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

Ana Gabriela Pandrea, Juan S. Gómez Cañón, Perfecto Herrera

ISMIR MTL2020

In this work, we address **music emotion recognition with a context-based end-to-end model**. Two main research questions are addressed:

- Are there differences/correlations between the emotions perceived in music by listeners raised with different mother tongues?
- Can an end-to-end model trained on raw waveforms, that was efficient for speech tasks, be employed for language-oriented MER?

Our hypothesis is that perceived emotion depends on cultural characteristics and models should be trained in the target test language in order to obtain sensitive results. The end-to-end architecture should outperform state-of-the-art models since it allows for more information to be processed.

Problem definition: music emotion recognition

- **Valence-Arousal plane** - 4 quadrants classification [1]
- **Raw waveform** vs. handcrafted features
- End-to-end deep learning: **SincNet** [2]
- Musical emotion complexity relates to **cultural-specific associations**

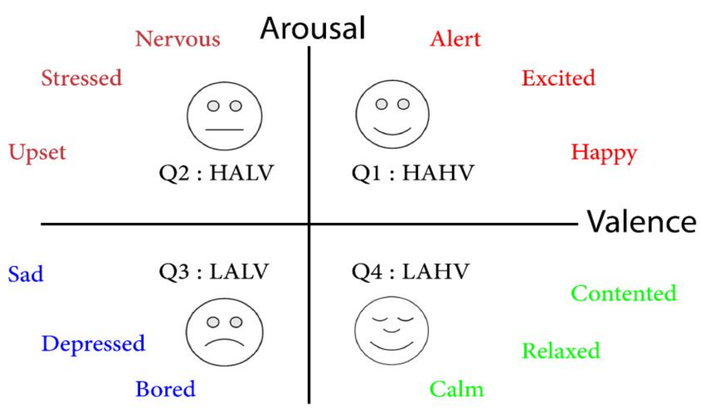

Fig 1. Valence-Arousal plane, taken from [3]. L is low, H is high, A is arousal, V is valence .

Methodology: data & models

- 3 languages: **English** [4], **Mandarin** [5], **Turkish** [6]
- Baseline feature sets: Essentia [7] & IS13 ComParE [8]
- Baseline Multi-Layer Perceptron (MLP) vs. End-to-end SincNet
- SincNet: CNN with **sinc filters** in the feature extractor
- Set-ups: **Within-dataset**, **Cross-dataset**, **Mixed training**, **Transfer Learning**
- Evaluation based on: weighted averages of precision (P), recall (R), f-score (F), accuracy (A) and confusion matrices

