# DUAL ATTENTION-BASED MULTI-SCALE FEATURE FUSION APPROACH FOR DYNAMIC MUSIC EMOTION RECOGNITION

Liyue Zhang, Xinyu Yang, Yichi Zhang, Jing Luo

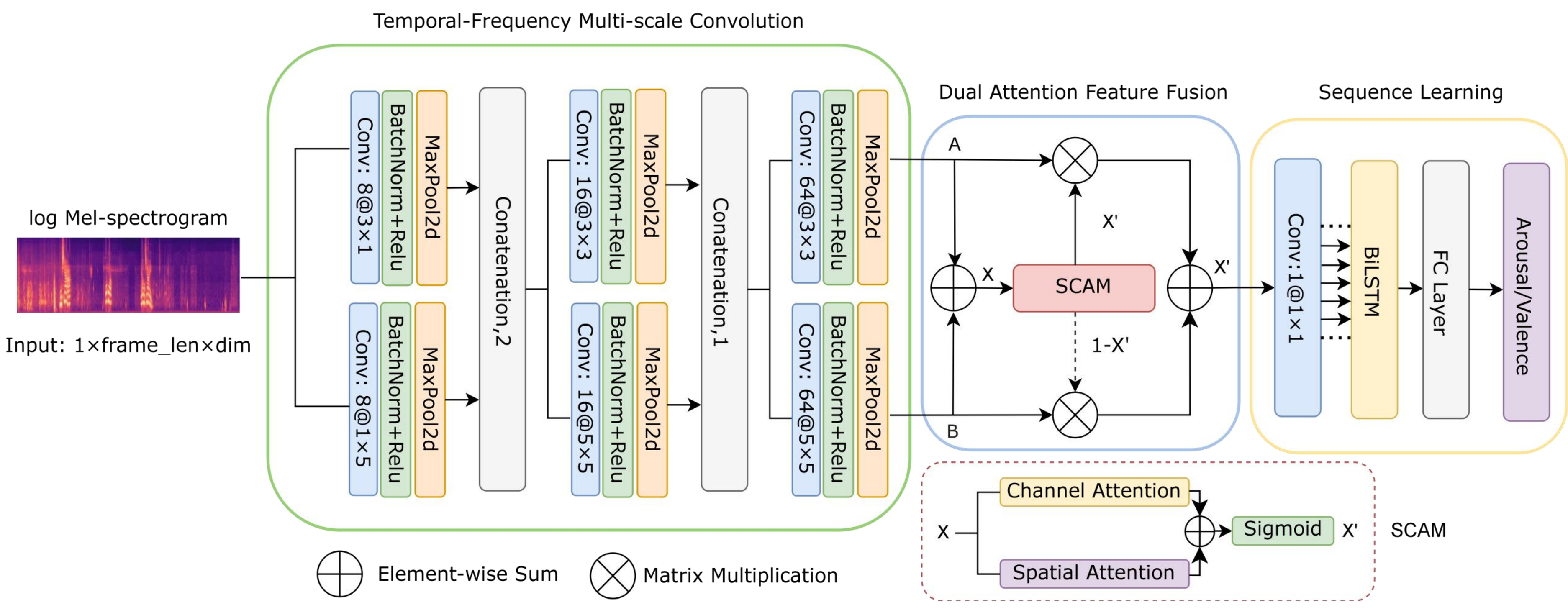
ISMIR 2023 Milan, Italy Nov. 5-9, 2023 POLITECNICO MILANO 1863

## Motivation & Contributions

- There are some issues with the current DMER models. LSTM-based models still use handcrafted features as input, and some widely used handcrafted feature operations will lose high-level features. The CNN-RNN based model mainly uses a fixed-scale CNN. Due to its fixed receptive field, the learned CNN features are limited, and the emotional crucial features of different fields of view are not extracted. Various problems exist in existing music emotion datasets, which also hinder the progress of DMER.
- This paper proposes a novel Dual Attention-based Multi-scale Feature Fusion (DAMFF) network, which extracts multi-scale convolutional features from spectrograms and exploits the dual-attention mechanism to capture important channel and spatial information.
- The music emotion dataset MER1101 we developed contains 1101 music audio with 16 genres, 5 languages and a balanced distribution of emotion labels.

