TOWARDS EXPLAINABLE MUSIC EMOTION RECOGNITION: THE ROUTE VIA MID-LEVEL FEATURES

Shreyan Chowdhury Andreu Vall Verena Haunschmid Gerhard Widmer
Institute of Computational Perception, Johannes Kepler University Linz, Austria
firstname.lastname@jku.at

ABSTRACT

Emotional aspects play an important part in our interaction with music. However, modelling these aspects in MIR systems have been notoriously challenging since emotion is an inherently abstract and subjective experience, thus making it difficult to quantify or predict in the first place, and to make sense of the predictions in the next. In an attempt to create a model that can give a musically meaningful and intuitive explanation for its predictions, we propose a VGG-style deep neural network that learns to predict emotional characteristics of a musical piece together with (and based on) human-interpretable, mid-level perceptual features. We compare this to predicting emotion directly with an identical network that does not take into account the mid-level features and observe that the loss in predictive performance of going through the mid-level features is surprisingly low, on average. The design of our network allows us to visualize the effects of perceptual features on individual emotion predictions, and we argue that the small loss in performance in going through the mid-level features is justified by the gain in explainability of the predictions.

1. INTRODUCTION

Emotions – portrayed, perceived, or induced – are an important aspect of music. MIR systems can benefit from leveraging this aspect because of its direct impact on human perception of music, but doing so has been challenging due to the inherently abstract and subjective quality of this feature. Moreover, it is difficult to interpret emotional predictions in terms of musical content. In our quest for computer systems that can give musically or perceptually meaningful justifications for their predictions [17], we turn to the notion of *'mid-level perceptual features'* as recently described and advocated by several researchers [1, 5]. These are musical qualities (such as rhythmic complexity, or perceived major/minor harmonic character) that are supposed to be musically meaningful and intuitively recognizable by most listeners, without requiring music-theoretic knowledge. It has been shown previously that there is considerable consistency in human perception of these features; that they can be predicted relatively well from audio recordings; and that they also relate to the perceived emotional qualities of the music [1].

That is the motivation for the work to be reported here. Our goal is to use mid-level features as a basis for providing explanations of (and thus get further insights into, or handles on) a model's emotion predictions, by training it to recognize mid-level qualities from audio, and predict emotion ratings from the mid-level predictions. Further, we wish to quantify the cost – in terms of loss of predictive performance – incurred by this detour. We will call this the 'cost of explainability'.

Focusing our study on a specific benchmark dataset labelled with both perceived emotional qualities and mid-level perceptual features, we first establish our basic VGG-style model architecture [16], showing that it can learn the two individual prediction tasks (mid-level from audio, emotion from mid-level) from appropriate ground-truth data, with accuracies that are at least on par with previously published models. We then present a network with nearly identical architecture that learns to predict emotional characteristics of a piece by explicitly going through a mid-level feature prediction layer. We compare this to predicting emotion directly, using an identical network with the exception of the mid-level layer, and find that the cost of going through the mid-level features is surprisingly low, on average. Finally, we show that by training the network to learn to predict mid-level and emotions jointly, the results can be further improved. A graphical overview of the general scenario is shown in Figure 1.

The fact that in our model, emotions are predicted from the mid-level by a single fully-connected layer, allows us to measure the effects of each of the features on each emotion prediction, providing the basis for interpretability and simple explanations. There are a number of application scenarios in which we believe this could be useful; some of these will be briefly discussed in the final section.

The remainder of this paper is structured as follows: Section 2 briefly discusses related work on which this research is based. In Section 4 the datasets that provide us with emotion and mid-level annotations are described. The three different approaches to modelling emotion are summarized in Section 5. Experimental results and a demonstration of interpretability are given in Sections 6 and 7.

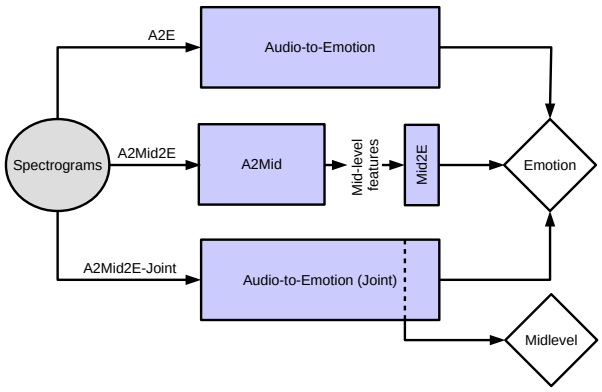

Figure 1: Three different architectures are compared for predicting emotion from audio.

