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

ABSTRACT

Despite the manifold developments in music emotion recognition and related areas, estimating the emotional impact of music still poses many challenges. These are often associated to the complexity of the acoustic codes to emotion and the lack of large amounts of data with robust golden standards. In this paper, we propose a new computational model (EmoMucs) that considers the role of different musical voices in the prediction of the emotions induced by music. We combine source separation algorithms for breaking up music signals into independent song elements (vocals, bass, drums, other) and end-to-end state-of-the-art machine learning techniques for feature extraction and emotion modelling (valence and arousal regression). Through a series of computational experiments on a benchmark dataset using source-specialised models trained independently and different fusion strategies, we demonstrate that EmoMucs outperforms state-of-the-art approaches with the advantage of providing insights into the relative contribution of different musical elements to the emotions perceived by listeners.

1. INTRODUCTION

The ability of music to express and induce emotions [15, 21] and act as a powerful tool for mood regulation [28] are well-known and demonstrable. Indeed, research shows that music listening is a commonly used, efficacious, and adaptable device to achieve regulatory goals [31], including coping with negative experiences by alleviating negative moods and feelings [17].

Crucial to this process is selecting the music that can facilitate the listener to achieve a determined mood regulation target, which often is not an easy task. In order to support listeners in this process, emotion-aware music recommendation systems became popular as they offer the possibility to explore large music libraries using affective cues. Indeed, recommending music based on the emotion of the listener at home [10] or background music personalised for the ones present in a restaurant would provide a more personal and enjoyable user experience [14].

At the core of these systems is music emotion recognition (MER), an active field of research in music information retrieval (MIR) for the past twenty years. The automatic prediction of emotions from music is a challenging task due to the subjectivity of the annotations and the lack of considerable data for effectively training supervised models. Song et al. [29] also argued that MER methods tend to perform well for genres such as classical music and film soundtracks, but not yet for popular music [25]. In addition, it is difficult to interpret emotional predictions in terms of musical content, especially for models based on deep neural networks. Although a few approaches exist for interpretable MER [5], the recognition accuracy of these methods is compromised, with the resulting performance loss often referred to as "cost of explainability".

As different voices within a composition can have a distinct emotional impact [13], our work leverages state-of-the-art deep learning methods for music source separation (MSS) to reduce the complexity of MER when limited training data is available. The proposed architecture (EmoMucs) is based on combining source separation methods with a parallel block of source-specialised models trained independently, subsequently aggregated with a fusion strategy. To benchmark our idea, we evaluated EmoMucs on the popular music with emotional annotations (PMEmo) dataset [38], and compared its performance with two reference deep learning models for MER. Experimental results demonstrate that our method achieves better performance for valence recognition, and comparable ones for arousal, while providing increased interpretability.

The main technical contributions are manifold: (i) we provide an in-depth evaluation of two reference models for MER, (ii) we propose a computational model that achieves an improved performance on the current baselines, under similar experimental conditions, and finally (iii) we show that our model provides no cost of interpretability.

The rest of the paper is structured as follows: Section 2 gives a primer on MER and an overview of related work on content-based methods, whilst Section 3 outlines the base-

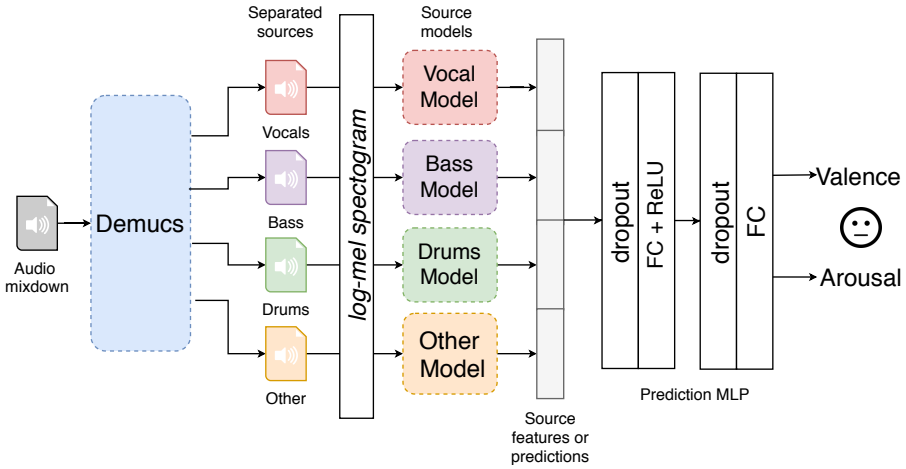

Figure 1. An overall architecture illustrating our proposed model, EmoMucs.

line architectures and our novel EmoMucs model. Section 4 details the experimental evaluation carried out, including our results on interpretability. Finally, Section 5 draws conclusions and gives direction for future work.

