# Cross-Dataset Music Emotion Recognition: an End-to-End Approach

**Ana Gabriela Pandrea** **Juan Sebastián Gómez-Cañón** **Perfecto Herrera**
Music Technology Group, UPF, Barcelona, Spain

## ABSTRACT

The topic of Music Emotion Recognition (MER) evolved as music is a fascinating expression of emotions, yet it faces challenges given its subjectivity. Because each language has its particularities in terms of sound and intonation, and implicitly associations made upon them, we hypothesize perceived emotions might vary in different cultures. To address this issue, we test a novel approach towards emotion detection and propose a language sensitive end-to-end model that learns to tag emotions from music with lyrics in English, Mandarin and Turkish.

## 1. INTRODUCTION

Music Emotion Recognition has become an important part of the Music Information Retrieval field, with applications in music recommendation and search, playlist creation, but also in therapy and marketing. Because the topic of emotion is very ambiguous and subjective, there are many challenges to overcome: improving the quality of annotations, the confusion between felt and perceived emotions, and taking advantage of the extra-musical information, e.g., culture. Various pre-extracted feature sets were proposed in order to leverage these issues [1], as well as various deep learning architectures that directly retrieve information from spectrograms [2].

## 2. METHODS

Music emotion recognition has started to be explored with the end-to-end learning strategy that encodes and learns from the raw waveform input, therefore we propose it here as a novel contribution. In this way, we classify excerpts under a mixture of dimensional and categorical taxonomy, in one of the four quadrants of the Russell's Valence-Arousal plane [3]. Moreover, we aim to investigate the relevance of cultural and dataset specific characteristics by conducting experiments with music in three different languages, English - 4Q-EMOTION dataset [4], Mandarin - CH-818 dataset [5] and Turkish - TR-MUSIC dataset [6].

At first, we examined several traditional machine learning algorithms and feature sets in order to establish a proper baseline model for the end-to-end proposal. From several classifiers built with Scikit-Learn [1] (K-Nearest-Neighbors, Support Vector Machine with linear kernel, Support Vector Machine with Radial Basis Function, Gaussian Process, Multi-Layer Perceptron, Gaussian Naive Bayes, Random Forest), we found that the most accurate results were generally obtained with the Multi-Layer Perceptron (MLP), having one hidden layer made of 100 neurons. We also compared the use of low-level descriptors extracted with Essentia [7] to the IS13 ComParE feature set [8], a well-evolved feature set for automatic recognition of audio emotion. The latter performed better thus it was used for the baseline model.

The end-to-end architecture is called SincNet [9], a Convolutional Neural Network originally designed for speaker recognition. We hypothesize that the speaker cues detected by this architecture are prone to play a role in Music Emotion Recognition with language considerations. The outstanding feature of this architecture is its first convolutional layer based on Sinc functions that implement rectangular band-pass filters [2].

## 3. RESULTS

We conducted 3 main experiments with both the baseline model and the end-to-end model, followed by a fourth only on SincNet. The first is a within-dataset evaluation, i.e., the models were trained and tested with music in the same language independently for each of the 3 sets, confirming research by [5]. While none of our models manages to learn from the Mandarin set, for English and Turkish several similarities were observed with the baseline model in terms of the best selected features and the common confusions between quadrants. We also observed that SincNet under this configuration does not outperform neither the baseline nor the state-of-the-art MER models.

Secondly, cross-dataset evaluations were performed, i.e., the model was trained with music from one culture and tested with music from another. As expected, results cross-dataset are worse than within-dataset with both MLP and SincNet. Thirdly, a mixed dataset configuration was employed under the assumption that training with more data should aim at better results if the learning is not language specific. Results do not improve under this consideration and training a single model with music from different languages only worsens the performance metrics of the MER system. Finally, with SincNet we also considered a trans-

**Attribution:** Ana Gabriela Pandrea, Juan Sebastián Gómez-Cañón, Perfecto Herrera, "Cross-Dataset Music Emotion Recognition: an End-to-End Approach", *Extended Abstracts for the Late-Breaking Demo Session of the 21st Int. Society for Music Information Retrieval Conf.,* Montréal, Canada, 2020.