We discuss our findings and conclude in Section 8.

2. RELATED WORK

In the MIR field, audio-based music emotion recognition (MER) has traditionally been done by extracting selected features from the audio and predicting emotion based on subsequent processing of these features [7]. Methods such as linear regression, regression trees, support vector regression, and variants have been used for prediction as mentioned in the systematic evaluation study by Huq et al [6]. Techniques using regression-like algorithms have generally focused on predicting arousal and valence as per the well-known Russell's *circumplex model* of emotion [15]. Deep learning methods have also been employed for predicting arousal and valence, for example [18], that investigated BLSTM-RNNs in tandem with other methods, and [3], that used LSTM-RNNs. Others such as [12] and [9] use support vector classification to predict the emotion class. Aljanaki et al. [2] provide a summary of entries to the MediaEval emotion characterization challenge and quote results for arousal and valence prediction.

Deep neural networks are preferable for many tasks due to their high performance but can be considered black boxes due to their non-linear and nested structure. While in some fields such as healthcare or criminal justice the use of predictive analytics can have life-affecting consequences [14], the decisions of MIR models are generally not as severe. Nevertheless, also in MIR it would be desirable to be able to obtain explanations for the decisions of a music recommendation or search system, for various reasons (see also Section 8). Many current methods for obtaining insights into deep network-based audio classification systems do not explain the predictions in a human understandable way but rather design special filters that can be visualized [13], or analyze neuron activations [8]. To the best of our knowledge, [19] is the only attempt to build an interpretable model for MER. They performed the task of feature extraction and selection and built models from different model classes on top of them. The only interpretation offered is the reporting of coefficients from their logistic regression models, without further explanation.

3. MID-LEVEL PERCEPTUAL FEATURES

The notion of *('mid-level') perceptual features* for characterizing music recordings has been put forward by several authors, as an alternative to purely sound-based or statistical low-level features (e.g., MFCCs, ZCR, spectral centroid) or more abstract music-theoretic concepts (e.g., meter, harmony). The idea is that they should represent musical characteristics that are easily perceived and recognized by most listeners, without any music-theoretical training. Research on such features has quite a history in the fields of music cognition and psychology (see [5] for a compact discussion). Such features are attractive for our purposes because they could provide the basis for intuitive explanations of a MIR system's decisions, relating as they do to the musical experience of most listeners.

Various sets of such perceptual features have been proposed in the literature. For instance, Friberg et al.'s set [5] contains such concepts as speed, rhythmic clarity/complexity, articulation, dynamics, modality, overall pitch hight, etc. In our study, we will be using the seven mid-level features defined by Aljanaki & Soleymani [1], because they come with an openly available set of annotated audios (see below). We recapitulate the features and their definitions in Table 1, for convenience.

4. DATASETS

For our experiments, we need music recordings annotated both with mid-level perceptual features, and with human ratings along some well-defined emotion categories. Our starting point is Aljanaki & Soleymani's *Mid-level Perceptual Features* dataset [1], which provides mid-level feature annotations. For the actual emotion prediction experiments, we then use the *Soundtracks* dataset, which is contained in the Aljanaki collection as a subset, and comes with numeric emotion ratings along 8 dimensions.

4.1 Mid-level Perceptual Features Dataset

The *Mid-level Perceptual Features Dataset* [1] consists of 5000 song snippets of around 15 seconds each annotated according to the seven mid-level descriptors listed in Table 1. The annotators were required to have some musical education and were selected based on passing a musical test. The ratings range from 1 to 10 and were scaled by a factor of 0.1 before being used for our experiments.

4.2 Emotion Ratings: The Soundtracks Dataset

The *Soundtracks (Stimulus Set 1)* [1] dataset, published by Eerola and Vuoskoski [4], consists of 360 excerpts from 110 movie soundtracks. The excerpts come with expert ratings for five categories following the discrete emotion model (happy, sad, tender, fearful, angry) and three categories following the dimensional model (valence, energy,

[1] https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/projects2/past-projects/coe/materials/emotion/soundtracks

Perceptual Feature	Question asked to human raters
Melodiousness	To which excerpt do you feel like singing along?
Articulation	Which has more sounds with staccato articulation?
Rhythmic Stability	Imagine marching along with the music. Which is easier to march along with?
Rhythmic Complexity	Is it difficult to repeat by tapping? Is it difficult to find the meter? Does the rhythm have many layers?
Dissonance	Which excerpt has noisier timbre? Has more dissonant intervals (tritones, seconds, etc.)?
Tonal Stability	Where is it easier to determine the tonic and key? In which excerpt are there more modulations?
Modality (‘Minorness’)	Imagine accompanying this song with chords. Which song would have more minor chords?