2. BACKGROUND AND RELATED WORK

2.1 A primer on music emotion recognition

Prior to introducing content-based methods for MER, we provide the reader with the fundamentals concepts of the task and refer to [3, 16, 34, 35] for a detailed overview.

Induced vs perceived emotions. Perceived emotions refers to the recognition of emotional meaning in music [29]. Induced (or felt) emotions refer to the feelings experienced by the listener whilst listening to music.

Annotation system. The conceptualisation of emotion with its respective emotion taxonomy remains a longstanding issue in MER [29]. There exist numerous emotion models, from miscellaneous [19] to domain specific [36], categorical [11] and dimensional [24], with the latter two being the prevailing ones. Whilst the categorical model focuses on all the emotions evolving from universal innate emotions like happiness, sadness, fear and anger [11], the dimensional model typically comprises an affective two-dimensional *valence-arousal* space. Valence represents a pleasure-displeasure continuum, whilst arousal outlines the activation-deactivation continuum [24].

Time scale of predictions. Predictions can either be static or dynamic. In the former case, the representative emotion of a song is given by one valence and arousal value [16]. Emotion annotations can also be obtained over time (e.g. second-by-second valence-arousal labels), thus resulting in dynamic predictions [27].

Audio features. Musical compositions consist of a rich array of features such as harmony, tempo, loudness and timbre and these all have an effect on emotion. Previous work in MIR has fuelled around developing informative acoustic features [16]. However, as illustrated in other works [18, 22, 26] and to the best of our knowledge, there exists no dominant single feature for MER.

2.2 Methods for content-based MER

The field of MIR has followed a similar path to other machine learning ones. Prior to the deep learning era, most methods relied on manual audio feature extraction. Huq et al. [12] give an overview on how musical features were traditionally extracted and fed into different architectures such as support vector machines, k-nearest neighbours, random forests, deep belief networks and other regression models. These methods were tested for MER on Russell's well-established arousal and valence emotion model [24].

These were succeeded with deep learning methods. Such techniques, like binarised and bi-directional long short-term memory recurrent neural networks (LSTM-RNN) and deep belief networks, have also been successfully employed for valence and arousal prediction [33]. Other methods again used LSTM-RNNs for dynamic arousal and valence regression, on the acoustic and psychoacoustic features obtained from songs [6]. Most of these works stemmed from entries in the MediaEval emotion challenge [2]. Multimodality has also been an interest for this research community, where [9] looked into MER based on both the audio signal and the lyrics of a musical track. Again, deep learning methods such as LSTM-RNNs are at the core of the architectures proposed.

An important factor in machine learning has been to build interpretable models, to make them applicable to a wider array of applications. To the best of our knowledge, only a few works have attempted to build an interpretable model for MER. In [37], different model classes were built over the extracted and selected features. These vital features were filtered and wrapped, followed by shrinkage methods. In [5], a deep network based on two-dimensional convolutions is trained to jointly predict "*mid-level perceptual features*", related to emotional qualities of music [1], with emotion classes in a categorical annotation space.

Our work focuses on the prediction of induced emotions at a global time scale (static predictions). This is done in a continuous annotation space, as adopting a categorical one would not exhibit the same richness in induced human emotion [35]. The idea of using MSS methods for MER was first investigated in [32]. Our work differs in

the following: (i) we focus on valence-arousal MER and address the former as a regression task; (ii) our methods rely on state-of-the-art deep learning methods with no need of traditional methods for audio feature extraction; (iii) we investigate different fusion strategies and training approaches, and (iv) we provide an insightful analysis for the interpretability of our models.

