

Oregon State University

Time is Not on Your Side: Assessing the Value of Dynamic Audio Features for Static Music Emotion Labels

Shaun Leib

April 28, 2026

soundbendor lab

Introduction

Music Emotion Recognition (MER)

- Computational methods to identify emotional responses from music
- Combines several disciplines
- Bridge between emotional experiences and machine understanding
- Music recommendation systems, mood based playlist generation, affective music therapy, generative music composition

Russell's Valence and Arousal Circumplex

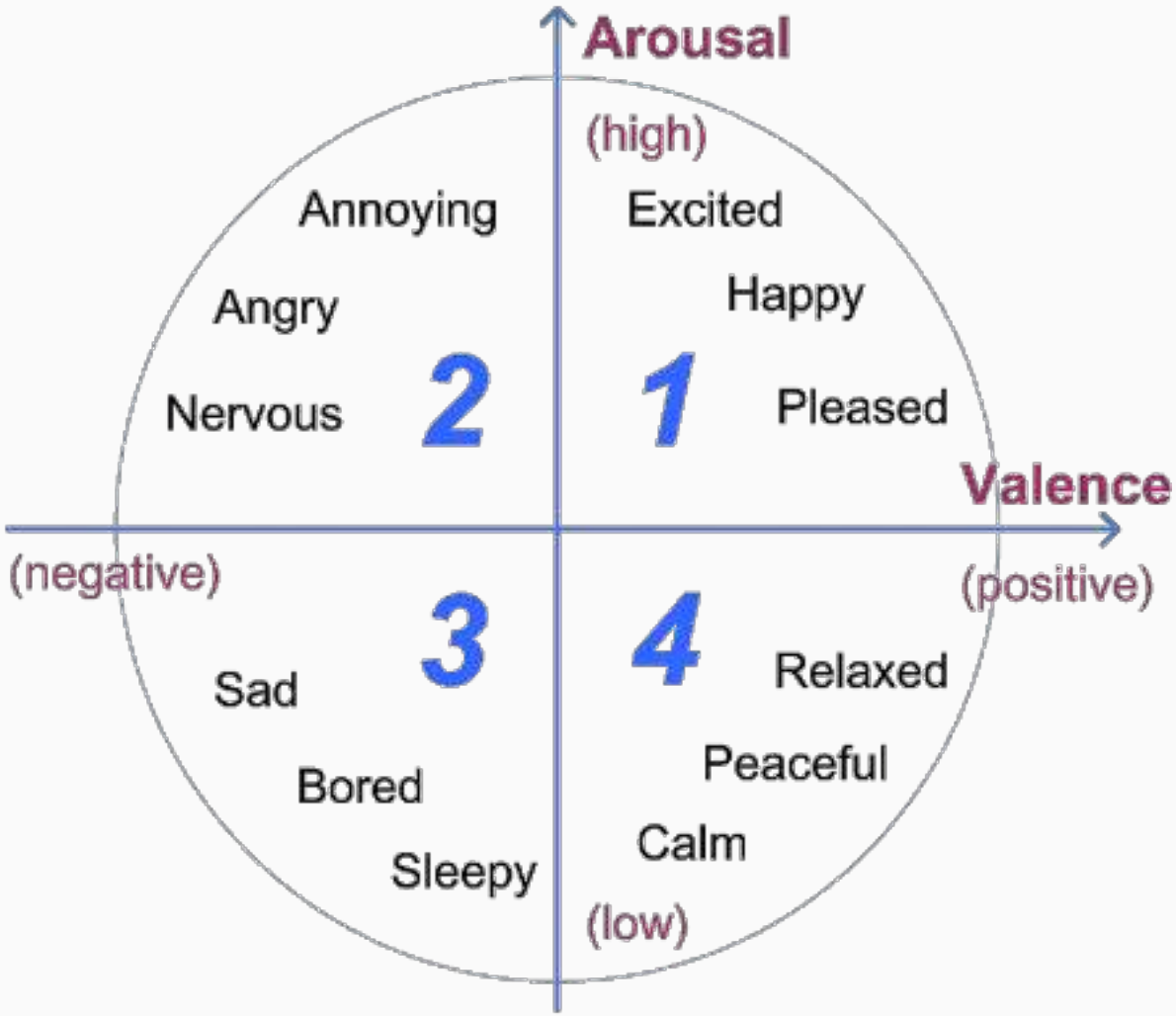


Figure: Russel's Valence and Arousal Circumplex [14].

Related Works

DEAM

- MediaEval 2013 - 2015
- Valence and Arousal static and dynamic regression
- Low-level audio features and spectrograms
- Highly imbalanced and high disagreement between annotators
- BLSTM and CRNN models outperform other models [5, 6, 12]
- Attention pooling before classification helps performance [11]

Related Works

DEAM

- MediaEval 2013 - 2015
- Valence and Arousal static and dynamic regression
- Low-level audio features and spectrograms
- Highly imbalanced and high disagreement between annotators
- BLSTM and CRNN models outperform other models [5, 6, 12]
- Attention pooling before classification helps performance [11]

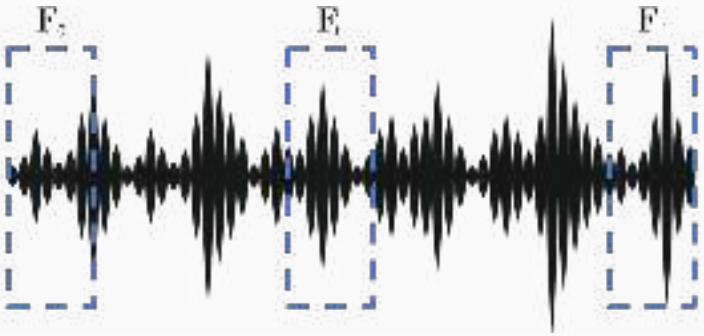
Issues in MER

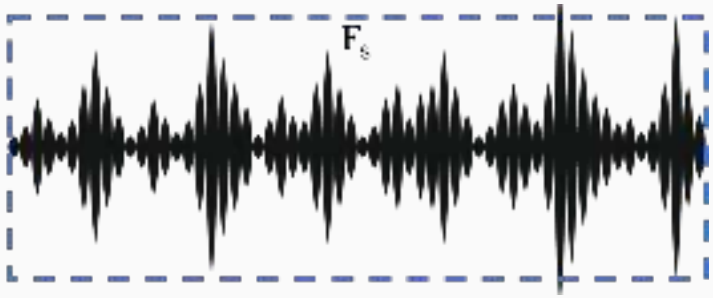
- Subjectivity of human emotion [15]
- Manual annotation is expensive, time consuming, and unreliable [17]
- Music changes across its duration and sections

