The multiple voices of musical emotions: source separation for improving music emotion recognition models and their interpretability

Jacopo de Berardinis[1,2] Angelo Cangelosi[1] Eduardo Coutinho[2]
[1] Machine Learning and Robotics Group, University of Manchester
[2] Applied Music Research Lab, University of Liverpool
jacopo.deberardinis@manchester.ac.uk https://github.com/jonnybluesman/emomucs

TLDR. EmoMucs is a deep neural network that considers the role of different musical voices in the prediction of the emotions induced by music. A source separation algorithm breaks up music signals into independent song elements (vocals, bass, drums, other) and end-to-end machine learning techniques are used for feature extraction and emotion modelling (valence and arousal regression). Results demonstrate that EmoMucs outperforms our baselines whilst providing insights into the relative contribution of different musical elements to the emotions perceived by listeners.

Music Emotion Recognition (MER)

MER: automatically predicting emotions from music.

- Perceived vs induced emotions
- How to represent emotions: categorical vs continuous space
- Static vs dynamic MER

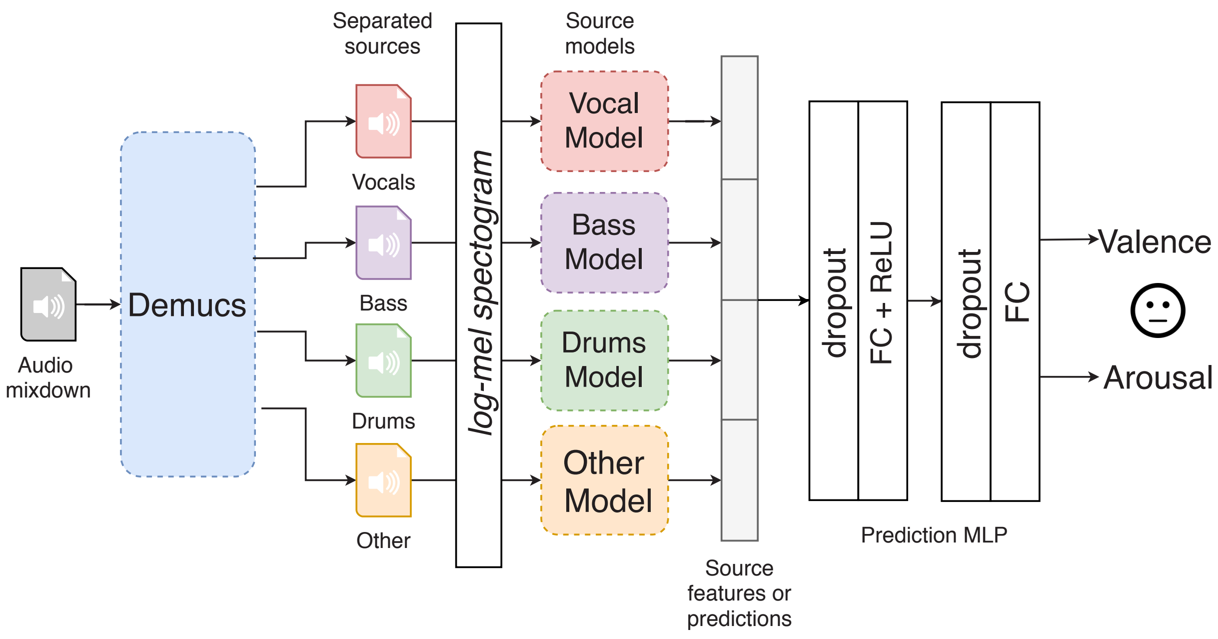
Why is MER difficult?

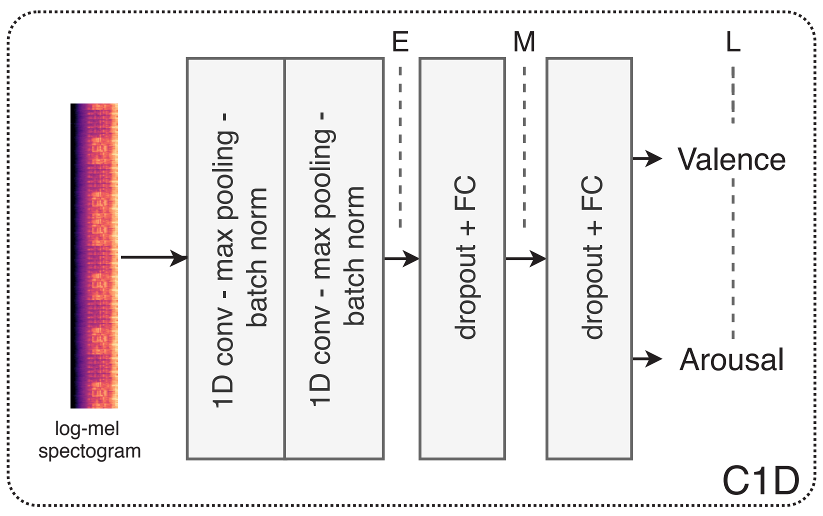
- Subjectivity
- Limited amount of data
- Data augmentation
- Interpretability comes at a cost