3. METHODOLOGY

Our approach is based on the observation that different musical sources in a composition can evoke distinct emotional responses from the listeners [13]. Given a music piece, they can contribute differently to the overall induced emotion. For instance, the bass and the vocal lines of a track can be more informative to predict valence, whereas drums might have more impact on arousal. Nevertheless, our aim here is not to provide a general explanation of the emotional influence of musical parts, as this is often an individualistic property belonging to each track.

Instead, we propose a computational model for MER based on a decomposition of the original audio signal to the possible sources (e.g. vocals, drums, bass) that can be detected from it. By doing so, it will be easier for the model to process the audio stream whilst searching for emotion-related patterns in every single source. The aggregation of the resulting source-specific models within a single architecture would also account for the possible inter-source relationships. This approach can thus be considered as a way to provide prior knowledge to a model in order to reduce the complexity of the learning task when limited data is available – a recurring issue in MER.

Considering the technical challenges in MER, the design of a computational model based on music source separation has the potential to (i) improve the performance of the current solutions with the same amount of training data; (ii) provide a modular architecture which can be further adapted and fine-tuned with respect to each source-specific module, and (iii) quantify the contribution of each source to the final prediction for improved interpretability.

Our model, EmoMucs, achieves this through a multiplexed framework for emotion recognition. The architecture of our model is illustrated in Figure 1, with its building blocks explained in the following subsections.

3.1 Music source separation module

In the final step of music production, the tracks corresponding to each individual instrument [1] are mixed together in a single audio file known as mix-down. Music source separation (MSS) aims at reconstructing the individual sources from a mix-down. A reference categorisation of these sources is the *SiSec Mus* evaluation campaign [30], which is based on the following classes: (i) *vocals*, (ii) *drums*, (iii) *bass* and (iv) *other*. Given a mix-down, the goal of a MSS model is to generate a waveform for each of the four original sources.

Most of the approaches for MSS operate on the spectrograms generated by the short-time Fourier transform (STFT). They are trained to produce a mask on the magnitude spectrums for each frame and source [8]. The output audio is then obtained through an inverse STFT on the masked spectrograms, reusing the input mixture phase. However, a technical limitation of these approaches is the information loss resulting from the mix of sources, which cannot be easily recovered through masking.

For this reason we use *Demucs* [8], a recent deep learning model for MSS directly operating on the raw input waveform. Instead of relying on a masking approach, Demucs is inspired by models for music generation in the waveform domain. It implements a U-net architecture with a convolutional encoder-decoder, and a bidirectional LSTM between them to increase the number of channels exponentially with depth [7].

Given an audio track, our system starts by feeding it to Demucs. This results into four different source tracks – one for each SiSec Mus class. To ensure comparability of our architecture with the baseline methods for MER, we compute the log-mel spectogram of each source track.

3.2 Source models and fusion strategies

As illustrated in Figure 1, the log-mel spectogram of each source track is then passed to the specific model associated to that source (e.g. the vocal's spectogram is fed to the vocal model). By disentanglement, each sub-model processes a single voice independently, and learns the corresponding source-specific musical features for emotion recognition. This approach thus provides a high degree of flexibility, as it makes it possible to design the architecture of each sub-model specifically for the corresponding source. Nonetheless, to guarantee a fair comparison of our architecture with the current methods for MER, we use one of our baselines as the architecture for all source models.

The baseline models are based on two common deep learning architectures for MER, illustrated in Figure 2. The first is a one-dimensional convolutional neural network (C1D) that was proposed in [9] as an audio model for multimodal MER. The architecture consists of two one-dimensional convolutional layers followed by max-pooling and batch normalisation. Its resulting feature maps are then passed to two fully-connected layers with dropout masks to improve generalisation. The second baseline comprises a VGG-style network, demonstrated to be effective in several MIR tasks [4]. This musically-engineered model, C2D, consists of 5 two-dimensional convolutional blocks, each separated by max pooling and dropout layers. The two-dimensional pooling operators progressively decrease the size of each feature map, while keeping the same number of kernels (32) after each block. After the convolutional blocks, two-dimensional average pooling is applied to ensure that the resulting feature map is of size 32×1. Following dropout, a single fully-connected layer is then employed to predict arousal and valence.

The architecture of our source models can be either C1D or C2D, resulting in two different implementations of EmoMucs – *EmoMucs-C1D* and *EmoMucs-C2D*. To yield a final prediction of valence (V) and arousal (A), the features from the source models are concatenated and passed

[1] We use *voice*, *instrument* and *source* interchangeably.

