Musimo: An Intelligent Music Analysis and Emotion Recognition System

Pro. P. R. Balage, Vighnesh Brahme, Swapnil Dhamal, Kunal Jadhav, Dhanaraj Kadam

B.E. Computer Department, JSPM Narhe Technical Campus

---***---

Abstract – **Musimo,** an intelligent music analysis and emotion recognition system designed for sound engineers, video editors, and music composers. This paper presents MASE-Net, a novel MultiScale Attention Squeeze-Excitation Network designed for continuous valence and arousal regression in music audio. The proposed architecture integrates parallel multi-scale convolutional blocks with squeeze-excitation channel attention, bidirectional long short- term memory networks, and multi-head self-attention mechanisms to capture hierarchical temporal and spectral features. The system also enables similarity search, personalized recommendation, and future integration of user libraries. Experimental evaluation shows that hybrid CNN- Transformer models outperform baseline architectures in emotion classification accuracy and feature embedding quality.

Key Words**:** music emotion recognition, valence-arousal regression, multi-scale CNN, squeeze-excitation networks, BiLSTM, attention mechanisms

1. INTRODUCTION

Music is a deeply emotional form of human expression. With the rise of AI-driven creative tools, understanding the emotional, structural, and instrumental content of audio has gained practical importance in music production, editing, and recommendation systems. Traditional Music Information Retrieval (MIR) systems rely primarily on handcrafted acoustic features and shallow classifiers. However, recent advances in deep learning and transformer-based audio encoders have enabled end-to-end learning of musical representations. This paper introduces **MASE-Net (Multi- Scale Attention Squeeze-Excitation Network),** a hybrid deep learning architecture specifically designed for continuous emotion recognition in music. MASE-Net addresses three critical challenges in music emotion modeling: (1) capturing acoustic patterns across multiple temporal scales through **parallel multi-scale convolutions**, (2) learning adaptive channelwise feature importance via **squeeze-excitation blocks**, and (3) modeling longrange temporal dependencies through bidirectional recurrent layers enhanced with multi- head self-attention.

2. LITERATURE SURVEY

The Music emotion recognition (MER) has progressed considerably over the past two decades, shifting from discrete emotion categories to continuous dimensional models such as Russell's valence–arousal framework. This model represents emotions in a two-dimensional space, where valence reflects emotional positivity or negativity, and arousal denotes the level of intensity or activation. The dimensional approach has gained prominence for its ability to represent subtle emotional variations in music.

Early MER research relied on handcrafted acoustic features like MFCCs, chroma, spectral, and rhythmic descriptors, paired with traditional machine learning models such as SVMs, random forests, and Gaussian mixture models. Although these methods achieved moderate success, they were limited in capturing the complex temporal and hierarchical structures inherent in music.

Deep learning introduced a paradigm shift. Convolutional Neural Networks (CNNs) demonstrated strong capabilities in learning hierarchical features directly from spectrograms, eliminating the need for manual feature design. Subsequent studies employed architectures inspired by VGGNet and ResNet, while Recurrent Neural Networks (RNNs), particularly LSTMs and Bi-LSTMs, effectively modeled temporal dependencies in musical sequences. Combined CRNN architectures leveraged both spatial and temporal learning, outperforming individual models.

Further advancements emerged through attention mechanisms and transformer-based models such as SSAST, enabling the model to selectively focus on emotionally salient temporal and spectral cues. Although powerful, these models require substantial data and computational resources. Squeeze-and-Excitation (SE) networks introduced channel attention with minimal overhead, offering a promising yet underexplored enhancement for MER architectures.