Table 1: Perceptual mid-level features as defined in [1], along with questions that were provided to human raters to help them interpret the concepts. (The ratings were collected in a pairwise comparison scenario.) In the following, we will refer to the last one (*Modality*) as *‘Minorness’*, to make the core of the concept clearer.

tension). This makes it a suitable dataset for musically conveyed emotions [4]. The ratings in the dataset range from 1 to 7.83 and were scaled by a factor of 0.1 before being used for our experiments. As stated above, all the songs in this set are also contained in the Mid-level Features Dataset, so that both kinds of ground truth are available.

5. AUDIO-TO-EMOTION MODELS

In the following, we describe three different approaches to modeling emotion from audio, all based on VGG-style convolutional neural networks (CNNs). The architectures are summarized in Figure 2. For all models, we use an Adam optimizer with a learning rate of 0.0005 and a batch size of 8, and employ early stopping with a patience of 50 epochs to prevent overfitting.

In terms of preprocessing, the audio samples are first converted into 149-point spectrograms calculated on randomly selected 10-second sections of the original snippets. The audio is resampled at 22.05 kHz, with a frame size of 2048 samples and a frame rate of 31.25 frames per second, and amplitude-normalized before computing the logarithmic-scaled spectrogram. This results in input vectors of size 313×149. These spectrograms are used as inputs for the following model architectures.

5.1 A2E Scheme

The first model, which we term “A2E", is the most straightforward one. The spectrograms are fed into a VGG-style CNN to directly predict emotion values from audio. This is the leftmost path in Figure 2. This model is not interpretable due to its black box architecture and is used as a baseline and for computing the *cost of explainability* when comparing to more interpretable but possibly worse performing models.

5.2 A2Mid2E Scheme

In order to obtain a more interpretable model, an intermediate step is introduced. First, a VGG-style network is used to predict mid-level features from audio. This model is trained on the mid-level features dataset described in Section 4.1 above. Next, a linear regression model is trained to predict the 8 emotion ratings in the Soundtracks dataset from the 7 mid-level feature values that we get as an output from the mid-level predictor network. This corresponds to a fully connected layer with 7 input units and 8 outputs and linear (identity) activation function – see the middle path in Figure 2. We call this scheme “A2Mid2E". A linear model is chosen because its weights can easily be interpreted to understand the importance of each mid-level feature in predicting the emotion ratings.

5.3 A2Mid2E-Joint Scheme

In an attempt to replace the step-wise training of two separate models with a single model that, ideally, could learn an internal representation useful for both prediction tasks, while keeping the interpretability of the linear weights, we propose a third architechture, called “A2Mid2E-Joint" (rightmost path in Figure 2). This network learns to predict mid-level features and emotion ratings jointly, but still predicts the emotions directly from the mid-level via a linear layer. This is achieved by the second last layer having exactly the same number of units as there are mid-level features (7), followed by a linear output layer with 8 outputs. From this network, we extract two outputs – one from the second last layer ("mid-level layer"), and one from the last layer ("emotion layer"). We compute losses for both the outputs and optimize the combined loss (summation of both the losses).

6. EXPERIMENTS

The audio clips are preprocessed as described in Section 5 to obtain the input spectrograms. During training, one ran-

	Valence	Energy	Tension	Anger	Fear	Happy	Sad	Tender	Avg.
Mid2E (Aljanaki)	0.88	0.79	0.84	0.65	0.82	0.81	0.73	0.72	0.78
Mid2E (Ours)	0.88	0.80	0.84	0.65	0.82	0.81	0.74	0.73	0.79
A2E	0.81	0.79	0.84	0.82	0.81	0.66	0.60	0.75	0.76
A2Mid2E	0.79	0.74	0.78	0.72	0.77	0.64	0.58	0.67	0.71
A2Mid2E-Joint	0.82	0.78	0.82	0.76	0.79	0.65	0.64	0.72	0.75
$CoE_{A2Mid2E}$	0.02	0.05	0.06	0.10	0.03	0.02	0.02	0.08	0.05
$CoE_{A2Mid2E\text{-}Joint}$	-0.02	0.01	0.02	0.06	0.02	0.01	-0.04	0.03	0.01

Table 2: Summaries of the different model performances on predicting emotion. The last two rows show the "cost of explainability" (CoE), as the difference between our baseline (A2E) and the newly proposed models (A2Mid2E, A2Mid2E-Joint). A positive cost indicates a loss in performance.

