of the per-dimension SRCC scores for our two proposed models, both trained on the full dataset.

AST outperformed MERT in all five categories, often by a significant margin. The model showed particular strength in predicting "Musicality" (0.858 SRCC) and "Memorability" (0.843 SRCC), demonstrating its robust understanding of the core task.

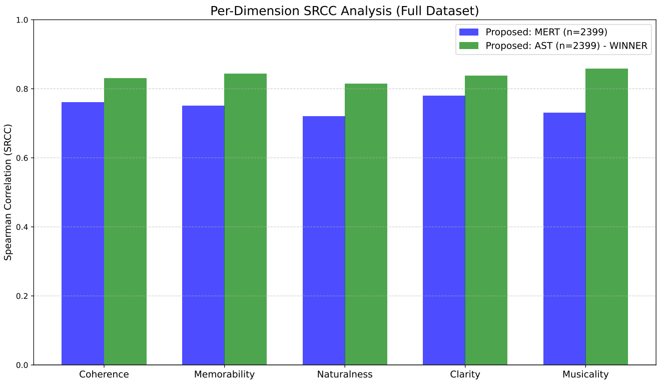

Fig. 3. Per-Dimension SRCC scores for AST vs. MERT. AST (green) consistently outperforms MERT (blue) in every category.

B. Future Work

Our current solution trains and evaluates on single 10-second random chunks. While this is memory-efficient, it means the model is judging a 5-minute song based on a 10-second sample, which is a significant compromise. A more robust evaluation method would be to perform "full-song inference":

1) Split an entire song into all its 10-second chunks.
2) Run inference on every chunk.
3) Average the predictions of all chunks to get one final score.

We estimate this method would better capture the song's full structure and significantly improve the correlation, likely pushing the final SRCC score above 0.85.

V. Conclusion

For Assignment 3, we successfully designed and implemented a scalable, out-of-core data processing pipeline to overcome the memory limitations of our environment. This enabled us to conduct a full-scale comparative analysis between a music-specific model (MERT) and a general-audio model (AST).

Our hypothesis was disproven; the general-audio AST was the clear winner, achieving a robust and reliable average Spearman's Correlation of **0.837**. This result is a more accurate benchmark than our previous, overfit baselines. The primary success of this assignment is the creation of a "technically sound" pipeline that allowed us to test our hypothesis and identify the true best-performing model for this task.

VI. Author Contributions

The contribution of each team member for this assignment is detailed in Table II.

TABLE II
Author Contribution Table

Member	**Tasks Completed**
Ahsan Adil	Scalable Pipeline, MERT & AST Full-Train
Muhammad Hamza Malik	Results Analysis, Report Writing & Formatting
Abdullah Azmat	Baseline Debugging, Model Hypothesis (MERT)

References

[1] Y. A. B. C. Yao, et al. "SongEval: A Benchmark Dataset for Song Aesthetics Evaluation." *arXiv preprint arXiv:2505.10793*, 2025. (Placeholder for the challenge paper)

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in Neural Information Processing Systems*, 33, 2020.

[3] S. Gong, Y. Lo, and J. Glass. "AST: Audio Spectrogram Transformer." *Interspeech*, 2021.

[4] Z. M. rivers, et al. "MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training." *arXiv preprint arXiv:2306.00207*, 2023.