Task

Hypothesis

Dynamic features will outperform static equivalent counterparts

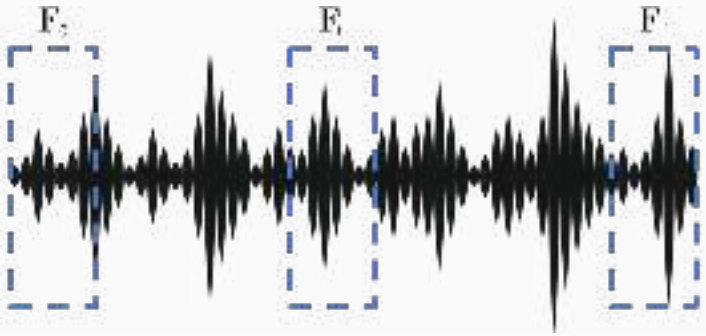


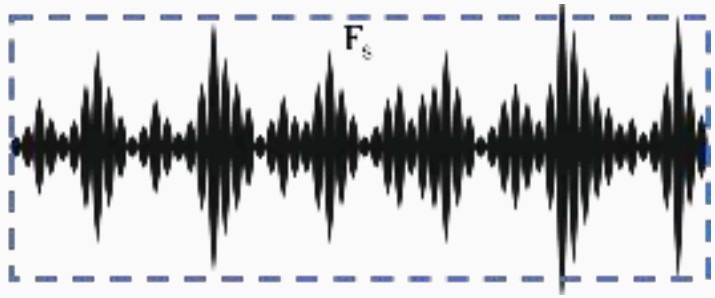


Task

Hypothesis

Dynamic features will outperform static equivalent counterparts





Contributions

- Handcrafted and pre-trained audio features
- First experiments with dynamic features on MERGE
- First experiments exploring fusion of pre-trained audio models
- State-of-the-art regression scores

Dataset

Dataset

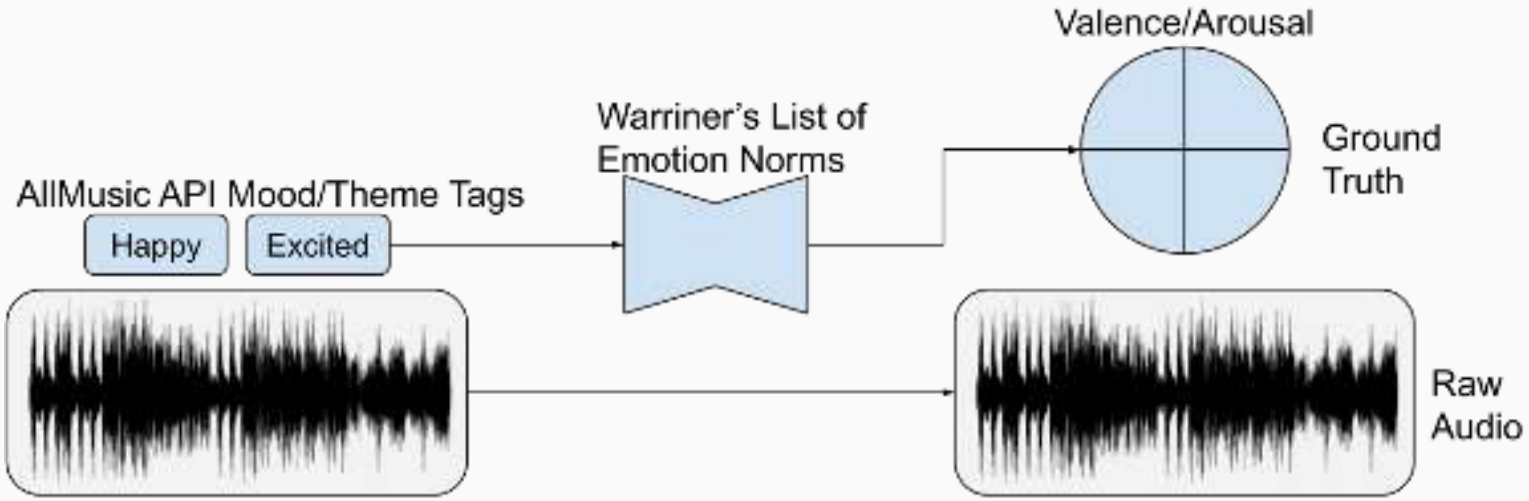
Dataset	Q1	Q2	Q3	Q4	Total
Audio Complete	875	915	808	956	3554
Audio Balanced	808	808	808	808	3232
Bimodal Complete	525	673	500	518	2216
Bimodal Balanced	500	500	500	500	2000

Table: The distribution of audio samples between the different versions of the MERGE dataset. Recreated from [10].

MERGE

- **M**usic **E**motion **R**ecognition next **GE**neration (MERGE)
- Developed by Louro et al. to improve upon previous datasets
- Relatively large, psychologically backed taxonomy, human validation quality controlled, audio provided, collective baselines

Dataset Creation



MERGE

- Automatic labeling: AllMusic API $\rightarrow$ Warriner's adjective list to map tags into regression values in V/A
- Human validation and quality control

Dataset Distribution

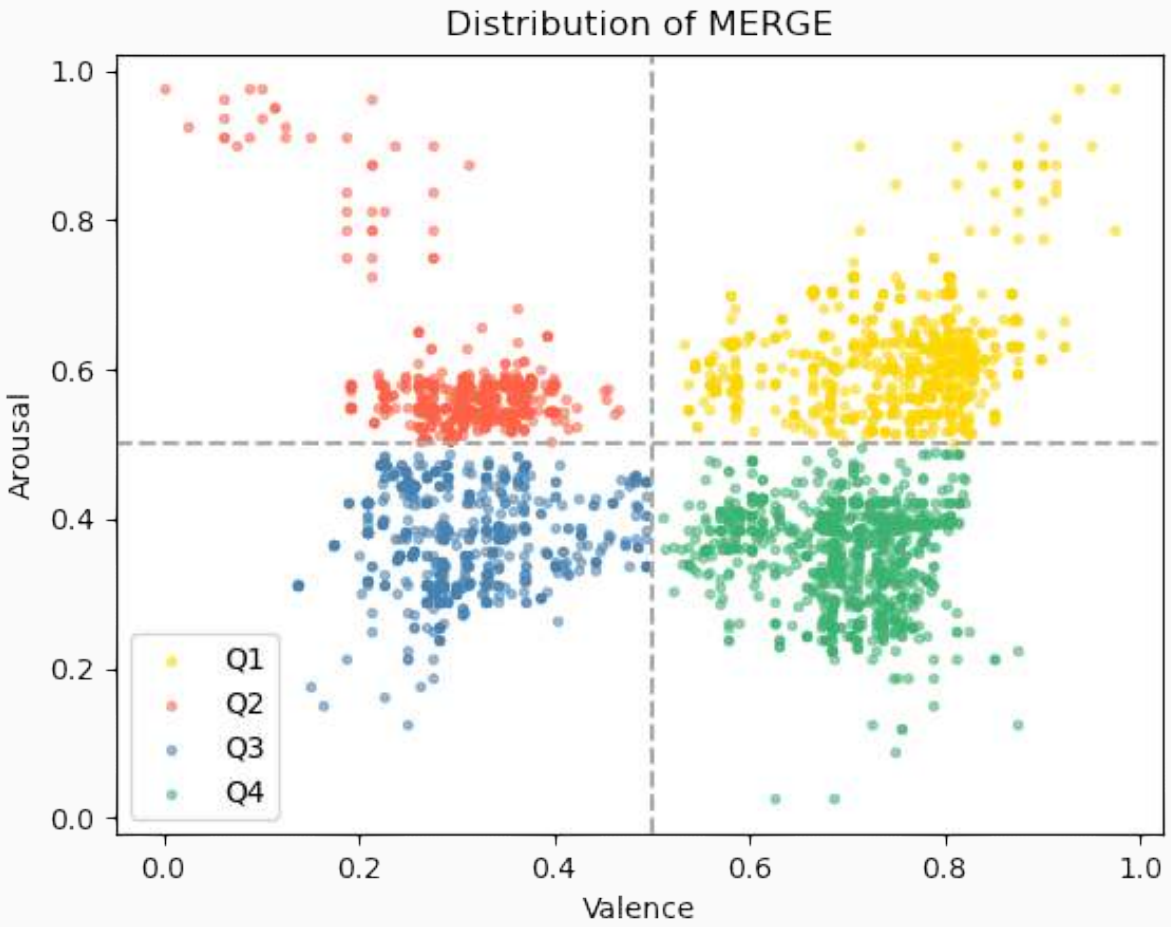


Figure: Distribution of the MERGE dataset. Regression values are discretized into the four quadrants for classification.