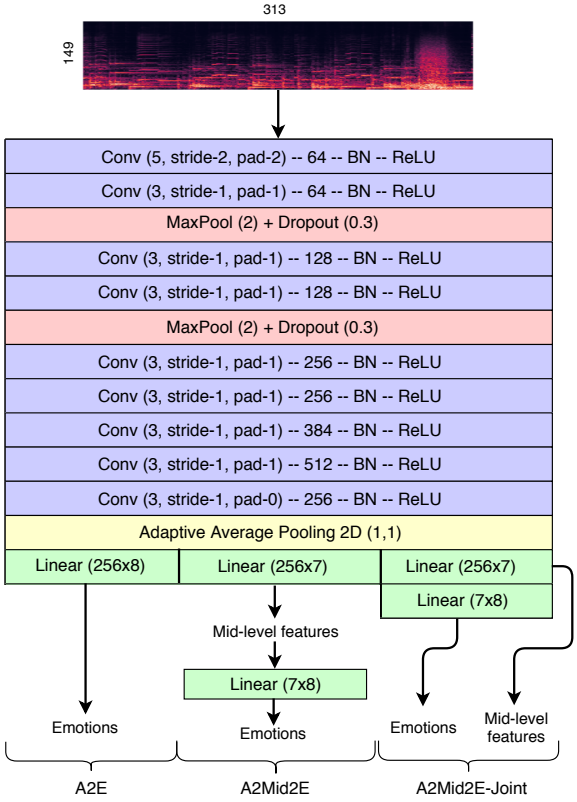

Figure 2: The same architecture from the first layer up to the 'Adaptive Average Pooling 2D' layer is shared by all networks.

dom 10-second snippet from each spectrogram is taken as input. We optimize the mean squared error, and use Pearson's correlation coefficient as the evaluation metric for emotion rating prediction. Each of the paths (A2E, A2Mid2E, A2Mid2E-Joint) is run ten times and the average correlation values are reported. Each run has a different seed which reshuffles the train-test split.

6.1 Verifying our Basic Architecture

Before going any further, we first want to verify that our VGG-style network model performs on par with comparable methods on the basic component tasks of predicting mid-level features from audio (A2Mid) and emotions from (given) mid-level features (Mid2E).

For the A2Mid scenario (Table 3), we train in two ways: first on the entire Mid-level features dataset with 8% test set selected as described in [1]. We call this A2Mid+. This is the result that should be directly compared to column 1. Second, we train only on the songs from the Soundtracks dataset with 20% test set, and call this A2Mid. The column 'Joint' in Table 3 gives the mid-level predictions produced by our A2Mid2E-Joint model. As can be seen, our models are broadly comparable to the results reported in [1], with A2Mid+ and *Joint* performing slightly better, on average.

Regarding the prediction of emotions from given mid-level feature annotations (Table 2, first two rows), there is not much space for deviation, as the models used by Aljanaki [1] and us are very simple.

Mid-level feature	Aljanaki	A2Mid+	A2Mid	Joint
Melodiousness	0.70	0.70	0.69	0.72
Articulation	0.76	0.83	0.84	0.79
R. Stability	0.46	0.39	0.39	0.34
R. Complexity	0.59	0.66	0.45	0.46
Dissonance	0.74	0.74	0.73	0.74
Tonal Stability	0.45	0.56	0.61	0.63
Minorness	0.48	0.55	0.51	0.57

Table 3: Correlation values for mid-level features predictions using our models, compared with those reported by Aljanaki et al. [1].

6.2 Quantifying the Cost of Explainability

We now compare our three model architectures (A2E, A2Mid2E, A2Mid2E-Joint) on the full task of predicting emotion from audio, We train on the Soundtracks dataset as described above, with 10 runs with randomly selected train-test 80:20 splits. The A2E model serves as reference for the subsequent models with explainable linear layers. The results can be found in Table 2.