## Methodology

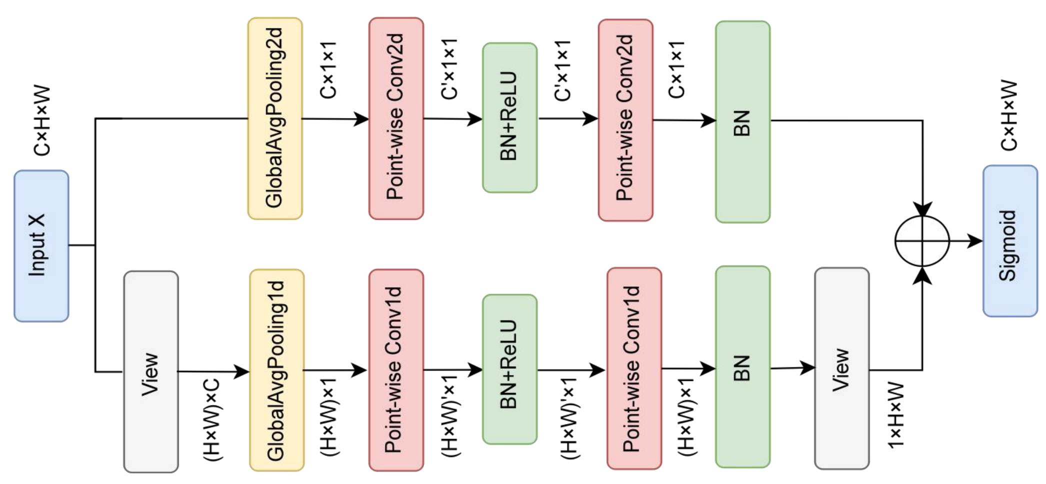
- This paper proposes a novel Dual Attention-based Multi scale Feature Fusion (DAMFF) network.




### Temporal-Frequency Multi-scale Convolution

- We extract features through three layers of parallel convolutional blocks in the Temporal-Frequency Multi-scale Convolution module.

### Dual Attention Feature Fusion




*Spatial Channel Attention Module (SCAM)*

- $C = \beta(Conv2d_2(\delta(\beta(Conv2d_1(Pool2d(X))))))$ *Channel Attention Module*
- $S = \beta(Conv1d_2(\delta(\beta(Conv1d_1(Pool1d(X))))))$ *Spatial Attention Module*
- $X' = Sigmoid(S \oplus C)$
- $Z = X' \otimes A + (1\text{-}X') \otimes B$ *Feature Fusion Strategy*

### Sequence Learning

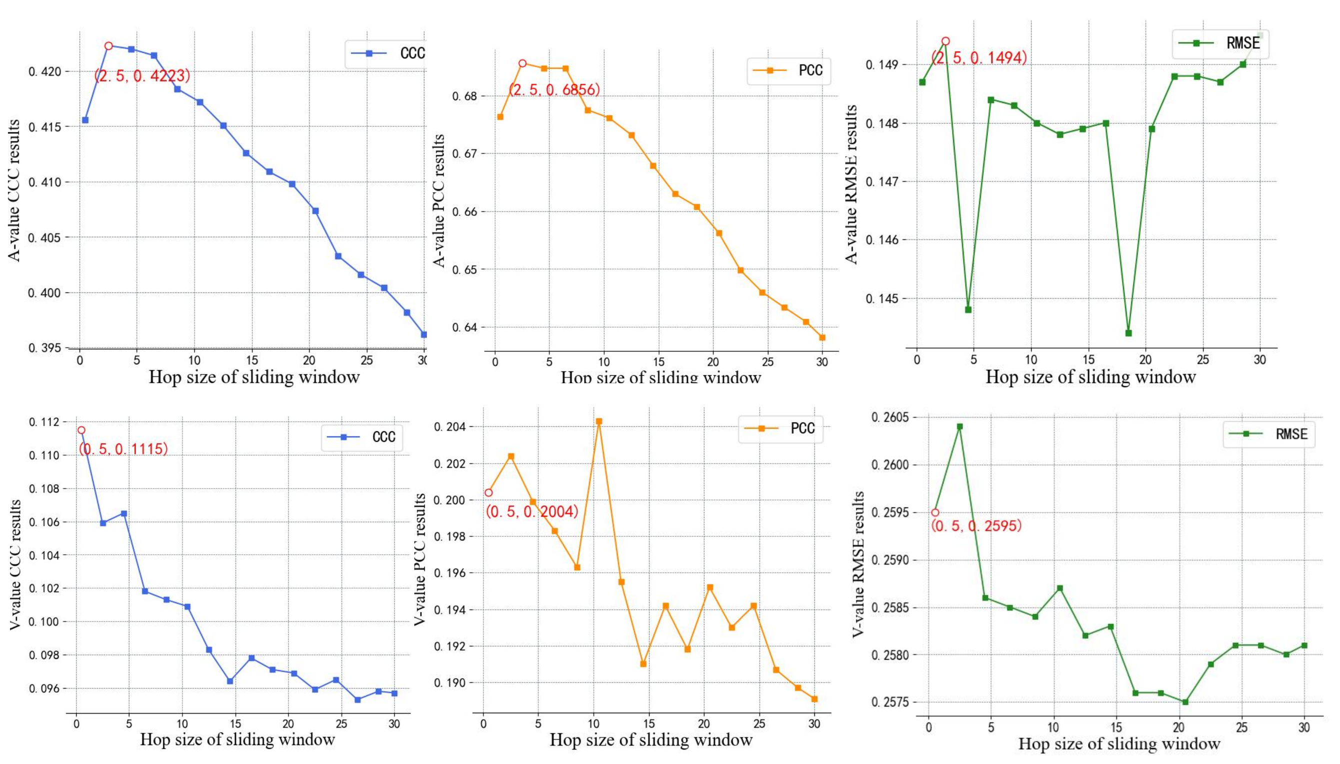
- Finally, we employ BiLTSM, building a map from emotion-crucial features to emotional space.