Our Efforts

Handcrafted Features

Musical Dimension	Feature Count
Melody	14
Harmony	99
Rhythm	46
Dynamics	63
Tone Color	449
Expressiveness	6
Form	2
High-Level	1

Table: The distribution of handcrafted features by musical dimension [13].

Extraction

- Five audio toolkits: Essentia, MIR TB, PsySound3, Marsyas, Yaafe
- Frame Size: 2048, Hop Size: 1024
- Feature post-processing removes constant and correlated features
- Total low/mid-level features:
  - Dynamic: 467 *time steps* $\times$ 608 *features*
  - Static: 300 *features*

Wav2Vec 2.0

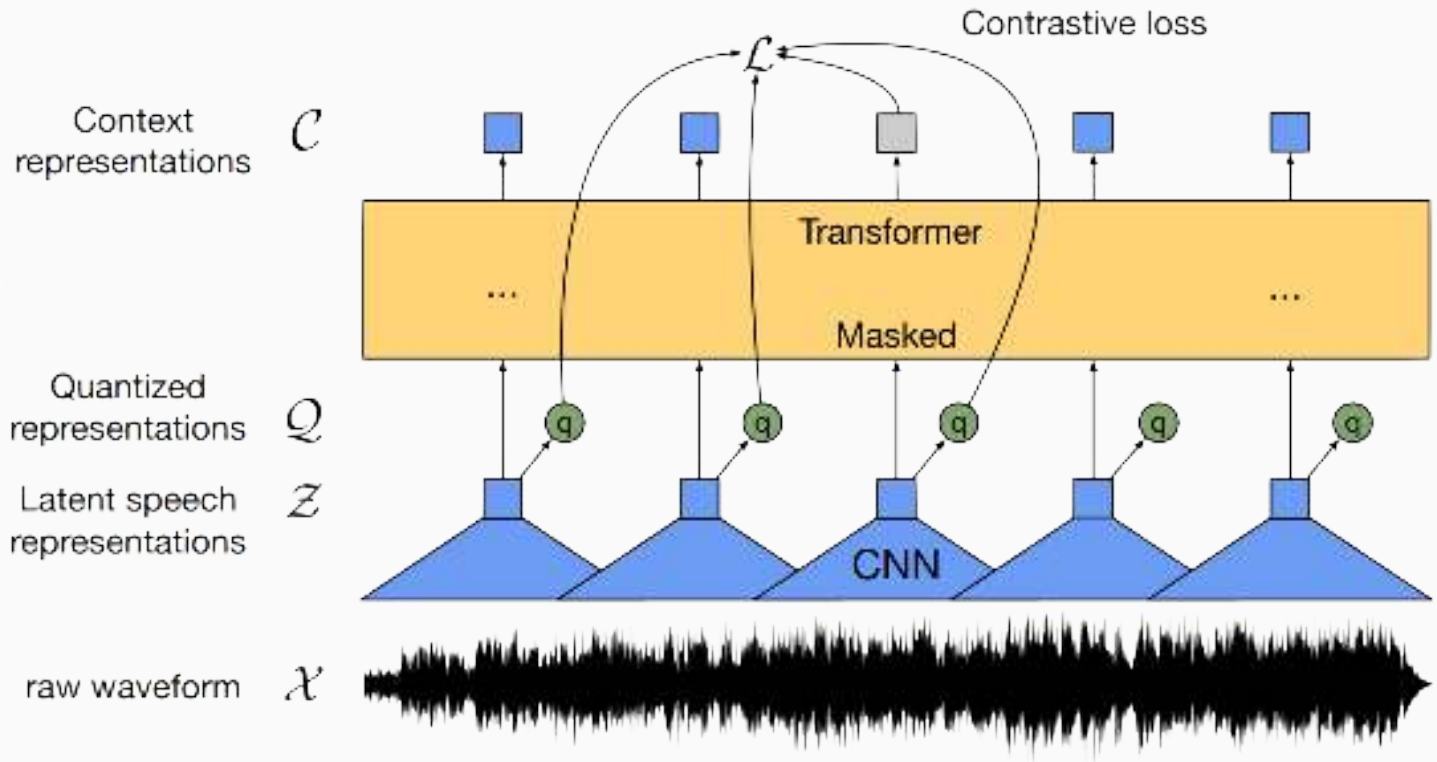


Figure: A visualization of the Wav2Vec 2.0 framework which learns contextualized representations from quantized speech units. Figure from [2].

