ON PERCEIVED EMOTION IN EXPRESSIVE PIANO PERFORMANCE: FURTHER EXPERIMENTAL EVIDENCE FOR THE RELEVANCE OF MID-LEVEL PERCEPTUAL FEATURES

Shreyan Chowdhury[1]   **Gerhard Widmer**[1,2]

[1]Institute of Computational Perception, Johannes Kepler University Linz, Austria
[2]LIT AI Lab, Linz Institute of Technology, Austria

`firstname.lastname@jku.at`

ABSTRACT

Despite recent advances in audio content-based music emotion recognition, a question that remains to be explored is whether an algorithm can reliably discern emotional or expressive qualities between different performances of the same piece. In the present work, we analyze several sets of features on their effectiveness in predicting arousal and valence of six different performances (by six famous pianists) of Bach's Well-Tempered Clavier Book 1. These features include low-level acoustic features, score-based features, features extracted using a pre-trained emotion model, and Mid-level perceptual features. We compare their predictive power by evaluating them on several experiments designed to test performance-wise or piece-wise variations of emotion. We find that Mid-level features show significant contribution in performance-wise variation of both arousal and valence – even better than the pre-trained emotion model. Our findings add to the evidence of Mid-level perceptual features being an important representation of musical attributes for several tasks – specifically, in this case, for capturing the expressive aspects of music that manifest as perceived emotion of a musical performance.

1. INTRODUCTION

A musical performance, particularly in the Western music tradition, is not merely a literal acoustic rendering of a notated piece or composition. Rather, the piece is transformed by the performer's own expressive performance choices, relating to such dimensions as the choice of tempo, expressive tempo and timing variations, dynamics, articulation, and so on. The emotional effect of a performance on a listener can be a consequence both of the composition itself, with its musical properties and structures, and of the performance, the way the piece was played. In fact, it has been convincingly demonstrated [1, 2] that performers are capable of communicating, with high accuracy, intended emotional qualities by their playing.

The analysis of emotion in music recordings has a long history in Music Information Retrieval, with many works addressing content-based emotion regression and classification typically using low-level or hand-crafted audio and musical features [3–6] or using deep learning based methods [7–9]. However, there has been little research on the more subtle problem of identifying emotional aspects that are due to the actual *performance*, and even less on models that can automatically recognize this from *audio* recordings. On the latter problem – the one to be addressed in this paper – the most directly relevant prior work we are aware of is [10], where 324 6-second audio snippets of different genres (classical, jazz, blues, metal, etc.) were annotated in terms of perceived emotion (valence and arousal), and various regressors were trained to predict these two dimensions from a set of standard audio features. The regression models were then used to predict valence-arousal trajectories over 5 different recordings of 4 Chopin pieces, but no ground truth in terms of human emotion annotations was collected. The relevance of the model predictions was evaluated only indirectly, by comparing similarity scores between predicted profiles with overall performance similarity ratings by three human listeners, which showed some non-negligible correlations.

In a recent focused study [11], Battcock & Schutz (referred to as "B&S" henceforth) investigate how three specific score-based cues (Mode, Pitch Height, and Attack Rate [1]) work together to convey emotion in J.S.Bach's preludes and fugues collected in his *Well-tempered Clavier (WTC)*. They used recordings of the complete WTC Book 1 (48 pieces) of one famous pianist (Friedrich Gulda) as stimuli for human listeners to rate each performance on perceived arousal and valence. Their findings suggest that within this set of performances, arousal is significantly correlated with attack rate and valence is affected by both the attack rate and the mode. However, that study was based on only one set of performances, making it impossible to decide whether the human emotion ratings used as ground truth really reflect aspects of the compositions themselves, or whether they were also (or even predominantly) affected by the specific (and, in some cases, rather unconventional)

---

[1] Actually, attack rate as computed by B&S is also informed by the average tempo of the performance; thus, it is not strictly a score-only feature.

Pianist	Recording	Year
Glenn Gould	Sony 88725412692	1962-1965
Friedrich Gulda	MPS 0300650MSW	1972
Angela Hewitt	Hyperion 44291/4	1997-1999
Sviatoslav Richter	RCA 82876623152	1970
András Schiff	ECM 4764827	2011
Rosalyn Tureck	DG 4633052	1952-1953

Table 1: Pianists and recordings.

way in Friedrich Gulda plays the pieces – that is, whether the emotion ratings reflect piece or performance aspects.