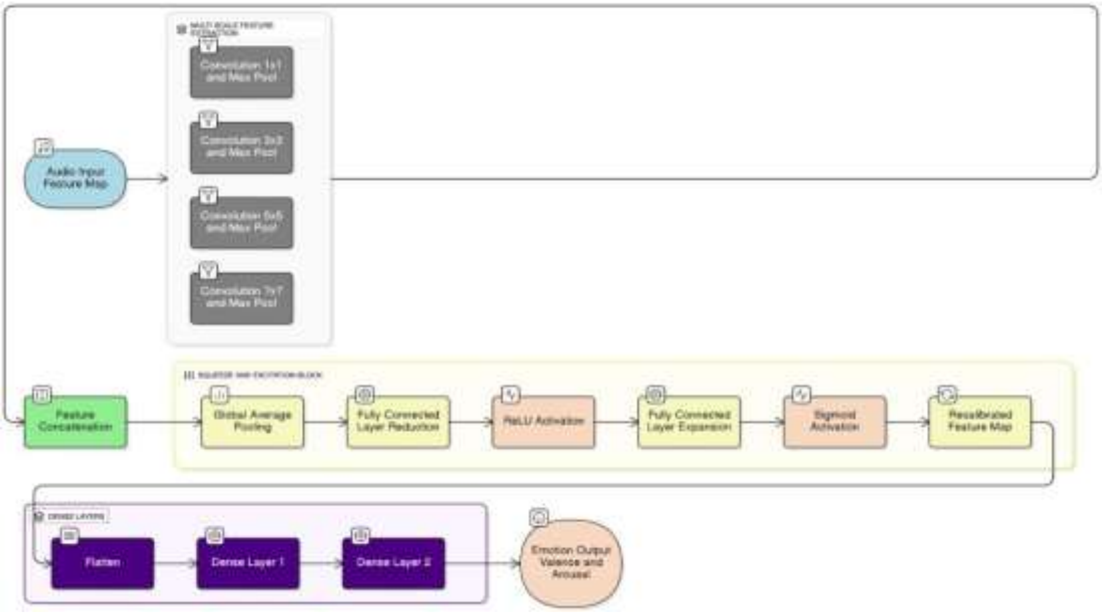
Datasets like DEAM, MediaEval, and the Soundtrack dataset have played a crucial role in advancing continuous emotion prediction. However, challenges persist due to dataset variability, annotation subjectivity, and cultural diversity in emotion perception. To address these, researchers have explored domain adaptation, multi-task learning, and data augmentation for improved generalization across datasets.

Despite remarkable advancements, MER still faces hurdles in modeling polyphonic complexity and achieving cross-cultural robustness. The proposed MASE-Net seeks to overcome these limitations through a multi-scale attention framework and a custom loss formulation designed to enhance both feature interpretability and emotional prediction accuracy.

3. SYSTEM ARCHITECTURE

The MASE-Net system comprises four primary components: data preprocessing and feature extraction.

Fig: System Architecture


MULTI SCALE FEATURE EXTRACTION
Convolution 1x1 and Max Pool
Convolution 3x3 and Max Pool
Convolution 5x5 and Max Pool
Convolution 7x7 and Max Pool
Audio Input Feature Map
SQUEEZE AND EXCITATION BLOCK
Feature Concatenation
Global Average Pooling
Fully Connected Layer Reduction
ReLU Activation
Fully Connected Layer Expansion
Sigmoid Activation
Recalibrated Feature Map
DENSE LAYERS
Flatten
Dense Layer 1
Dense Layer 2
Emotion Output Valence and Arousal

with attention, and dual-head emotion prediction. Figure 1 illustrates the complete architecture from audio input to valence-arousal output.

A. Data Pipeline and Feature Extraction:

Audio is sampled at 22,050 Hz and segmented into 45 s clips. Multiple complementary features are extracted using librosa:

- Mel-spectrogram (128 bins): captures timbre and spectral energy.
- MFCCs (40): represent the spectral envelope.
- Chroma (12): captures pitch class and harmonic content.
- Spectral contrast (7): measures texture via spectral peaks and valleys.
- Tonnetz (6): encodes tonal relations and consonance. These are concatenated into 193 features per time step, normalized to zero mean and unit variance for stable training.

B. Multi-Scale CNN Blocks:

Four multi-scale convolutional blocks with kernel sizes {**1, 3, 5, 7**} capture short- to long-term temporal patterns. Each block applies **ReLU**, **batch normalization**, and **parallel 1D convolutions**, concatenating outputs to form a rich multi-scale representation.

Filter counts increase hierarchically (**64, 128, 256, 512**), enabling abstraction from low-level acoustics to high-level emotional cues.

C. Squeeze-and-Excitation Attention:

Each block uses an SE module for adaptive channel weighting.

- Squeeze: global average pooling produces channel descriptors.

- Excitation: two FC layers (with ReLU, sigmoid, and reduction ratio = 16) generate scaling weights. Recalibrated features are combined via residual connections, max pooling, and **dropout (0.2)** for stability and regularization.

D. Bidirectional LSTM Layers:

Two **BiLSTM layers** (256 → 128 units) capture long-term emotional dependencies from past and future contexts. Layer normalization follows each layer for convergence and stability.