MERT

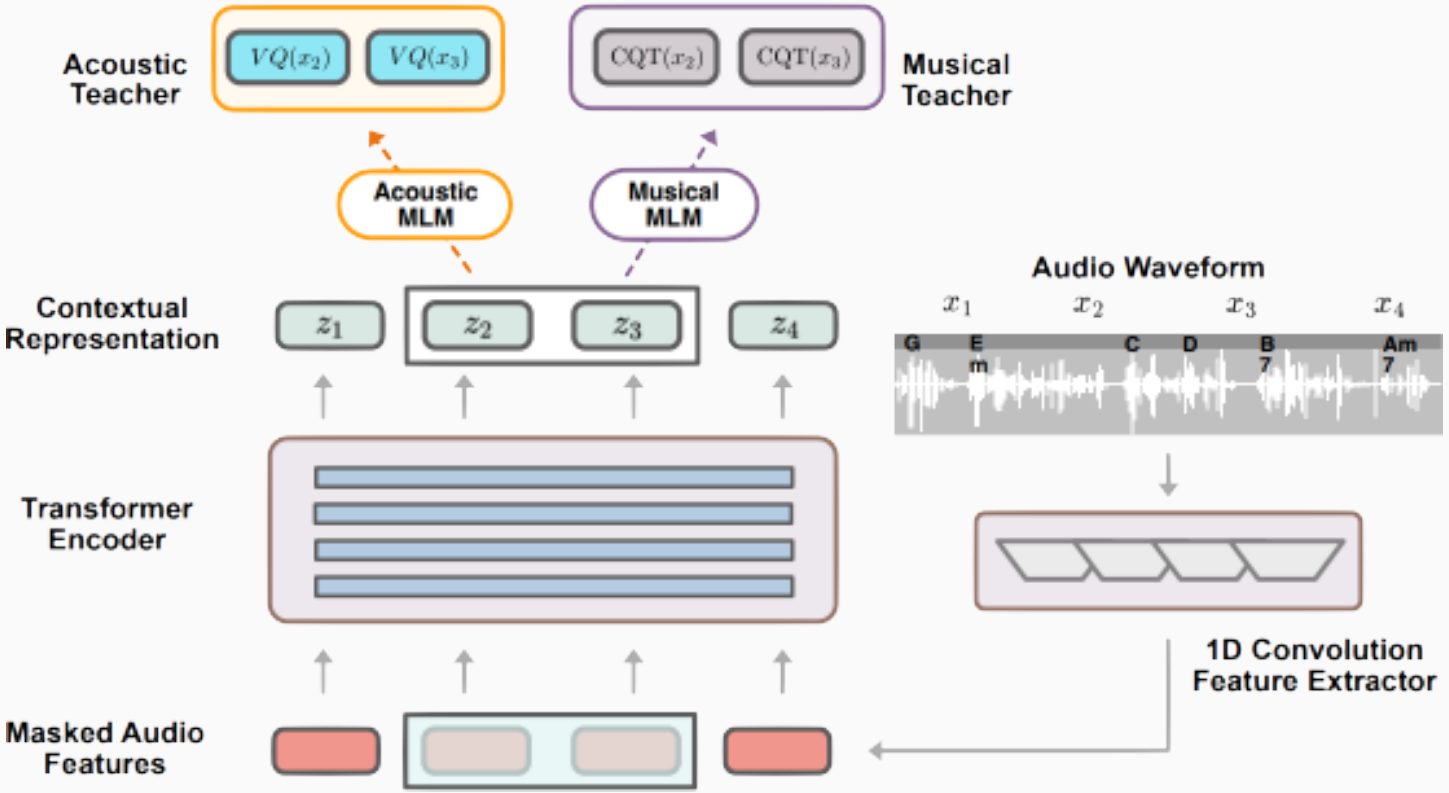


Figure: A visualization of the MERT training framework which utilizes a student teacher framework with acoustic and musical teachers. Figure from [7].

MAEST

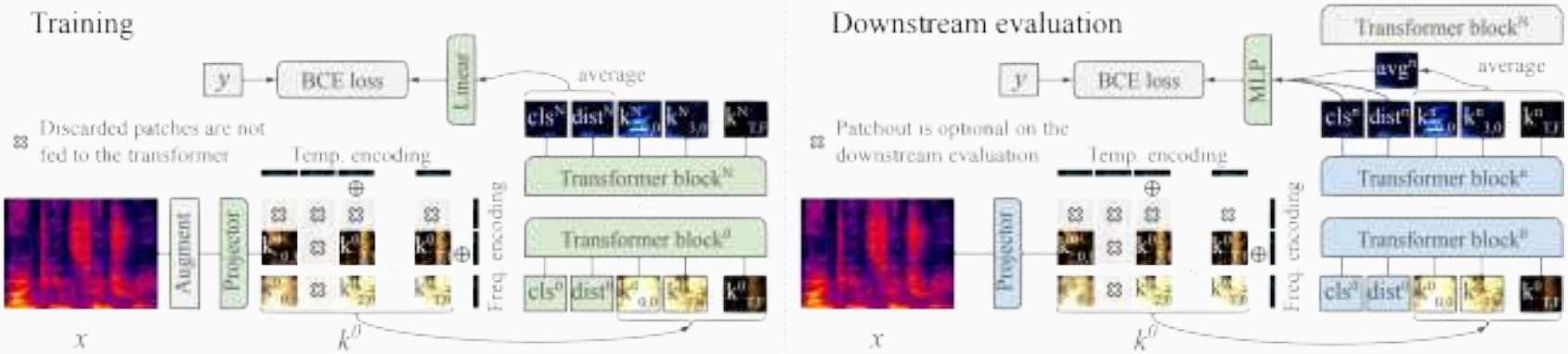


Figure: A visualization of the supervised training and downstream evaluation procedure of MAEST. Trainable blocks are shown in green and frozen blocks in blue. Figure from [1].

Our Implementation of Pre-trained Models

Embedding Extraction

- Wav2Vec 2.0 and MERT
  - 1D CNN to reduce 25 hidden layers
  - Average pooling and linear interpolation
  - Dynamic: 467 *time steps* $\times$ 1024 *features*
- MAEST
  - 7th hidden layer is best for downstream tasks [1]
  - Average pooling and linear interpolation
  - Dynamic: 467 *time steps* $\times$ 768 *features*
  - First two time steps are classification and distillation tokens

Experimental Design

- Ten-fold cross validation across 20 independent runs
- Classification: Macro/Weighted F_1-Score to evaluate performance

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

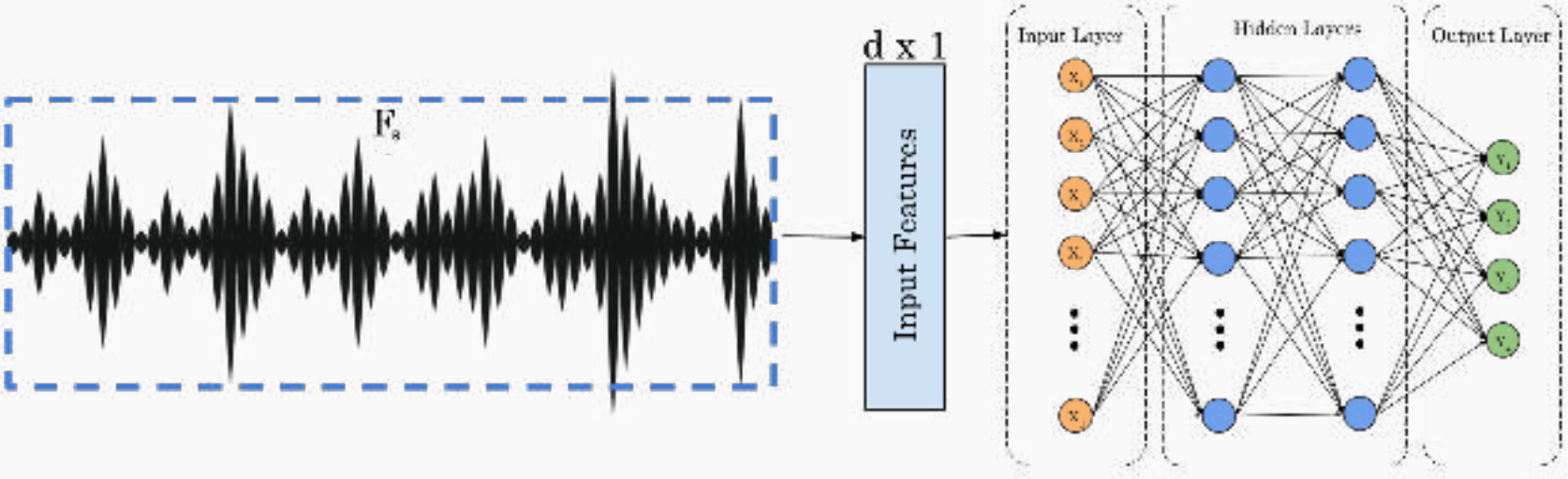
- Regression: RMSE, R, R^2

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N}}, \quad R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

- Bayesian optimization with early stopping for hyperparameter search

Static Features

Static Problem



Static Results

Feature Type	Macro F_1	Weighted F_1
Handcrafted	68.39	68.75
Wav2Vec 2.0	68.27	68.55
MERT	74.51	74.71
MAEST avg	**77.57**	**77.78**
MAEST cls	75.98	76.20
Handcrafted [10]	70.12	–
Handcrafted + Spectrogram [9]	68.84	–
MERT [8]	–	72.37
Lyrics + Spectrogram [8]	–	**79.21**

Table: Results from our static features (top) compared against static baselines (bottom).