In the case of direct emotion prediction (A2E), the final layer is connected to 256 input nodes. However, in the A2Mid2E scheme, due to the fact that we introduce a bottleneck (viz. the 7 mid-level predictions) as inputs to the subsequent linear layer predicting emotions, our hypothesis is that doing so should result in a decrease in the perfor-

mance of emotion prediction. We calculate this cost as the difference in correlation coefficients between the two models for each emotion. The results (rows A2E and A2Mid2E in Table 2) reflect the expected trend, but as can be seen, the decrease in performance is quite small (less than 7% of the original correlation coefficient on average).

6.3 Joint Learning of Mid-level and Emotions

Further improvements in the performance of the mid-level-based network can be obtained by training jointly on the mid-level and emotion annotations (as described in Section 5.3). This model reduces the cost even further, as can be seen in Table 2, row A2Mid2E-Joint. The decrease in performance is now less than 1.5% of the correlation coefficients for the A2E case on average. We believe this is acceptable in view of the possibility of obtaining explanations from this network (see below).

7. OBTAINING EXPLANATIONS

Since the mapping between mid-level features and emotions is linear in both proposed schemes (A2Mid2E, A2Mid2E-Joint), it is now straightforward to create human-understandable explanations. Linear models can be interpreted by analyzing their weights: increasing a numerical feature by one unit changes the prediction by its weight. A more meaningful analysis is to look at the *effects*, which are the weights multiplied by the actual feature values [10]. An effects plot shows the distribution, over a set of examples, of the effects of each feature on each target. Each dot in an effects plot can be seen as the amount this feature contributes (in combination with its weight) to the prediction, for a specific instance. Instances with effect values closer to 0 get a prediction closer to the intercept (bias term). Figure 3 shows the effects of the model A2Mid2E-Joint.

First we will show how this can be used to provide model-level explanations and then we will explain a specific example at the song level.

7.1 Model-level Explanation

Before a model is trained, the relationship between features and response variables can be analyzed using correlation analysis. The pairwise correlations between mid-level and emotion annotations in our data are shown in Figure 4. When we compare this to the effect plots in Figure 3, or the actual weights learned for the final linear layer (Figure 5) it can be seen that for some combinations (e.g., valence and melodiousness, happy and minorness) positive correlations go along with positive effect values and negative correlations with negative effect values, respectively. This is not a general rule, however, and there are several examples (e.g., tension and dissonance, energy and melody) where it is the other way around. The explanation for this is simple: correlations only consider one feature in isolation, while learned feature weights (and thus effects) also depend on the other features and must hence be interpreted in the overall context. Therefore it is not sufficient to look at the data in order to understand what a model has learned.

	predicted		annotated	
	#153	#322	#153	#322
valence	0.28	0.39	0.38	0.46
energy	0.37	0.50	0.37	0.54
tension	0.40	0.46	0.50	0.56
anger	0.28	0.23	0.15	0.22
fear	0.41	0.27	0.18	0.28
happy	0.17	0.21	0.17	0.17
sad	0.20	0.23	0.27	0.28
tender	0.18	0.23	0.10	0.10

Table 4: Emotion prediction profiles for the two example songs #153 and #322.

To get a better understanding, we will look at each emotion separately, using the effects plot given in Figure 3. In addition to the direction of the effect – which we can also read from the learned weights in Figure 5 (but only because all of our features are positive) – we can also see the spread of the effect which tells us more about the actual contribution the feature can have on the prediction, or how different combinations of features may produce a certain prediction.

7.2 Song-level Explanations

Effect plots also permit us to create simple example-based explanations that can be understood by a human. The feature effects of single examples can be highlighted in the effects plot in order to analyze them in more detail, and in the context of all the other predictions. To show an interesting case we picked two songs with similar emotional but different midlevel profiles. To do so we computed the pairwise euclidean distances between all songs in emotion (d_{E}) and midlevel space (d_{Mid}) separately, scaled both to the range $[0, 1]$ and combined them as $d_{\mathrm{comb}} = d_{\mathrm{E}} - (1 - d_{\mathrm{Mid}})$. We then selected the two songs from the Soundtracks dataset that maximised d_{comb}. The samples are shown in Figure 3 as a red square (song #153) and a blue dot (song #322). The reader can listen to the songs/snippets by downloading them from the Soundtracks dataset page [1].

