CISUC
CENTRO DE INFORMÁTICA E SISTEMAS DA UNIVERSIDADE DE COIMBRA

IMPROVING MUSIC EMOTION RECOGNITION BY LEVERAGING HANDCRAFTED AND LEARNED FEATURES

Pedro Lima Louro [1 †], pedrolouro@dei.uc.pt
Hugo Redinho [1 †], redinho@dei.uc.pt
Ricardo Malheiro [1, 2 †], rsmal@dei.uc.pt
Rui Pedro Paiva [1 †], ruipedro@dei.uc.pt
Renato Panda [1, 3 †], panda@dei.uc.pt

[1] Centre for Informatics and Systems of the University of Coimbra (CISUC)
[2] Polytechnic Institute of Leiria - School of Technology and Management
[3] Ci2 - Smart Cities Research Center - Polytechnic Institute of Tomar
[†] Intelligent Systems Associate Laboratory (LASI)

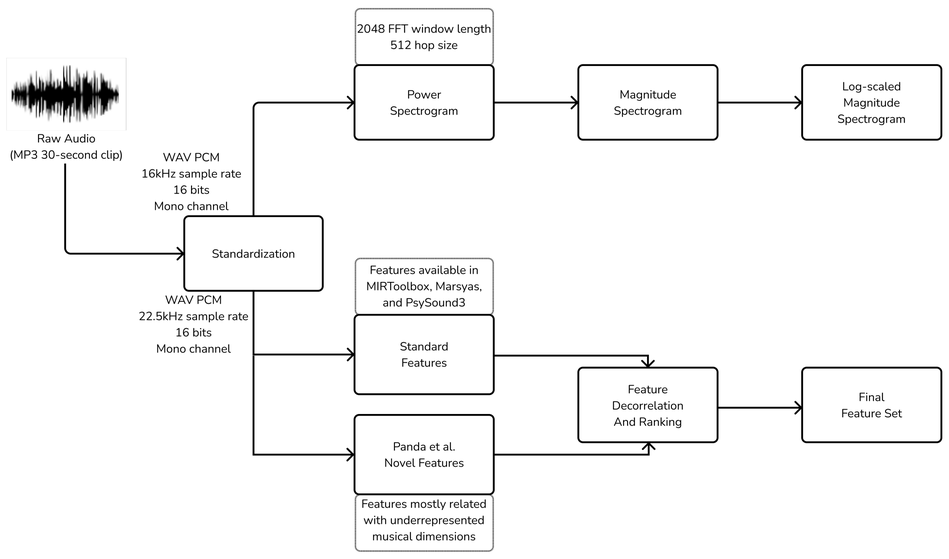

Figure 1

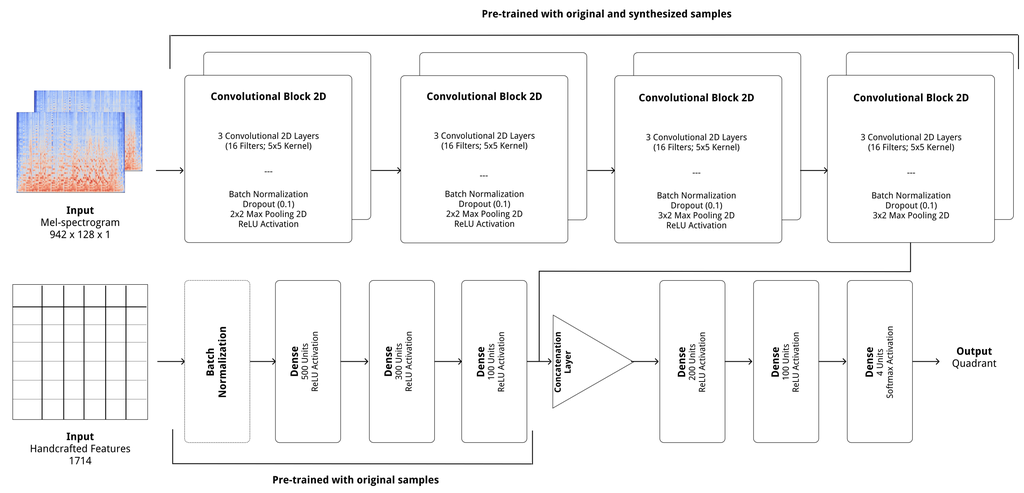

Figure 2

1. Introduction

- Feature engineering (FE) based methodologies provide emotionally-relevant patterns for Music Emotion Recognition (MER);
- These are very time-consuming to develop and require knowledge from a myriad of domains;
- Deep learning (DL) approaches are promissing for MER due to their feature learning (FL) capabilities;
- Lack of quality and sizeable datasets considerably reduce the generalization of these approaches;
- A previous study from our team [1] shows that a combination of both FE and FL increases the classification performance;
- We conduct further studies on the newly proposed MERGE datasets [2], to assess the impact of the sample size increase.

2. Methodology

- Samples are first standardized before obtaining Mel-spectrograms and handcrafted features, as shown in Figure 1;
- The neural network architecture is composed of an FL branch, extracting the most relevant patterns from Mel-spectrograms, and an FE branch, processing and selecting the most relevant hancrafted features;
- Concatenation of both branches' outputs is performed before a dense network conducts classification, depicted in Figure 2.

3. Datasets and Evaluation

- MERGE is a collection of audio, lyrics, and bimodal audio-lyrics datasets, each comprissing a complete and balaced set;
- For this study, we only consider the audio and bimodal sets since these provide audio samples;
- The class distribution of the four sets can be seen in Table 1;

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

Table 1

- The proposed 70-15-15 train-validate-test split was used for optimization and evaluation;
- Model optimization is conducted using bayesian optimization over 10 trials, optimizing batch size, optimizer, and corresponding learning rate;
- Weighted precision, recall, and F1-score metrics were computed for each dataset's test set.

4. Results and Discussion