## Experiments

### MER1101 dataset

- Compared with the existing publicly available datasets in the MER domain, MER1101 contains 1101 music snippets from 16 genres with richer languages, more extensive size, and more balanced emotion label distribution.

### Hop Size Selection of Sliding Window




*With the increase of the hop size, the emotion prediction effect decreased significantly, demonstrating that the shorter hop size performs better.*

### Impact of CNN filters

| Model | Arousal | | | Valence | | |
|---|---|---|---|---|---|---|
| | CCC↑ | PCC* ↑ | RMSE* ↓ | CCC↑ | PCC↑ | RMSE↓ |
| **Hybrid CNN** | **0.4223** | 0.6856 | 0.1494 | **0.1115** | **0.2004** | 0.2595 |
| T-F CNN | 0.4120 | 0.6787 | 0.1478 | 0.0846 | 0.1363 | 0.2684 |
| Square CNN | 0.4130 | 0.6894 | 0.1439 | 0.0732 | 0.1343 | 0.2703 |
| T-S CNN | 0.4090 | 0.6881 | 0.1458 | 0.1085 | 0.1959 | **0.2542** |
| F-S CNN | 0.4150 | 0.6804 | 0.1562 | 0.1046 | 0.1640 | 0.2800 |

* The result of the significance test (Student's t test) show that there is no significant difference between the results of this metric.

### Comparison with the Existing Models

- The proposed method DMAFF achieves state-of-the-art results!

MER1101 dataset

| Model | Arousal | | | Valence | | |
|---|---|---|---|---|---|---|
| | CCC↑ | PCC↑ | RMSE↓ | CCC↑ | PCC↑ | RMSE↓ |
| CRNN | 0.2798 | 0.5177 | 0.1625 | 0.0573 | 0.1033 | 0.2721 |
| BCRSN | 0.1741 | 0.3770 | 0.3063 | 0.0660 | -0.0647 | 0.4143 |
| DNN | 0.0529 | 0.0903 | 0.2372 | 0.0118 | 0.0017 | 0.2734 |
| MCRNN | 0.0564 | 0.0918 | 0.2401 | 0.0155 | 0.0028 | 0.2752 |
| **DAMFF** | **0.4223** | **0.6856** | **0.1494** | **0.1115** | **0.2004** | **0.2595** |

DEAM2015 dataset

| Model | Arousal | | | Valence | | |
|---|---|---|---|---|---|---|
| | CCC↑ | PCC↑ | RMSE↓ | CCC↑ | PCC↑ | RMSE↓ |
| CRNN | 0.3488 | 0.5885 | **0.2197** | 0.0053 | -0.0292 | 0.3542 |
| BCRSN | 0.3168 | 0.5148 | 0.2397 | 0.0125 | -0.0171 | 0.2914 |
| DNN | 0.2757 | 0.4282 | 0.2483 | 0.0075 | 0.0031 | 0.3353 |
| MCRNN | 0.2700 | 0.4396 | 0.2428 | 0.0137 | 0.0126 | 0.3135 |
| **DAMFF** | **0.4203** | **0.6866** | 0.2401 | **0.0151** | **0.0366** | **0.3403** |

### Ablation Study

| Model | Arousal | | | Valence | | |
|---|---|---|---|---|---|---|
| | CCC↑ | PCC* ↑ | RMSE* ↓ | CCC↑ | PCC↑ | RMSE↓ |
| DAMFF | **0.4223** | 0.6856 | 0.1494 | **0.1115** | **0.2004** | **0.2595** |
| w/o Fusion Strategy | 0.4097 | 0.6869 | 0.1563 | 0.0846 | 0.1363 | 0.2684 |
| w/o Channel Attention | 0.4061 | 0.6894 | 0.1439 | 0.0732 | 0.1343 | 0.2703 |
| w/o Spatial Attention | 0.4090 | 0.6881 | 0.1458 | 0.1085 | 0.1959 | 0.2542 |
| w/o DAFF | 0.4150 | 0.6804 | 0.1562 | 0.1046 | 0.1640 | 0.2800 |

* The result of the significance test (Student's t test) show that there is no significant difference between the results of this metric.