As can be seen from Figure 3 and from the emotion prediction profile of the two songs (see Table 4), both songs have relatively high predicted values for *tension* and *energy*, but apparently for different reasons: song #322 more strongly relies on “minorness" and “articulation" for achieving its “tense" character; on the other hand, its rhythmic stability counteracts this more strongly than in the case of song #153. The higher score on the “energy" emotion scale for #322 seems to be primarily due to its much more articulated character (which can clearly be heard: 153 is a saxophone playing a chromatic, harmonically complex line, 322 is an orchestra playing a strict, *staccato* passage).

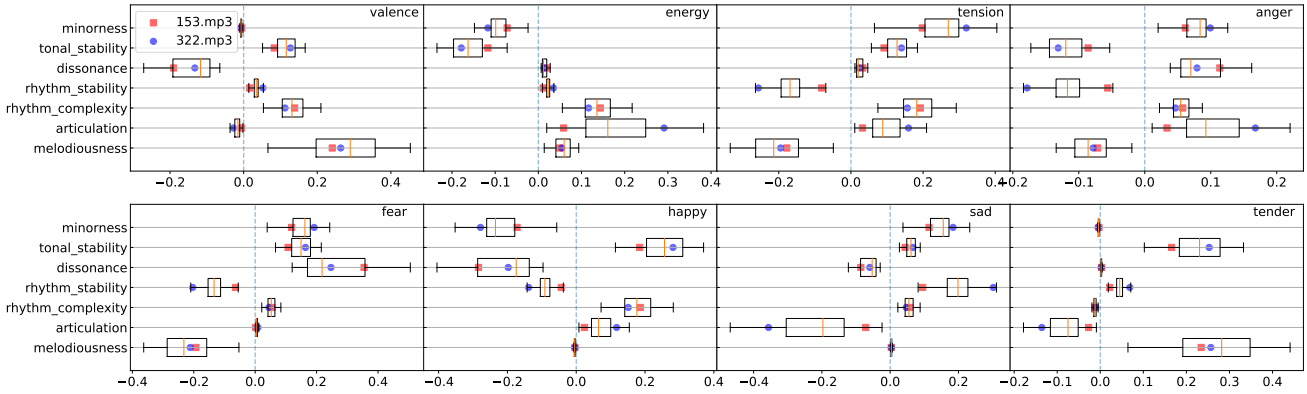

Figure 3: Effects of each mid-level feature on prediction of emotion. The boxplots show the distribution of feature effects of the model 'A2Mid2E-Joint' helping us to understand the model globally. Additionally, two example songs (blue dots, red squares) are shown to provide song-level explanations (see Section 7.2 for a discussion).

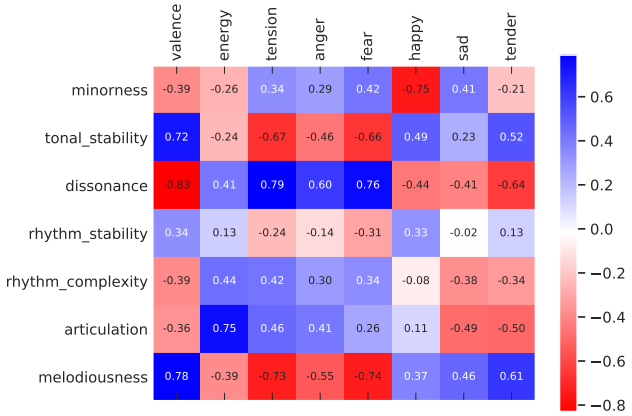

Figure 4: Pairwise correlation between mid-level and emotion annotations.

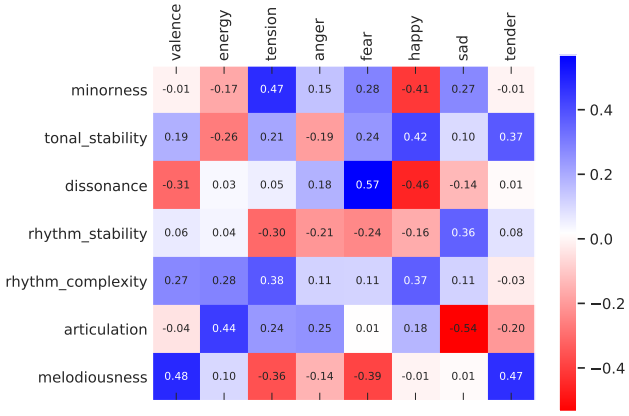

Figure 5: Weights from the linear layer of the 'A2Mid2E-Joint' model.