Regression

Feature Type	RMSE V	RMSE A	R V	R A	R^2 V	R^2 A
Static	**0.090**	**0.155**	**0.741**	**0.720**	**0.549**	0.519
Baseline [10]	0.305	0.169	–	–	0.505	**0.528**

Table: Results from static MAEST embeddings with regression targets. Baseline results were obtained with Wav2Vec 2.0 and RoBERTa embeddings.

Regression Plot

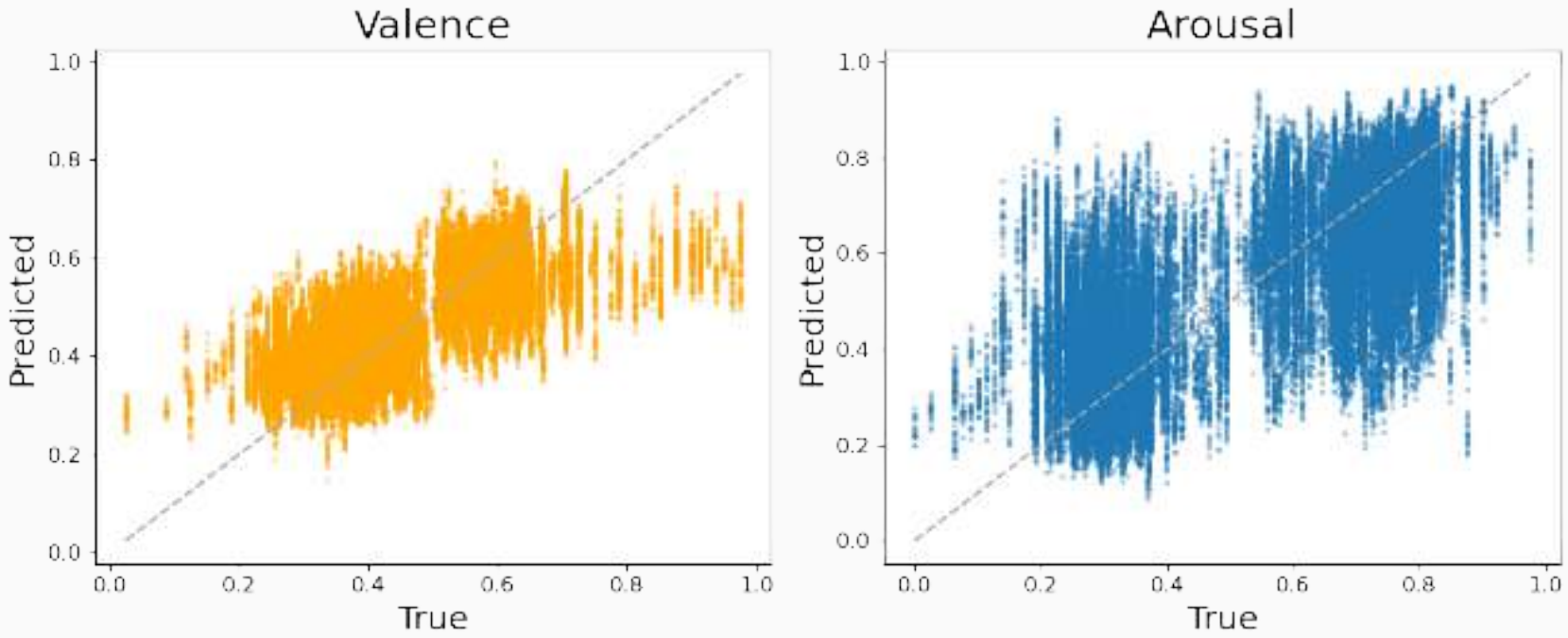
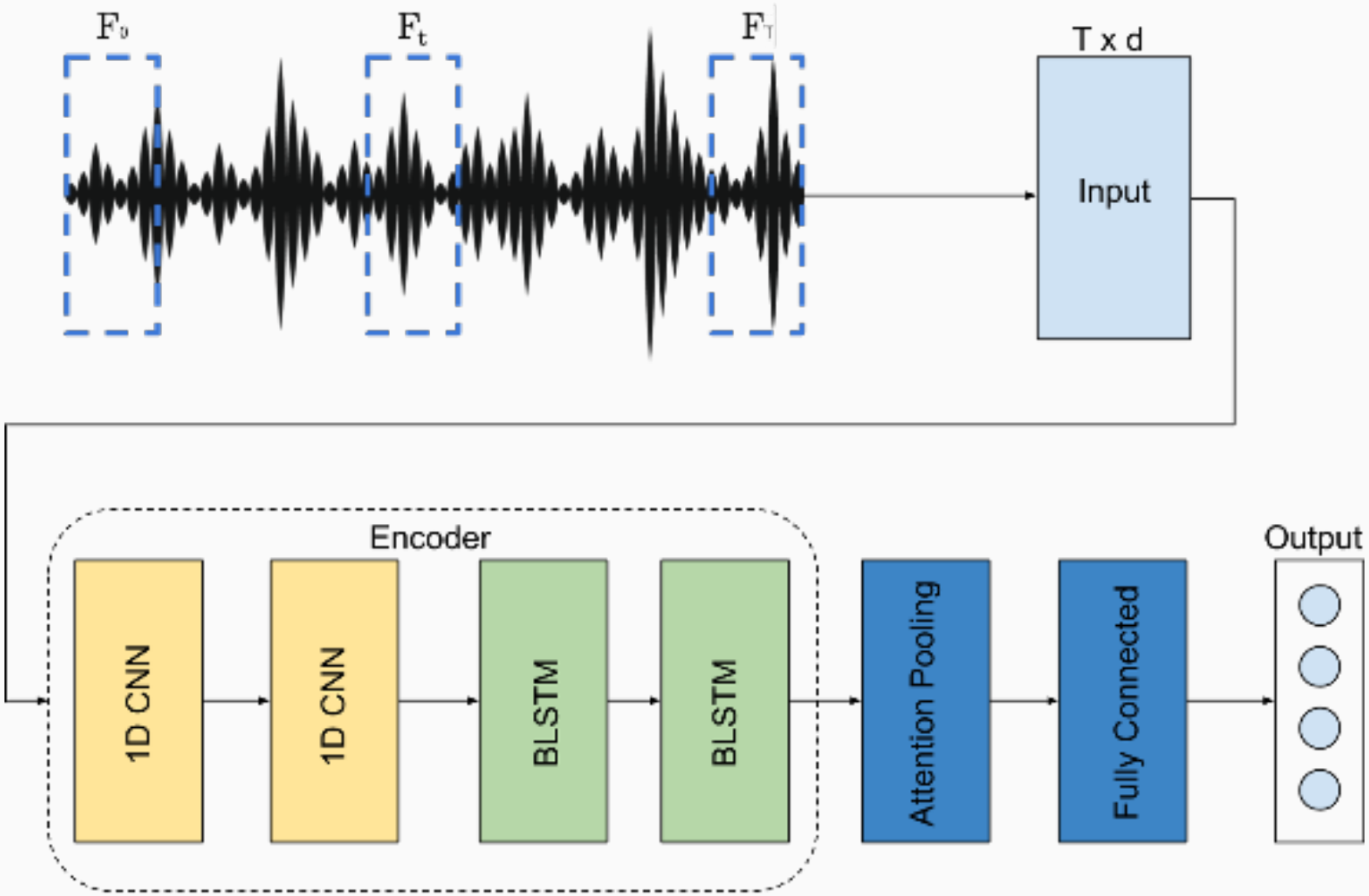