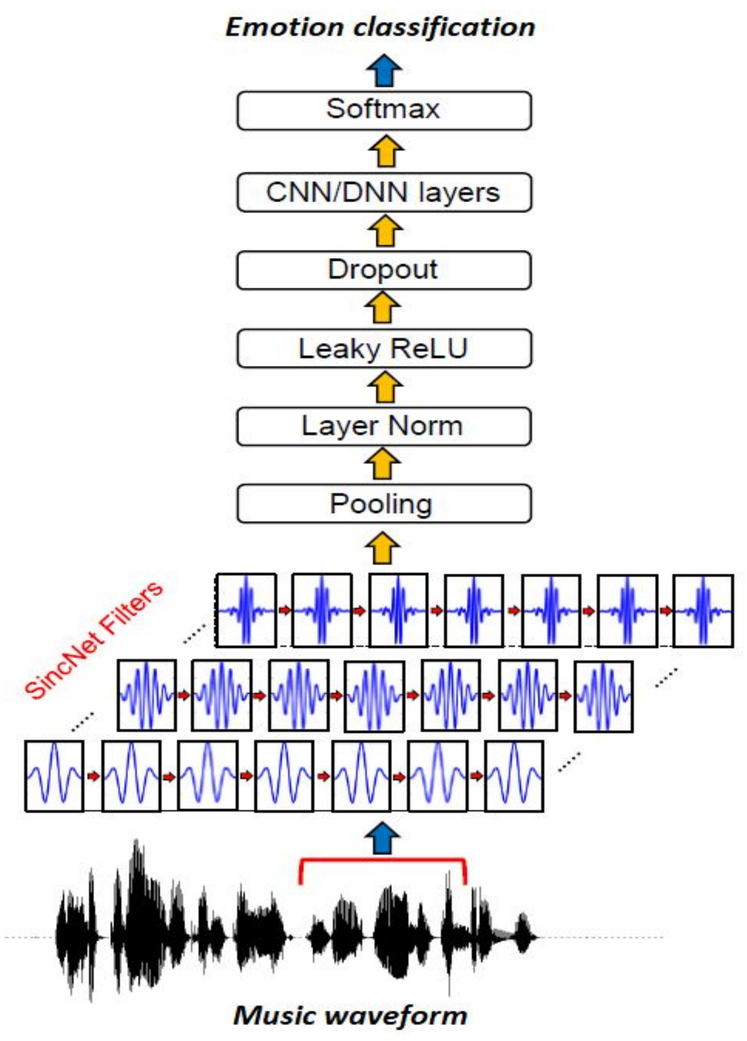

Fig 2. The SincNet architecture, adapted from [2]

Results: Baseline, SincNet, Cross-Dataset

- Baseline model: **MLP**, best from default traditional ML in Scikit-Learn [9]
- Baseline features: **IS13 ComParE**, with 65 low-level descriptors, statistics
- Best features were found to be more similar for English and Turkish
- Combined with Essentia => better scores for English, Turkish
- SincNet: 200ms frame size, 128 batch size, 100 epochs
- **Bigger frame size to better detect emotion**: 500ms frame size, 32 batch size, 100 epochs => better f-score and accuracy for English, Turkish
- Cross-dataset: train & test with different sets => general poor results
- Transfer learning 1* & 2*: **tune SincNet with the 3 sets in specific orders**: CH-TR-EN & TR-CH-EN for English, EN-TR-CH & TR-EN-CH for Mandarin, EN-CH-TR & CH-EN-TR for Turkish
- Transfer learning 3*: train with EN / TR & test with TR / EN => scores don't necessarily increase, thus CH also contributes to transfer learning
- Mixed dataset: train on combined datasets & test separately => **within-dataset is better, thus mixing brings noise**

Dataset		English (4Q-EMOTION)				Mandarin (CH-818)				Turkish (TR-MUSIC)			
Set-up	Model	P	R	F	A	P	R	F	A	P	R	F	A
Within dataset	MLP Essentia	58	56	55	56	26	24	**23**	24	71	70	70	70
Within dataset	MLP IS13 ComParE	65	63	63	63	23	**30**	**23**	**30**	74	71	71	71
Within dataset	MLP combined	**72**	**67**	**64**	**67**	18	19	18	19	**80**	**77**	**77**	**77**
Within dataset	SincNet 200ms	59	57	52	57	11	***27***	16	***27***	68	63	58	63
Within dataset	SincNet 500ms	58	58	56	58	3	18	6	18	68	66	66	66
Transfer learning 1*	SincNet 200ms	57	56	51	56	10	24	12	24	***75***	***75***	***75***	***75***
Transfer learning 2*	SincNet 200ms	***61***	***60***	***57***	***61***	**28**	26	17	26	72	71	71	71
Transfer learning 3*	SincNet 200ms	59	58	56	58	-	-	-	-	73	73	73	73
Mixed dataset	MLP IS13 ComParE	60	57	56	57	22	23	22	23	67	64	64	64
Mixed dataset	SincNet 200ms	58	57	56	57	26	23	***21***	23	61	51	50	51

Fig 3. Summary of results. Underlined are best scores for each dataset from all training set-ups, in italics are best scores with SincNet. *Transfer learning set-ups as described above.

- **Conclusions:** end-to-end models need more data and better fine-tuning & general cues are learned from different cultures, but **sensitivity is achieved with context-based fine-tuning**
- **Future work**: continuous valence and arousal values, better data curation, data augmentation, bigger chunk and batch sizes, fine-tuning SincNet, exploring other cultures

References

[1] Russell, "A circumplex model of affect," J. Personal. Soc. Psychol., 1980.
[2] Ravanelli and Bengio, "Speaker recognition from raw waveform with sincnet," IEEE Spoken Language Technology Workshop, 2018
[3] Soroush et al., "Emotion Classification through Nonlinear EEG Analysis Using Machine Learning Methods", Intern. Clinical Neuroscience Journal, 2018.
[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] Pedregosa et al., "Scikit-learn: Machine Learning in Python, JMLR 12, 2011.

Acknowledgments

This project was developed as part of a Master's Thesis at UPF and can be found and reproduced from: https://github.com/ana-pandrea/SincNet-MER

BCN Music Technology Group

upf. Universitat Pompeu Fabra Barcelona
MTG Music Technology Group