8. DISCUSSION AND CONCLUSION

Model interpretability and the possibility to obtain explanations for a given prediction are not ends in themselves. There are many scenarios where one may need to understand why a piece of music was recommended or placed in a certain category. Concise explanations in terms of mid-level features would be attractive, for example, in recommender systems or search engines for 'program music' for professional media producers, where mid-level qualities could also be used as additional search or preference filters [2] . As another example, think of scenarios where we want a music playlist generator to produce a music program with a certain prevalent mood, but still maintain musical variety within these limits. This could be achieved by using the mid-level features underlying the mood/emotion classifications to enforce a certain variability, by sorting or selecting the songs accordingly.

There are several obvious next steps that need to be taken in the research. The first is to extend this analysis to a larger set of diverse datasets and emotion-related dimensions. [3]

Second, we plan to extend the models and sets of perceptual features. One rather obvious (and obviously relevant) perceptual dimension that is conspicuously missing from our (Aljanaki & Soleymani's) set of mid-level features is *perceived speed* (which is not the same as tempo). Adding this intuitive musical dimension is an obvious next step towards improving our model. Of course, this will require an appropriate ground truth for training.

Generally, the relation between the space of musical qualities (such as our mid-level features) and the space of musically communicated emotions and affects deserves more detailed study. A deeper understanding of this might even give us means to control or modify emotional qualities in music by manipulating mid-level musical properties.

[2] A demonstration of mid-level explanations of emotional variability in multiple versions of songs can be found in `https://shreyanc.github.io/ismir_example.html`

[3] In fact, we do have preliminary results on a second dataset – the *MIREX-like Mood Dataset)* of [11], which is also covered by the Mid-level Perceptual Features Dataset of [1] and differs from *Soundtracks* in that is comes with discrete mood labels. The results confirm the general trends reported in the present paper, but because of the different emotion/mood encoding scheme, further optimisations on our models may still improve the results further.

9. ACKNOWLEDGMENTS

This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 670035 (project "Con Espressione").

10. REFERENCES

[1] Anna Aljanaki and Mohammad Soleymani. A data-driven approach to mid-level perceptual musical feature modeling. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, pages 615–621, 2018.

[2] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. *PLOS ONE*, 12(3):1–22, 03 2017.

[3] Eduardo Coutinho, George Trigeorgis, Stefanos Zafeiriou, and Björn W Schuller. Automatically estimating emotion in music with deep long-short term memory recurrent neural networks. In *MediaEval*, 2015.

[4] Tuomas Eerola and Jonna K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

[5] Anders Friberg, Erwin Schoonderwaldt, Anton Hedblad, Marco Fabiani, and Anders Elowsson. Using listener-based perceptual features as intermediate representations in music information retrieval. *The Journal of the Acoustical Society of America*, 136(4):1951–1963, 2014.

[6] 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.

[7] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010.

[8] Andreas Krug, René Knaebel, and Sebastian Stober. Neuron activation profiles for interpreting convolutional speech recognition models. In *NIPS 2018 Interpretability and Robustness for Audio, Speech and Language Workshop (IRASL'18)*, 2018. to appear.

[9] Chingshun Lin, Mingyu Liu, Weiwei Hsiung, and Jhihsiang Jhang. Music emotion recognition based on two-level support vector classification. In *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 375–389. IEEE, 2016.

[10] Christoph Molnar. *Interpretable Machine Learning*. 2019. `https://christophm.github.io/interpretable-ml-book/`.

[11] Renato Panda, Ricardo Malheiro, Bruno Rocha, António Oliveira, and Rui Pedro Paiva. Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis.

[12] Renato Panda, Ricardo Malheiro, Bruno Rocha, António Oliveira, and Rui Pedro Paiva. Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. In *International Symposium on Computer Music Multidisciplinary Research*, 2013.

[13] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18-21, 2018*, pages 1021–1028, 2018.

[14] Cynthia Rudin and Ustun Berk. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, 48(5):449–466, 2018.

[15] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[17] Gerhard Widmer. Getting closer to the essence of music: The Con Espressione Manifesto. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):19, 2017.

[18] 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.

[19] JiangLong Zhang, XiangLin Huang, Lifang Yang, and Liqiang Nie. Bridge the semantic gap between pop music acoustic feature and emotion: Build an interpretable model. *Neurocomputing*, 208:333 – 341, 2016. SI: BridgingSemantic.