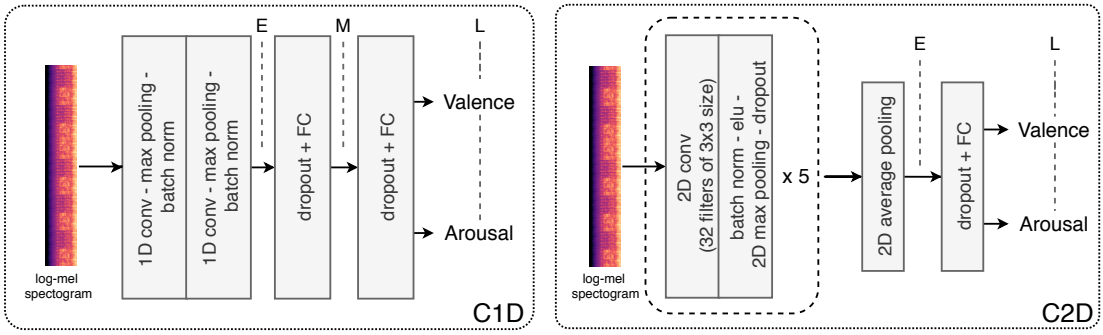

Figure 2. The baseline models our system C1D and C2D. These are the building blocks for the source models.

to two fully-connected layers with dropout. Assuming C1D is chosen as the architecture for our source models, there are three possible ways to access the features of a source model. This gives three fusion strategies: *early* (E), *mid* (M) and *late* (L), depicted in Figure 2. The first strategy considers the features obtained after the convolutional layers; *mid* fusion concatenates the features learned after the first fully-connected layers; and the *late* strategy considers the output of each source model, which correspond to the VA predictions. From the definition of C2D, the *mid*-level fusion strategy is not possible with this architecture.

3.3 Our training approach

Considering the different role of each source model, there are three main strategies for training EmoMucs: *full*, *freeze* and *fine-tune*. The first approach consists in training the whole network from scratch and propagating the gradient back to the source models from the last fully-connected layer of EmoMucs. In contrast, the last two strategies are based on pre-training each source model separately as a first step. The full network is then trained until the concatenation level, for the *freeze* mode, or until the first convolutional layer of each source model, for the *fine-tune* mode. This last choice can be considered as a sort of fine-tuning strategy and should be implemented with small learning rates to avoid the source models to catastrophically forget what they have already learned independently.

4. EXPERIMENTS

Our method is validated using the baseline models C1D and C2D trained on the mix-downs as reference models. These are denoted as C1D-M and C2D-M respectively. The performance of the baselines is then compared with each source model trained independently. In particular, we compare CXD-M with CXD-{V | B | D | O}, where V, B, D, O denote the *vocals*, *bass*, *drum*, and *other* sources respectively, and X is a placeholder for 1, 2. This allows to verify how informative each source model is, and whether one of them outperforms the correspondent mix-down baseline.

Secondly, we experiment with the different fusion and training strategies of EmoMucs using all the source models. Similarly to the previous case, the performances of EmoMucs with either C1D or C2D architectures for its source models (denoted as EmoMucs-C1D and EmoMucs-C2D) are compared to C1D-M and C2D-M.

We evaluate the accuracy of the valence-arousal predictions with the root-mean-squared error (RMSE) and the R^2 score. The latter is the coefficient of determination, with the best score being 1 when the variability of the target data is fully captured by the regressor. Conversely, a score equal to 0 corresponds to a model always predicting the expected value of the target. To avoid biasing our evaluations on a single test set, each run employs nested cross-validation with 5 splits for the outer and inner folds.

4.1 Dataset

As mentioned in [25, 29], the current methods for MER perform well for genres such as classical music and film soundtracks, but their performances are still poor for popular music. For this reason, we chose the *popular music with emotional annotations* (PMEmo) dataset [38] for our experiments. This collection contains valence-arousal induced emotions for 794 songs, annotated by 457 subjects, and also provides: song metadata, music chorus clips in MP3 format and pre-computed audio features.