Figure: The distribution of our regression predictions utilizing static MAEST embeddings. The labels were obtained from Warriner's adjective mapping.

Dynamic Features

Dynamic Problem



Handcrafted Feature Optimization

Feature Selection (d)	Macro F_1
All (608)	**67.97**
MI (430)	66.70
MI (225)	65.60
Feature Mask	66.93
PCA (200)	66.82
PCA (400)	67.19

Table: Results using different feature selection and dimensionality reduction techniques for our handcrafted set of features.

Dynamic Results

Feature Type	Macro F_1	Weighted F_1
Handcrafted	67.97	68.29
Wav2Vec 2.0	66.78	67.10
MERT	73.48	73.70
MAEST	**76.52**	**76.72**
MAEST seq tokens	76.37	76.57
Static MAEST	**77.57**	**77.78**

Model	Macro F_1	Weighted F_1
One Transformer Block	**76.53**	**76.75**
Two Transformer Blocks	76.40	76.62
Time-Distributed FC	75.52	75.73

Table: Results of our experiments using dynamic features extracted from handcrafted methods and three pre-trained models.

Dynamic Regression

Feature Type	RMSE V	RMSE A	R V	R A	R^2 V	R^2 A
Static	**0.090**	**0.155**	**0.741**	**0.720**	**0.549**	0.519
Dynamic	0.092	0.158	0.733	0.711	0.533	0.506
Baseline [10]	0.305	0.169	–	–	0.505	**0.528**

Table: Results from our comparison of dynamic and static MAEST features with continuous regression targets.

Fusion

Fusion

Past Work

- Fusing input representations has shown to help significantly
- Mel Spectrograms + Cochleograms with cross attention produced the best results on dynamic regression on DEAM [4]
- MIDI, music theory, lyrics, and high-level features [3, 16]

Late Fusion

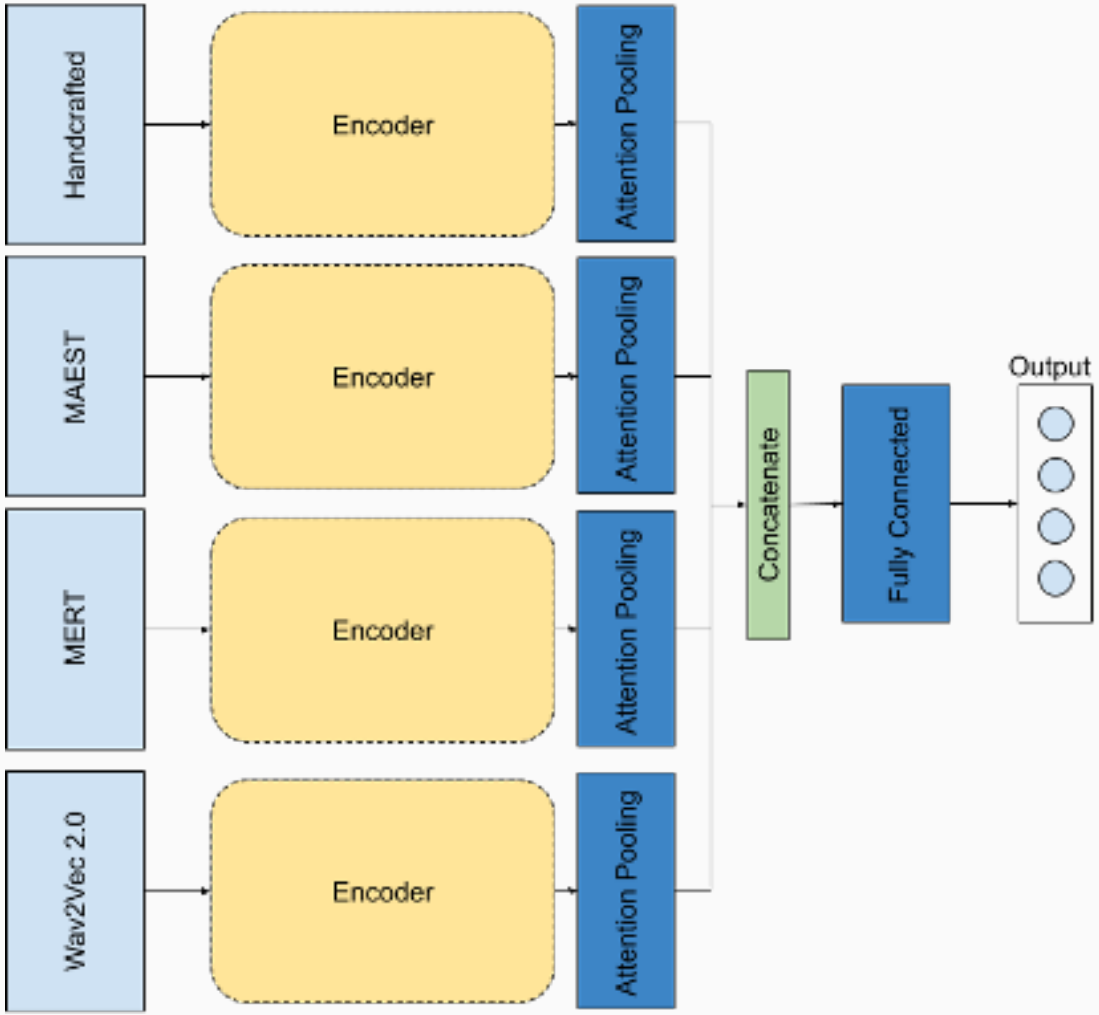


Figure: Our dynamic late fusion model with all four input types.

Fusion

Fusion Type	Macro F_1	Weighted F_1
Early	76.72	76.91
Late	**76.95**	**77.14**
Cross (Hand - embeddings)	74.45	74.66
Cross (all - all)	74.55	74.75
Dynamic MAEST	76.52	76.72
Early	76.72	76.92
Late	77.16	77.34
Static MAEST	**77.57**	**77.78**

Table: Results from different fusion strategies on our dynamic (top) and static (bottom) features.