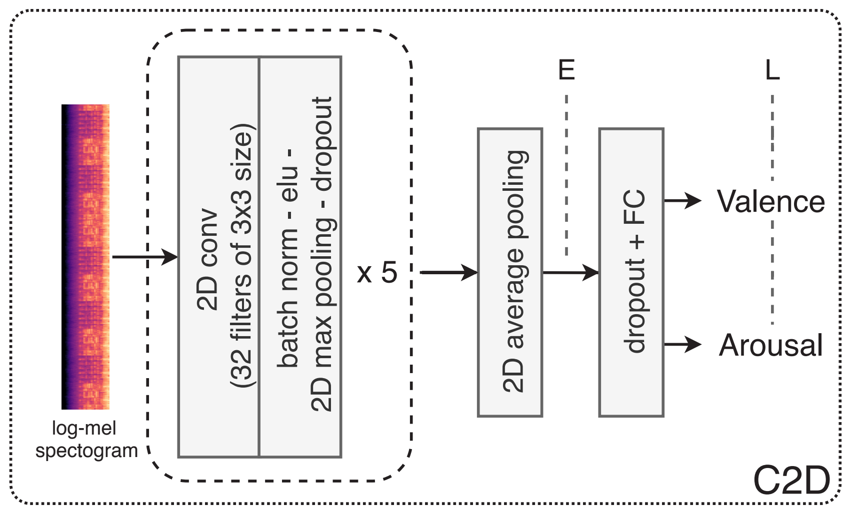
Pop music is harder for MER than classical music and soundtracks [1].

EmoMucs: a modular architecture based on the separation of sources and their distinct emotional impact

Method: *separate* with Demucs [2], *process* (each source separately), *aggregate* and *predict*. Each source model implemented as C1D or C2D.







How do we aggregate the output of each source model?
Three fusion strategies: early- (E), mid- (M), late- (L) level

Methodology and results

Dataset. PMEmo, the popular music with emotional annotations dataset [3], pre-processed as follows:

- 20s randomly selected clips from each chorus;
- zero-padding for 59 out of 794 tracks (≈4.3s);
- arousal and valence annotations in $[-1, 1]$;
- no data augmentation is performed.

Models under analysis

- EmoMucs-C1D and EmoMucs-C2D
- Baseline models on the mix-down: C1D-M, C2D-M
- Source models independently (e.g. C1D-V, C2D-V)
- Different combinations of source models

Training strategies for EmoMucs

- Full
- Fine-tune
- Freeze

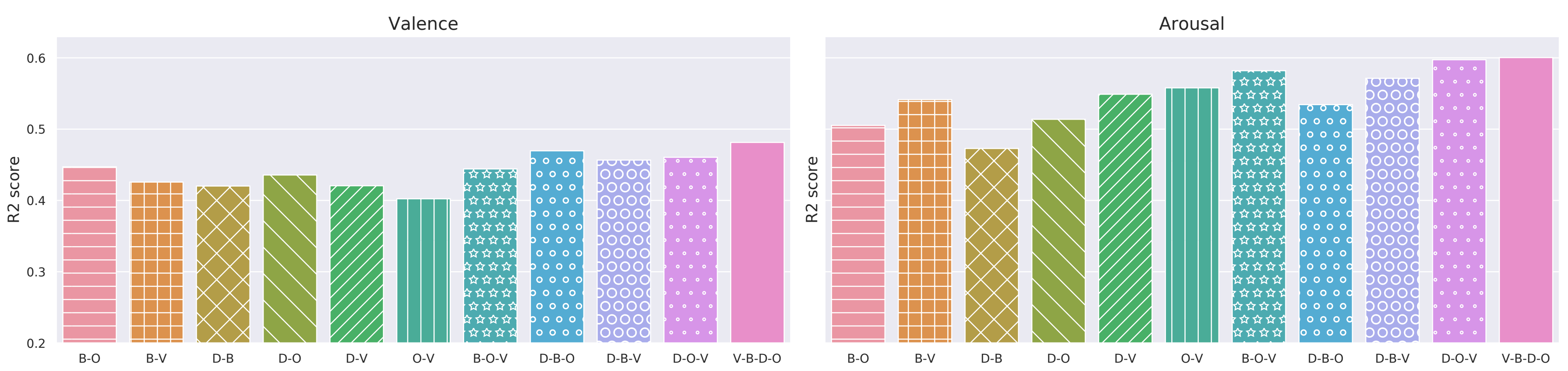
Evaluation metrics

- RMSE, the root-mean squared error (lower ⇒ better)
- R^2, the coefficient of determination (higher ⇒ better)

	RMSE		R^2	
Baseline	V	A	V	A
C1D-M	.2600	.2444	.3489	.5573
C2D-M	.2466	**.2285**	.4143	**.6100**

		Early				Mid				Late			
		RMSE		R^2		RMSE		R^2		RMSE		R^2	
EmoMucs	Training	V	A	V	A	V	A	V	A	V	A	V	A
w/ C1D	freeze	.2536	.2580	.3803	.5064	**.2428**	**.2435**	**.4332**	**.5615**	.2453	.2475	.4208	.5470
	finetune	.2562	.2624	.3655	.4878	.2516	.2492	.3875	.5395	na			
	full	.2536	.2628	.3787	.4850	.2625	.2651	.3371	.4794				
w/ C2D	freeze	.2373	**.2307**	.4584	**.6046**	na				**.2320**	**.2322**	**.4814**	**.6004**
	finetune	.2444	.2442	.4256	.5560					na			
	full	.2541	.2543	.3793	.5212								

→ Better performance for *valence*, comparable performance for *arousal*.



Conclusions

→ Same data, improved performance compared to current solutions
→ Tracing the relative contribution of each source at no cost
→ A modular architecture which can be further adapted for each source

Future work

- Specialisation of the source models (hyper-parameters)
- Attention mechanisms as aggregation strategy
- Alternatives to *Demucs* and singing voice separation techniques

Pasi Saari, Tuomas Eerola, and Olivier Lartillot.
Generalizability and simplicity as criteria in feature selection.
IEEE Transactions on Audio, Speech, and Language Processing, 19(6), 2010.

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach.
Music source separation in the waveform domain.
arXiv preprint arXiv:1911.13254, 2019.

Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun.
The pmemo dataset for music emotion recognition.
In *Proceedings of the 2018 International Conference on Multimedia Retrieval*