For our experiments, we consider the static valence-arousal annotations. As our model needs raw-audio data to feed Demucs and generate the separated sources from a given mix-down, we use 20-second randomly selected clips from each chorus. For 59 tracks with duration shorter than 20 seconds, we apply zero padding at the end of the clip to ensure fixed-size input. On average, the padding operation is used to compensate for 4.35 seconds. The arousal and valence annotations are scaled to the $[-1, 1]$ interval for improving the stability of the model. We choose not to augment the dataset as such strategy can potentially affect the emotional impact of music on listeners.

4.2 Implementation details

We use Librosa 0.7.2 [20] for computing the log-mel spectograms from the tracks, with a fast Fourier transform (FFT) window size of 512, 256 samples between successive frames and 96 Mel bands. Our models are implemented in PyTorch [23], and the source code to replicate these experiments is available at `github.com/jonnybluesman/emomucs`.

4.3 Experimental results

The results of our experiments are reported in Tables 1 and 2. From Table 1, we notice that the C2D architecture is

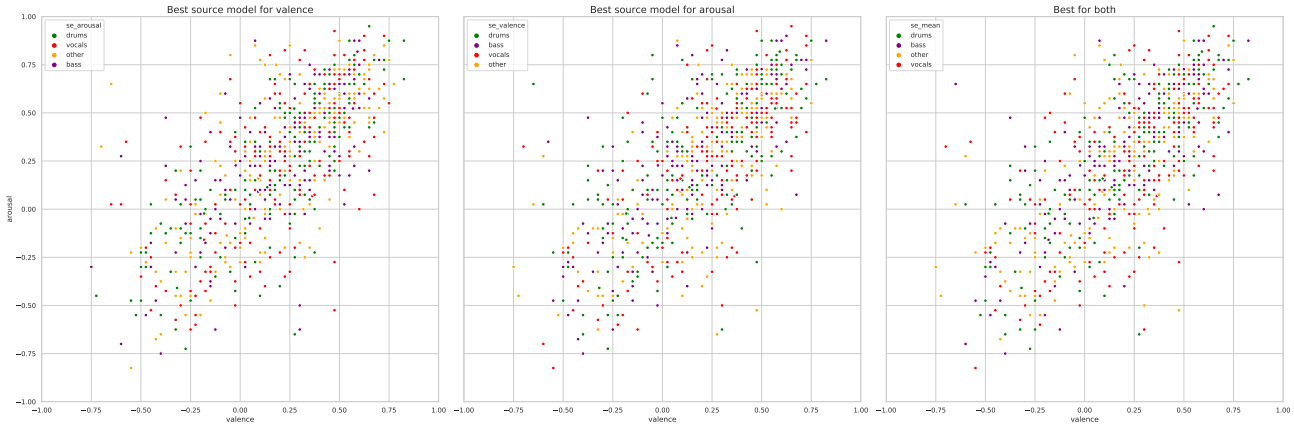

Figure 3. Target valence-arousal annotations of PMEmo in the $[-1, 1]$ interval, using colours to distinguish the C2D-based source model that better predicted each single annotation w.r.t. valence and arousal RMSE, together with their mean.

more accurate than C1D in all scenarios. In addition, the former baseline is also a more parsimonious model, due to the less number of parameters. All the baselines trained on the mix-down are considerably better than the source models considered independently. For both the architectures, we find that the drums model is the best between the source models at predicting valence, whereas the bass model is the worst for arousal. The performance disagreement for the source models using different architectures (e.g. C1D-V and C2D-V) suggests that the convolutional architecture plays a crucial role for the type of musical features that can be learned.

			RMSE		R2	
Baseline	# params	Input	V	A	V	A
C1D	86354	**M**	**.2600**	**.2444**	**.3489**	**.5573**
		V	.3048	.3214	.1131	.2426
		B	.2890	.3311	.2029	.1963
		D	.2710	.2961	.3000	.3572
		O	.2723	.2936	.2925	.3679
C2D	37698	**M**	**.2466**	**.2285**	**.4143**	**.6100**
		V	.2701	.2750	.3039	.4455
		B	.2762	.2924	.2718	.3732
		D	.2587	.2855	.3613	.4024
		O	.2633	.2748	.3381	.4462

Table 1. Evaluation of the baseline models trained on the mix-down (C1D-M, C2D-M) together with the corresponding source models. V, B, D, O denote *vocals*, *bass*, *drums* and *other* and they refer to the source models. Bold text highlights the best results for each baseline model. For both cases, the mix-down model achieves the best results.