Fusion Ablation

Feature Subset	Macro F_1	Weighted F_1
Hand., MERT, Wav2Vec 2.0, MAEST	77.16	77.34
Hand., MERT, Wav2Vec 2.0	74.08	74.33
Hand., Wav2Vec 2.0	70.62	70.90

Table: Late fusion ablation results with different subsets of static features.

Discussion

Static vs. Dynamic

- Static representations consistently outperforms dynamic ones
- Mismatch in resolution between features and labels
- Dynamic features introduce unnecessary fluctuations not relevant
- Static aggregated features emphasize most relevant cues
- Other studies produce competitive results using dynamic features
- Difference in dataset labeling methods
- Dynamically-labeled, modernized dataset

Limitations

Limitations

- No dynamically labeled songs
- 3554 songs limit large model training
- No fine-tuning

Future Work

- Compare results with a dynamic quality-controlled dataset
- Fine-tune a pre-trained model
- Incorporating lyrics or a different modality (MIDI) for fusion

Conclusion

Takeaways

- Static features outperform dynamic features for global level labels
- Aggregated information is sufficient to capture predominant emotion
- Dynamic features may introduce unnecessary noise
- Modernized dataset with time-varying labels to advance MER

Conclusion

Takeaways

- Static features outperform dynamic features for global level labels
- Aggregated information is sufficient to capture predominant emotion
- Dynamic features may introduce unnecessary noise
- Modernized dataset with time-varying labels to advance MER

Contributions

- State-of-the-art regression RMSE and R values
- Achieve the best audio-only performance on MERGE with results comparable to bimodal
- MLP model trained with static MAEST embeddings
- First experiments with dynamic features on MERGE
- Paper submitted to ISMIR 2026

References i

[1] Pablo Alonso-Jiménez, Xavier Serra, and Dmitry Bogdanov. **Efficient Supervised Training of Audio Transformers for Music Representation Learning.** 2023. arXiv: 2309.16418 [cs.SD]. URL: https://arxiv.org/abs/2309.16418.

[2] Alexei Baevski et al. **wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.** 2020. arXiv: 2006.11477 [cs.CL]. URL: https://arxiv.org/abs/2006.11477.

[3] Ziyi Huang et al. **"Cross-Level Feature Fusion for Music Emotion Recognition: Integrating Mel-Spectrogram, High-Level Musical Descriptors and Symbolic MIDI Data with Attention Mechanisms".** In: *2025 4th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*. 2025, pp. 124–134. DOI: 10.1109/AIoTC66747.2025.11198797.

[4] Hanchen Jiang and Qin Zhang. **"Dual-Stream Transformer with Psychoacoustic Synergy for Dynamic Music Emotion Recognition".** In: *Submitted to International Conference on Machine Intelligence Theory and Applications*. under review. 2025. URL: https://openreview.net/forum?id=qwsUZKhH5g.

[5] Xinxing Li et al. **"A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction".** In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 544–548. DOI: 10.1109/ICASSP.2016.7471734.

[6] Xinxing Li et al. **"DBLSTM-based multi-scale fusion for dynamic emotion prediction in music".** In: *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6. DOI: 10.1109/ICME.2016.7552956.

[7] Yizhi Li et al. **MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training.** 2024. arXiv: 2306.00107 [cs.SD]. URL: https://arxiv.org/abs/2306.00107.

[8] Pedro Lima Louro et al. **"BEE-MER: Bimodal Embeddings Ensamble for Music Emotion Recognition".** In: *Sound and Music Computing Conference (SMC)* (July 2025). DOI: 10.5281/zenodo.15837364.

[9] Pedro Lima Louro et al. **"Improving Music Emotion Recognition by Leveraging Handcrafted and Learned Features".** In: *1st Latin American Music Information Retrieval Workshop*. Rio de Janeiro, Brazil, 2024.

[10] Pedro Lima Louro et al. **MERGE – A Bimodal Audio-Lyrics Dataset for Static Music Emotion Recognition.** 2025. arXiv: 2407.06060 [cs.SD]. URL: https://arxiv.org/abs/2407.06060.

[11] Ye Ma et al. **"Multi-scale Context Based Attention for Dynamic Music Emotion Prediction".** In: *Proceedings of the 25th ACM International Conference on Multimedia*. MM '17. Mountain View, California, USA: Association for Computing Machinery, 2017, pp. 1443–1450. ISBN: 9781450349062. DOI: 10.1145/3123266.3123408. URL: https://doi.org/10.1145/3123266.3123408.

References ii

[12] Miroslav Malík et al. **"Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition".** In: *Sound and Music Computing Conference (SMC)*. Vol. 14. Espoo, Finland, July 2017. DOI: 10.48550/arXiv.1706.02292.

[13] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. **"Novel Audio Features for Music Emotion Recognition".** In: *IEEE Transactions on Affective Computing* 11.4 (2020), pp. 614–626. DOI: 10.1109/TAFFC.2018.2820691.

[14] James Russell. **"A Circumplex Model of Affect".** In: *Journal of Personality and Social Psychology* 39 (Dec. 1980), pp. 1161–1178. DOI: 10.1037/h0077714.

[15] Markus Schedl et al. **"On the Interrelation Between Listener Characteristics and the Perception of Emotions in Classical Orchestra Music".** In: *IEEE Transactions on Affective Computing* 9.4 (2018), pp. 507–525. DOI: 10.1109/TAFFC.2017.2663421.

[16] Xiaohan Shi, Xingfeng Li, and Tomoki Toda. **"Multimodal Fusion of Music Theory-Inspired and Self-Supervised Representations for Improved Emotion Recognition".** In: *Interspeech* (2024).

[17] Ying Zhang, Xianghua Ding, and Ning Gu. **"Understanding Fatigue and its Impact in Crowdsourcing".** In: *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*. 2018, pp. 57–62. DOI: 10.1109/CSCWD.2018.8465305.