- Results for each dataset are presented in Table 2;
- The best result is a 77.62% F1-score in the MERGE Audio Balanced dataset;
- This model is considerably sensitive to the quadrant distribution of the dataset, as can be seen by the difference between both MERGE Audio datasets;

Dataset	F1-score	Precision	Recall
MERGE Audio Complete	68.84%	69.52%	68.80%
MERGE Audio Balanced	**77.62%**	**78.11%**	**77.89%**
MERGE Bimodal Complete	73.13%	75.45%	74.40%
MERGE Bimodal Balanced	70.00%	69.99%	70.33%

Table 2

- Analyzing the performance over each quadrant, a significant performance increase is observed in the third and fourth quadrants;
- This increase points to better valence prediction on low arousal quadrants.

5. Conclusions and Future Work

- We conducted further studies on an hybrid FE and FL architecture for MER;
- The best attained results were a 77.62% F1-score, better modelling valence on low arousal quadrants than in other datasets;
- Future work will consider data synthetization on the FE portion, consider new data augmentation strategies, and improving the classification portion by incorporating recurrent layers.

References

[1] Louro, P.L., Redinho, H., Malheiro, R., Paiva, R.P. and Panda, R. 2024. *A Comparison Study of Deep Learning Methodologies for Music Emotion Recognition*. Sensors. 24, 7 (2024), 2201. DOI:https://doi.org/10.3390/s24072201.
[2] Louro, P.L., Redinho, H., Santos, R., Malheiro, R., Panda, R. and Paiva, R.P. 2024. *MERGE -- A Bimodal Dataset for Static Music Emotion Recognition*. arXiv.
for publication at A2I: Affective Artificial Intelligence (ICPR 2024).

Captions of Figures

Figure 1. Sample pre-processing phase. Each sample is standardized before either obatining a Mel-spectrogram representation for the FL portion of the architecture, or extracting emotionally-relevant features for the FE portion.

Figure 2. Hybrid augmented model, adapted from [1]. The architecture has a dedicated branch for both FE and FL to find the most relevant patterns for the input data. The latter undergoes a pre-training phase with synthesized samples that significantly improves its standalone performance.

Captions of Tables

Table 1. MERGE datasets sample distribution over the four quadrants of the model used for annotation. A semi-automatic protocol first computes continuous values on said model, which are then discritized and validated by human subjects.

1290 UNIVERSIDADE D COIMBRA
REPÚBLICA PORTUGUESA
fct Fundação para a Ciência e a Tecnologia
LASI Intelligent Systems Associate Laboratory
MERGE Music Emotion Recognition Next Generation