As can be seen from Table 2, combining all source models in a single network has a crucial impact on the performance of the model. We conjecture that this is achieved by the architecture of EmoMucs, which makes it possible to account for all the possible inter-source relationships. In particular, EmoMucs-C1D with mid-level feature fusion and *freeze* mode training achieves a R^2 score of 0.4332 for valence, which is a considerable increase compared to 0.3489 for C1D. This is also reflected with a decrease of the RMSE for valence, accounting for 0.2428 instead of 0.26. Considering that the valence-arousal annotations are scaled to the $[-1, 1]$ interval, we divide these values by 2 for a more intuitive interpretation of the error. Hence, 0.2428 and 0.26 can be considered as errors of 12.14% and 13% in the annotation interval. Analogously, EmoMucs C2D with late fusion and *freeze* mode training achieves valence $R^2 = 0.4814$ and RMSE $= 0.2320$ (11.6%), instead of $R^2 = 0.4143$ and RMSE $= 0.2466$ (12.33%) for the C2D baseline. On the other hand, the arousal predictions of EmoMucs are comparable to those of the baselines.

4.4 Interpretability

As the architecture of EmoMucs is based on a concatenation of features learned by each source model at a specific layer, it is possible to trace the contribution of each voice as well as those emerging from their interrelated connections. This form of interpretability is architecturally supported by our deep neural network, and it comes at no performance loss. This contrasts the work of Chowdhury et al. [5], who measured the "cost of explainibility" of their model by trading accuracy for interpretability.

A simple way to interpret EmoMucs is to isolate the performance of each model independently, as done in Table 1. It is also compelling to analyse the regression accuracy for each track in the dataset and visualise them together with the target annotations in the valence-arousal space. In Figure 3, this is done separately for valence and arousal by associating each target data point with a colour related to its best source model (the one with lowest valence and arousal RMSE for that target). If source models specialise in certain regions of the annotation space, e.g. drums and high arousal, we would expect them to form clusters in the annotation space. However, this hypothesis is rejected as Figure 3 does not suggest any clear specialisation of the source models in the annotation space. This supports our previous observation that each track has intrinsic features related to its emotional impact. For instance, two distinct tracks with very similar annotations can have a considerably different emotional influence from their sources.

Figure 4 reports the performance (R^2 score for valence

		Early				Mid				Late			
		RMSE		R2		RMSE		R2		RMSE		R2	
Model	Training	V	A	V	A	V	A	V	A	V	A	V	A
EmoMucs-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				
EmoMucs-C2D	freeze	.2373	**.2307**	.4584	**.6046**	na				**.2320**	**.2322**	**.4814**	**.6004**
	finetune	.2444	.2442	.4256	.5560					na			
	full	.2541	.2543	.3793	.5212								

Table 2. Comparison of EmoMucs models with different fusion and training strategies.

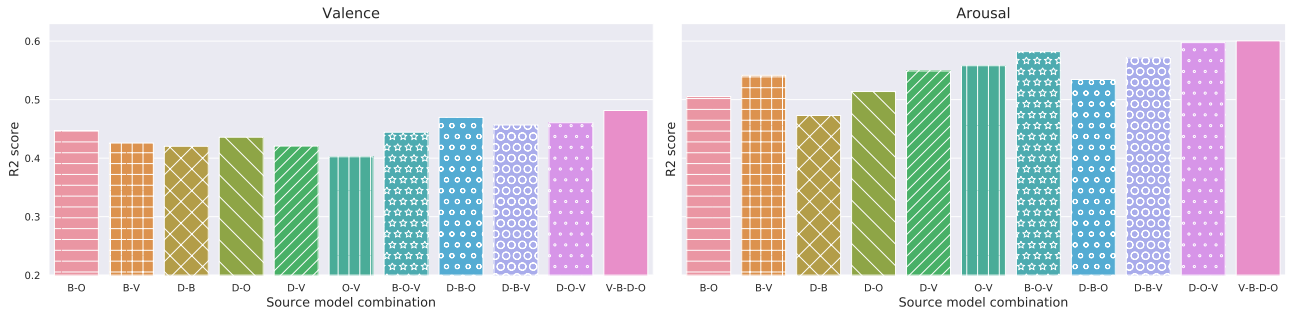

Figure 4. EmoMucs-C2D trained with different combinations of source models.