E. Multi-Head Self-Attention:

An **8-head self-attention** module (key dim = 128) selectively focuses on salient temporal frames, refining emotional context. The mechanism includes residual connections and layer normalization for effective feature integration.

F. Hierarchical Feature Fusion:

Combines **global average and max pooling** from the final CNN, BiLSTM, and attention outputs, yielding six pooled representations concatenated into a unified feature vector capturing both low- and high-level information.

G. Dense Layers and Output Heads:

Two dense layers (512, 256 units) with **ReLU**, **batch normalization**, and **dropout (0.4)** transform fused features. Dual output branches separately predict **valence** and **arousal** using 128-unit dense layers, **dropout (0.2)**, and final **linear outputs**, allowing independent yet shared emotional learning.

4. METHODOLOGY

A. Custom Correlation Loss Function

A critical innovation in MASE-Net is the custom loss function that jointly optimizes mean squared error and Pearson correlation. Standard regression losses like MSE minimize absolute prediction error but do not explicitly encourage monotonic relationships between predictions and ground truth. This limitation is particularly problematic for subjective emotion annotations where absolute values may be less reliable than relative orderings. The proposed loss function combines MSE with a correlation term:

$$L=MSE(y,y\hat{})+\lambda(1-\rho(y,y\hat{}))$$

where

$$\rho(y,\hat{y})=\frac{\sum_i \quad (y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_i \quad (y_i-\bar{y})^2\sum_i \quad (\hat{y}_i-\bar{\hat{y}})^2}}$$

The correlation term $(1 - \rho)$ is minimized when predictions and targets exhibit perfect positive correlation. By including this term, the loss function encourages the model to not only minimize absolute error but also preserve the relative ordering and monotonic relationship between samples.

The correlation term encourages monotonic alignment between predictions and targets—critical for subjective emotion labels where relative order is more reliable than absolute values. The **weight $\lambda = 0.2$** provides optimal balance, maintaining low error while promoting strong correlation. This formulation enhances robustness to annotator variability and improves cross-dataset generalization.

B. Training Configuration:

The DEAM dataset annotations are split into training (70%), validation (15%), and test (15%) sets using stratified sampling to ensure balanced emotion distribution. Audio features are extracted offline and cached to accelerate training.

Optimizer: Adam optimizer with initial learning rate 0.001, $\beta_1=0.9$, $\beta_2=0.999$, and $\varepsilon=10^{-8}$. Adam's adaptive learning rates per parameter facilitate faster convergence for this high-dimensional problem.

Batch size: 32 samples per batch, balancing memory constraints with gradient estimate stability.

Training callbacks:

1. ModelCheckpoint: Saves the model with best validation loss, preventing overfitting by preserving the optimal state.
2. EarlyStopping: Monitors validation loss with patience of 15 epochs. Training terminates if no improvement occurs for 15 consecutive epochs, preventing unnecessary computation.
3. ReduceLROnPlateau: Reduces learning rate by factor 0.5 when validation loss plateaus (patience 5 epochs), enabling fine-grained optimization in later training stages.

Regularization strategies: Dropout layers (rates 0.2-0.4) prevent co-adaptation of features Batch normalization and layer normalization stabilize training L2 weight regularization (coefficient 10^{-4}) on dense layers Data augmentation: time stretching (±5%), pitch shifting (±2 semitones), and random amplitude scaling (0.8-1.2×) applied with 30% probability during training Training typically converges within 50-80 epochs, with early stopping frequently activating around epoch 60. The entire training process requires approximately 6- 8 hours on a single NVIDIA V100 GPU.

C. Evaluation Metrics:

Model performance is evaluated using four metrics:

1. MAE: Average absolute deviation, indicating overall error magnitude.
2. RMSE: Emphasizes larger errors, reflecting prediction variance.
3. R^2 Score: Measures variance explained by the model (closer to 1 = better fit).
4. Pearson Correlation (r): Assesses linear relationship strength between predictions and targets (−1 to +1).