[1] https://scikit-learn.org/

[2] We refer the reader to [9] for more information on the architecture.

| | | | Baseline: MLP | | | | End-to-end: SincNet | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Configuration** | **Dataset (train / test)** | **S** | **A** | **P** | **R** | **F** | **A** | **P** | **R** | **F** |
| Within dataset (English) | 4Q / 4Q * | 720 | 0.63 | 0.65 | 0.63 | 0.63 | 0.57 | 0.59 | 0.57 | 0.52 |
| TL 1 - finetune 4Q | CH-TR-4Q / 4Q | 720 | - | - | - | - | 0.56 | 0.57 | 0.56 | 0.51 |
| TL 2 - finetune 4Q | TR-CH-4Q / 4Q ** | 720 | - | - | - | - | 0.60 | 0.61 | 0.60 | 0.57 |
| Mixed dataset | ALL / 4Q | 720 | 0.57 | 0.60 | 0.57 | 0.56 | 0.57 | 0.58 | 0.57 | 0.56 |
| Within dataset (Chinese) | CH / CH * | 288 | 0.30 | 0.23 | 0.30 | 0.23 | 0.27 | 0.11 | 0.27 | 0.16 |
| TL 1 - finetune CH | 4Q-TR-CH / CH | 288 | - | - | - | - | 0.24 | 0.10 | 0.24 | 0.12 |
| TL 2 - finetune CH | TR-4Q-CH / CH | 288 | - | - | - | - | 0.26 | 0.28 | 0.26 | 0.17 |
| Mixed dataset | ALL / CH ** | 288 | 0.23 | 0.22 | 0.23 | 0.22 | 0.23 | 0.26 | 0.23 | 0.21 |
| Within dataset (Turkish) | TR / TR | 320 | 0.71 | 0.74 | 0.71 | 0.71 | 0.63 | 0.68 | 0.63 | 0.58 |
| TL 1 - finetune TR | 4Q-CH-TR / TR | 320 | - | - | - | - | 0.75 | 0.75 | 0.75 | 0.75 |
| TL 2 - finetune TR | CH-4Q-TR / TR * ** | 320 | - | - | - | - | 0.71 | 0.72 | 0.71 | 0.71 |
| Mixed dataset | ALL / TR | 320 | 0.64 | 0.67 | 0.64 | 0.64 | 0.51 | 0.61 | 0.51 | 0.50 |

**Table 1**. Summary of results from baseline and end-to-end models. S stands for samples, A for accuracy, P for precision, R for recall, and F for F-score. * stands for best overall results for each language, ** stands for best SincNet results.

fer learning approach, from one source culture to the second and fine-tuning on the third. This seemed to be the most promising set-up for this architecture, especially with the Turkish data, where one of the transfer learning set-ups gave the best results from all our experiments.

Furthermore, we extended this first set of experiments with 3 more set-ups that were decided based on these preliminary results:

- We combined both feature sets that were proposed and extracted the best 50 features from the whole set, among which there were slightly more from Essentia. This worked even better than IS13 ComParE, with increasing scores for both English (increase of 1 percent points in F-score) and Turkish music (increase of 5 percent points in F-score).

- As English and Turkish appeared to be more similar in terms of relevant features, we considered the transfer learning experiments with only these two sets. Although differences are small, it appears that when considering the Chinese set as an intermediary tuning set, scores improve.

- Since the initial SincNet configuration considered frames of only 200 milliseconds and this amount of time might be questionable for emotion depiction, we wanted to use a longer fragment. According to the available computing power, the highest window we tested was 500 milliseconds, with a smaller batch size of 32 samples. Results are comparable, with an increase of 4 percent points in F-score for English and 8 percent point for Turkish, suggesting that similar set-ups could be further explored.

## 4. DISCUSSION

We addressed the problem of automatic music emotion recognition with the end-to-end SincNet architecture. We found that our approach is limited with respect to the tuning of the network and the size and structure of the datasets as none of our results is sensitive enough. Our findings suggest that traditional methods remain the best choice w.r.t. both within- and mixed-dataset set-ups. The end-to-end architecture might be promising provided that it is better adapted to the task. Transfer learning and fine-tuning SincNet on the target language gives better results than SincNet within-dataset, suggesting that the architecture could be more suitable for with this approach and that more training data improves performance in this set-up. Future work in this area should consider larger fragments and adaptations of SincNet, and the extension of similar studies to other cultures and datasets.

## 5. REFERENCES

[1] Yang, Dong, and Li, "Review of data features-based MER methods," *Multimedia Systems*, 2017.

[2] Dong, Yang, Zhao, and Li, "Bidirectional convolutional recurrent sparse network (bcrsn): An efficient model for MER," *IEEE Trans. on multimedia*, 2019.

[3] Russell, "A circumplex model of affect," *J. Personal. Soc. Psychol.*, 1980.

[4] Panda, Malheiro, and Paiva, "Novel audio features for MER," *IEEE Trans. on Affective Computing*, 2018.

[5] Hu and Yang, "Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs," *IEEE Trans. on Affective Computing*, 2017.

[6] B. Er and B. Aydilek, "MER by using chroma spectrogram and deep visual features," *Intern. J. of Computational Intelligence Systems*, 2019.

[7] Bogdanov and et al., "Essentia: an audio analysis library for mir," *Intern. Soc. for MIR Conf.*, 2013.

[8] Schuller and et al., "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech Language*, 2018.

[9] Ravanelli and Bengio, "Speaker recognition from raw waveform with sincnet," *IEEE Spoken Language Technology Workshop*, 2018.