and arousal) of EmoMucs-C2D using late fusion and *freeze* training mode for different combinations of source models. The best performance is achieved when using all sources (V-B-D-O). The contribution of source models varies with their combination. When using two sources, the combination of the *bass* and the *other* models gives better performance in the valence space, but for the arousal space an improved result is achieved when combining *vocals* with *other*. When three sources are considered, the combination of *drums*, *bass* and *other* achieves the best R^2 for valence, whereas, for arousal, excluding *bass* gives comparable results to EmoMucs-C2D with all the sources.

5. CONCLUSIONS

The task of computational music emotion recognition (MER) is particularly challenging due to several factors such as subjectivity within annotations, scarcity of labelled data for training supervised models, and inadequate data augmentation strategies. There is common belief that the current models perform well for classical music and film soundtracks, but their performances are still poor for popular music. To the best of our knowledge, improving the interpretability of MER models jeopardises their performance, thus introducing a "cost of explainability".

In this paper we introduced EmoMucs, a deep learning architecture for MER based on music source separation. First, our method separates the audio signal into different sources associated to vocals, drums, bass and other voices of the mix-downs. Different sub-models are then used to process each source independently, with their features being aggregated according to a fusion strategy.

We evaluated EmoMucs on the popular music with emotional annotations (PMEmo) dataset, and compared its performance with two common deep learning models for MER trained on the mix-downs. Our results demonstrate that EmoMucs outperforms the baseline models for valence, and achieves comparable performance for arousal, while providing increased interpretability.

Our work achieves the following: (i) improved performance of the current solutions with the same amount of training data; (ii) a modular architecture which can be further adapted and fine-tuned with respect to each source-specific module, and (iii) a quantified contribution of each source to the final prediction for more interpretability.

The implementation of EmoMucs considered in our experiment is designed to prioritise the comparability of our approach to other baselines. In our future endeavours, we plan on optimising the architecture and hyper-parameters of each source model in order to specialise their design to the corresponding sources. Additionally, a study based on the analysis of the activations of the fusion layer would provide more detailed insights regarding the contribution of each source-model, thereby increasing the interpretability of our method and its potential applications.

6. REFERENCES

[1] Anna Aljanaki and Mohammad Soleymani. A data-driven approach to mid-level perceptual musical feature modeling. *arXiv preprint arXiv:1806.04903*, 2018.

[2] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. *PLOS ONE*, 12:e0173392, 03 2017.

[3] Mathieu Barthet, György Fazekas, and Mark Sandler. Music emotion recognition: From content- to context-based models. In *CMMR*, 2012.

[4] Keunwoo Choi, George Fazekas, Mark B. Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks. *ArXiv*, abs/1703.09179, 2017.

[5] Shreyan Chowdhury, Andreu Vall, Verena Haunschmid, and Gerhard Widmer. Towards explainable music emotion recognition: The route via mid-level features. In *ISMIR*, 2019.

[6] Eduardo Coutinho, George Trigeorgis, Stefanos Zafeiriou, and Björn Schuller. Automatically estimating emotion in music with deep long-short term memory recurrent neural networks. pages 1–3, 01 2015.

[7] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 933–941. JMLR. org, 2017.

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

[9] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. Music mood detection based on audio and lyrics with deep neural net. In *ISMIR*, 2018.

[10] S. Dornbush, K. Fisher, K. McKay, A. Prikhodko, and Z. Segall. Xpod - a human activity and emotion aware mobile music player. In *2005 2nd Asia Pacific Conference on Mobile Technology, Applications and Systems*, pages 1–6, 2005.

[11] Paul Ekman. An argument for basic emotions. 1992.

[12] Arefin Huq, Juan Pablo Bello, and Robert Rowe. Automated music emotion recognition: A systematic evaluation. *Journal of New Music Research*, 39(3):227–244, 2010.

[13] David Huron, Neesha Anderson, and Daniel Shanahan. "you can't play a sad song on the banjo:" acoustic factors in the judgment of instrument capacity to convey sadness. *Empirical Musicology Review*, 9(1):29–41, 2014.