Cross Fusion

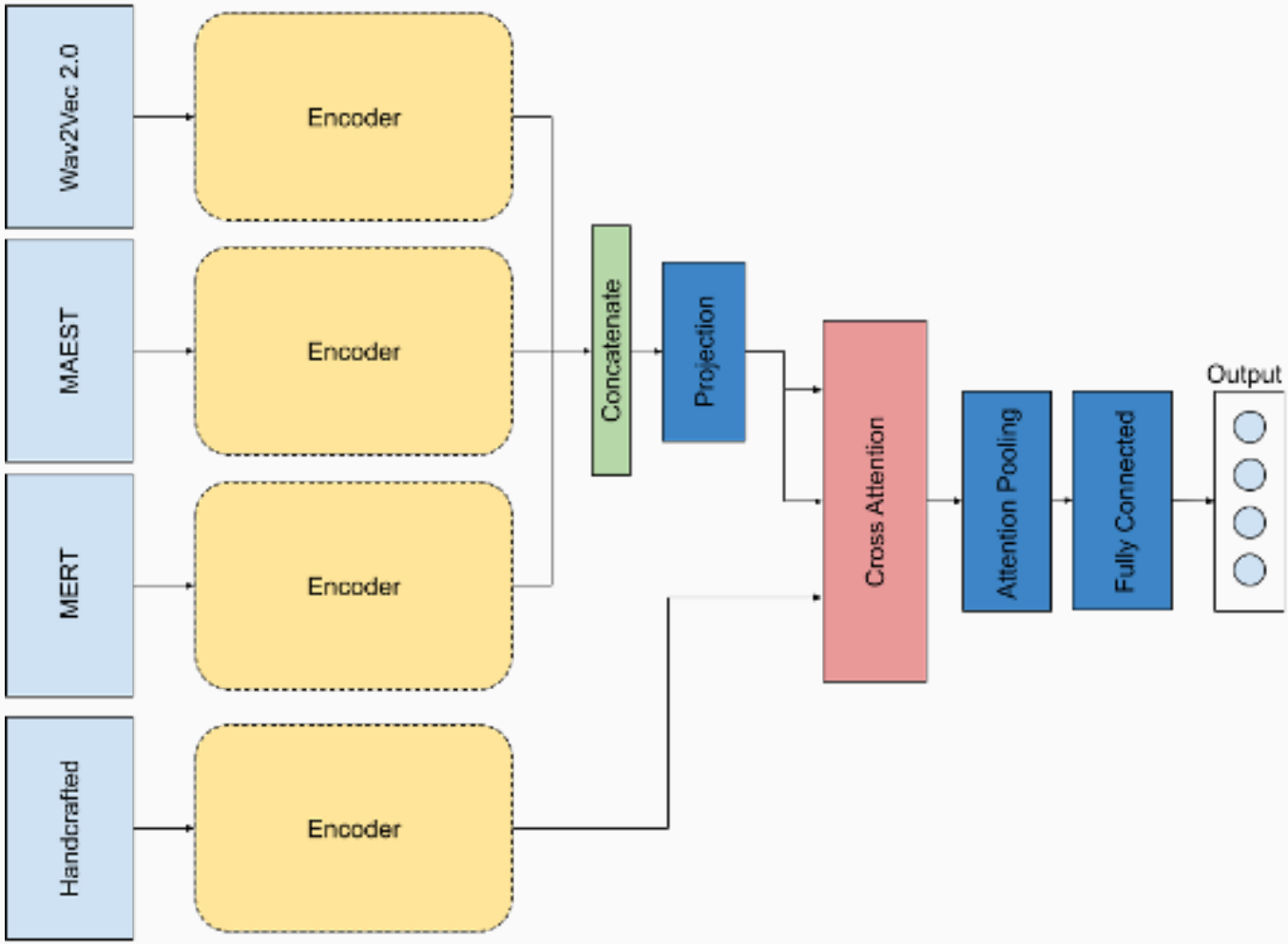


Figure: Our mid-level fusion model which utilizes cross attention to fuse inputs.

Static Improvement

Feature Type	Δ Macro F_1	Δ Weighted F_1
Handcrafted	0.42	0.46
Wav2Vec 2.0	1.49	1.45
MERT	1.03	1.01
MAEST	1.05	1.06

Table: The amount of improvement between static and dynamic features.

Downsampling

Number of Frames	**Seconds Per Frame**	Weighted $\mathbf{F}_1$
467	0.064	76.72
60	0.5	76.53
30	1	76.63

Table: Results from downsampling our MAEST embeddings to 60 and 30 time steps. Besides the number of time steps, all else remains the same.

Static vs. Dynamic

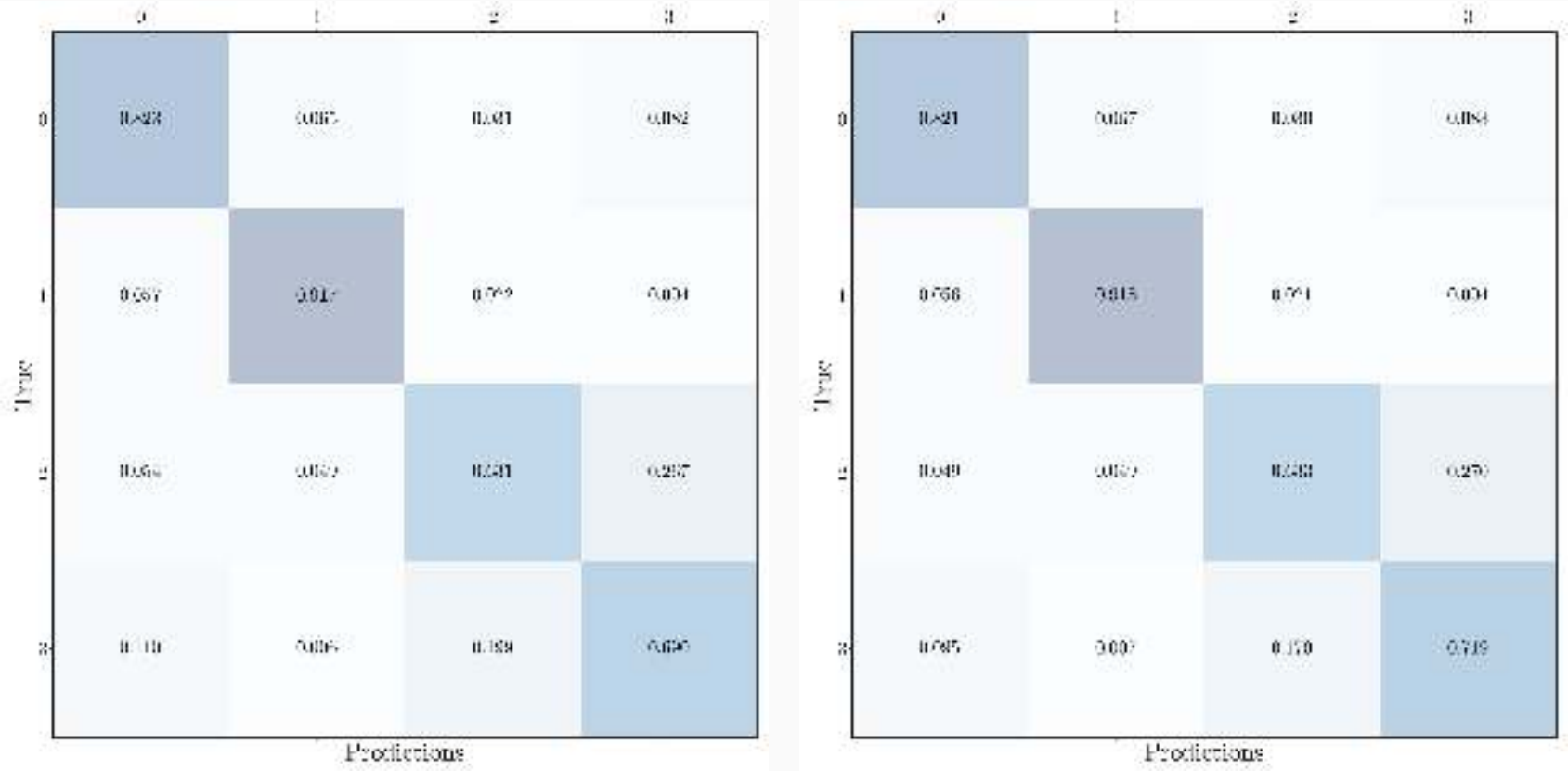


Figure: These confusion matrices show the distribution of results from our dynamic (left) and static (right) MAEST embeddings, or our best results.