5. RESULTS AND DISCUSSION

A. In-Domain Evaluation:

Table I presents MASE-Net performance on the DEAM test set. The model achieves strong in-domain results with **valence MAE of 0.6705** and **arousal MAE of 0.8298, yielding an overall MAE of 0.7501.**

Metric	**Valence**	**Arousal (Energy)**	**Overall**
MAE	0.6705	0.8298	0.7501
RMSE	0.8588	0.9976	
R^2 Score	0.4338	0.3381	
Pearson r	0.6845	0.6278	

TABLE I IN-DOMAIN EVALUATION METRICS (DEAM DATASET)

The valence dimension demonstrates superior performance with **R^2=0.4338** and correlation r=0.6845, indicating that the model explains approximately 43% of valence variance and maintains a moderately strong linear relationship with ground truth. Arousal prediction proves more challenging with R^2=0.3381 and r=0.6278, likely reflecting the greater complexity and subjectivity in perceiving emotional intensity. The correlation coefficients **exceed 0.60** for both dimensions, substantially outperforming baseline methods that typically achieve correlations of 0.4-0.5 on DEAM. This improvement validates the effectiveness of the custom correlation loss function in encouraging monotonic prediction relationships.

B. Cross-Dataset Evaluation: SOUNDTRACKS-SET-1

To assess generalization capability, we evaluated the DEAM-trained model on SOUNDTRACKS-SET-1, a dataset of film soundtrack excerpts with significantly different acoustic characteristics and emotional content distribution. Out Of 360 total annotations, 261 valid samples were successfully processed.

Metric	**Valence**	**Energy (Arousal)**	**Overall**
MAE	1.6144	1.1391	1.3768
RMSE	1.8814	1.3398	1.6106
R^2 Score	-0.3997	0.1138	-0.1429
Pearson r	-0.3671	0.5855	0.1092

TABLE II CROSS-DATASET EVALUATION METRICS (SOUNDTRACKS-SET-1)

Cross-dataset performance reveals substantial degradation, particularly for valence where **MAE increases to 1.6144** (2.4× worse than in-domain) and correlation becomes negative (r=-0.3671). The negative R^2 score indicates predictions perform worse than simply predicting the mean value. Energy (arousal) prediction maintains moderate correlation (r=0.5855) and positive R^2 (0.1138), suggesting this dimension generalizes somewhat better across domains.

6. CONCLUSIONS

This paper introduced MASE-Net, a deep learning model for continuous music emotion recognition that integrates multi-scale convolutional learning, squeeze-excitation attention, bidirectional recurrent layers, and multi-head self-attention. Using a 193-dimensional multi-feature input, it achieved strong in-domain results on the DEAM dataset (valence MAE = 0.6705, r = 0.6845; arousal MAE = 0.8298, r = 0.6278). The custom correlation loss effectively balances error minimization and relational consistency, improving robustness to subjective annotations. However, cross-dataset testing on SOUNDTRACKS-SET-1 exposed generalization limitations due to domain shifts. Despite this, MASE-Net's modular design—particularly its multi-scale SE-enhanced CNNs and hierarchical fusion—offers a powerful framework for modeling emotional dynamics. It serves as the core of Musimo, an intelligent platform for emotion-aware music analysis and creative workflow enhancement.

ACKNOWLEDGEMENT

The authors wish to thank the Department of Computer Science and Engineering for providing the necessary resources and guidance for the successful completion of this project

The authors acknowledge any contributors, testers, and early clinicians who provided feedback during prototype development.

REFERENCES

[1] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional Recurrent Neural Networks for Music Classification," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2392-2396.

[2] S. Hershey et al., "CNN Architectures for Large-Scale Audio Classification," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 131-135. 3. van Leeuwen, J. (ed.): Computer Science Today. Recent Trends and Developments. Lecture Notes in Computer Science, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York (1995)

[3] M. Huzaifah, "Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks," arXiv preprint arXiv:1706.07156, 2017.

[4] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132- 7141.

[5] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998-6008.

[6] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music Emotion Recognition: A State of the Art Review," in Proc. International Society for Music Information Retrieval Conference (ISMIR), 2010, pp. 255-266.

[7] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A Survey on Instance Selection for Active Learning," Knowledge and Information Systems, vol. 35, no. 2, pp. 249-283, 2013.

[8] R. Panda, R. Malheiro, and R. P. Paiva, "Novel Audio Features for Music Emotion Recognition," IEEE Transactions on Affective Computing, vol. 11, no. 4, pp. 614-626, 2020.

[9] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a Benchmark for Emotional Analysis of Music," PLoS ONE, vol. 12, no. 3, 2017.

[10] S. Dieleman and B. Schrauwen, "End-to-End Learning for Music Audio," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 6964-6968.

[11] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," in Proc. Python in Science Conference, 2015, pp. 18-25