[14] Alejandro Jaimes, Nicu Sebe, and Daniel Gatica-Perez. Human-centered computing: A multimedia perspective. pages 855–864, 01 2006.

[15] Patrik N Juslin and John A. Sloboda. Handbook of music and emotion: Theory, research, applications. 2011.

[16] Youngmoo Kim, Erik Schmidt, Raymond Migneco, Brandon Morton, Jeffrey Scott, Jacquelin Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010*, 01 2010.

[17] Emmanuel Kuntsche, Lydie Le Mével, and Irina Berson. Development of the four-dimensional motives for listening to music questionnaire (mlmq) and associations with health and social issues among adolescents. *Psychology of Music*, 44:219–233, 2016.

[18] Karl F MacDorman, Stuart Ough Chin-Chang Ho. Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research*, 36(4):281–299, 2007.

[19] Stephen Mcadams, Bradley Vines, Sandrine Vieillard, B. Smith, and R. Reynolds. Influences of large-scale form on continuous ratings in response to a contemporary piece in a live concert setting. *Music Perception*, 22:297–350, 12 2004.

[20] Brian McFee, Vincent Lostanlen, Matt McVicar, Alexandros Metsai, Stefan Balke, Carl Thomé, Colin Raffel, Ayoub Malek, Dana Lee, Frank Zalkow, Kyungyun Lee, Oriol Nieto, Jack Mason, Dan Ellis, Ryuichi Yamamoto, Scott Seyfarth, Eric Battenberg, Rachel Bittner, Keunwoo Choi, Josh Moore, Ziyao Wei, Shunsuke Hidaka, nullmightybofo, Pius Friesch, Fabian-Robert Stöter, Darío Hereñú, Taewoon Kim, Matt Vollrath, and Adam Weiss. librosa/librosa: 0.7.2, January 2020.

[21] Leonard B Meyer. *Emotion and meaning in music*. University of chicago Press, 2008.

[22] Luca Mion and Giovanni De Poli. Score-independent audio features for description of music expression. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):458–466, 2008.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[24] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

[25] Pasi Saari, Tuomas Eerola, and Olivier Lartillot. Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1802–1812, 2010.

[26] Erik M Schmidt and Youngmoo E Kim. Prediction of time-varying musical mood distributions from audio. In *ISMIR*, pages 465–470, 2010.

[27] Erik M Schmidt, Douglas Turnbull, and Youngmoo E Kim. Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the international conference on Multimedia information retrieval*, pages 267–274, 2010.

[28] Thomas Schäfer, Peter Sedlmeier, Christine Städtler, and David Huron. The psychological functions of music listening. *Frontiers in Psychology*, 4:511, 2013.

[29] Yading Song and D Simon. How well can a music emotion recognition system predict the emotional responses of participants? In *Sound and Music Computing Conference (SMC)*, pages 387–392, 2015.

[30] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 293–305. Springer, 2018.

[31] Annelies van Goethem and John Sloboda. The functions of music for affect regulation. *Musicae Scientiae*, 15(2):208–228, 2011.

[32] Jieping xu, Xirong Li, Yun Hao, and Gang Yang. Source separation improves music emotion recognition. 04 2014.

[33] Mingxing Xu, Xinxing Li, Haishu Xianyu, Jiashen Tian, Fanhang Meng, and Wenxiao Chen. Multi-scale approaches to the mediaeval 2015 "emotion in music" task. In *MediaEval*, 2015.

[34] Xinyu Yang, Yizhuo Dong, and Juan Li. Review of data features-based music emotion recognition methods. *Multimedia Systems*, 24:365–389, 2017.

[35] Yi-Hsuan Yang and Homer H. Chen. Machine recognition of music emotion: A review. *ACM Trans. Intell. Syst. Technol.*, 3:40:1–40:30, 2012.

[36] Marcel Zentner, Didier Grandjean, and Klaus Scherer. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion (Washington, D.C.)*, 8:494–521, 09 2008.

[37] JiangLong Zhang, Huang Xianglin, Lifang Yang, and Liqiang Nie. Bridge the semantic gap between pop music acoustic feature and emotion: build an interpretable model. *Neurocomputing*, 208, 06 2016.

[38] Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. The pmemo dataset for music emotion recognition. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR '18, page 135–142, New York, NY, USA, 2018. Association for Computing Machinery.