The purpose of the present paper is to try to disentangle the possible contributions and roles of different features in capturing composer-(piece-)specific and performer-(recording-)specific aspects. To this end, we collected human ratings of perceived valence and arousal in six complete sets of recordings of WTC Book 1, and then performed a systematic study with feature sets derived from various levels of musical abstraction, including some extracted by pre-trained deep neural networks.

2. DATA COLLECTION

2.1 Pieces and Recordings

J.S.Bach's *Well-tempered Clavier (WTC)* is ideally suited for systematic and controlled studies of this kind, as it comprises a stylistically coherent set of keyboard pieces from a particular period, evenly distributed over all keys and major/minor modes, with a pair of two pieces (a prelude, followed by a fugue) in each of the 24 possible keys, for a total of 48 pieces. Each piece has its own distinctive musical character, and despite being written in a rather strict style and not meant to be played in 'romantic' ways, the music offers pianists (or pianists take) lots of liberties in ornamentation, but also overall performance parameters (e.g., tempo and articulation). For example, there are pieces in our set of recordings that one pianist takes more than twice (!) as fast as another.

For a broad set of diverse performances, we selected six recordings of the complete WTC Book 1, by six famous and highly respected pianists, all of whom can be considered Bach specialists to various degrees. The recordings are listed in Table 1.

2.2 Emotion Annotations and Pre-processing

In accordance with B&S, we will only use the first 8 bars of each recording for the annotation process and our experiments. These were cut out manually. The participants of our annotation exercise were students of a course at a university, without a specifically musical background. Each participant heard a subset of the recordings (all 48 pieces as played by one pianist) and was asked to rate the valence on a scale of -5 to +5 (11 levels) and the arousal on a scale of 0 to 100 (increments of 10; a total of 11 levels). They could listen to a recording as many times as they liked. Each recording was rated by 29 participants. In total, we collected 8,352 valence-arousal annotation pairs.

For the purposes of this paper, we take the mean arousal and mean valence ratings for each recording, and these values serve as our ground-truth values for all following experiments. The distributions (over the 6 performances) of these mean ratings for each piece are summarised in Figure 1.
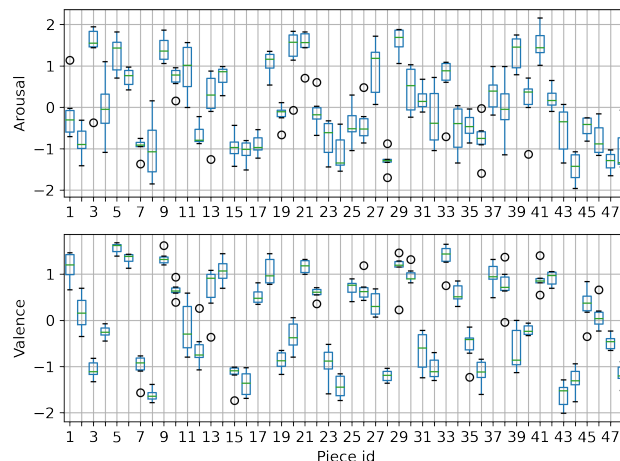

Figure 1: Distribution of emotion ratings across pianists for every piece.

3. FEATURE SETS

In this section, we briefly describe the four feature sets that we use to model arousal and valence.

3.1 Low-level Features

These consist of hand-crafted musical features (such as onset rate, tempo, pitch salience) as well as generic audio descriptors (such as spectral centroid, loudness). Taken together, they reflect several musical characteristics such as tone colour, dynamics, and rhythm. A brief description of all low-level features that we use is given in Table 2. We use Essentia [12] and Librosa [13] for extracting these. The audio is sampled at 44.1kHz and the spectra computed (when required) with a frame size of 1024 samples and a hop size of 512 samples. Each feature is aggregated over the entire duration of an audio clip by computing the mean and standard deviation over all the frames of the clip (a 'clip' being an 8 bar initial segment from a recording).

3.2 Score Features

The following set of features was computed directly from the musical score (i.e., sheet music) of the pieces instead of the audio files. The unit of score time, "beat", is defined by the time signature of the piece (e.g., 4/4 means that there are 4 beats of duration 1 quarter in a bar). The score information and the audio files were linked using automatic score-to-performance alignment. Table 3 describes the score features in detail.

Dissonance	Total harmonic dissonance computed from pairwise dissonance of all spectral peaks.
Dynamic Complexity	The average absolute deviation from the global loudness level estimate in dB.
Loudness	Mean loudness of the signal computed from the signal amplitude.
Onset Rate	Number of onsets (note beginnings or transients) per second.
Pitch Salience	A measure of tone sensation, computed from the harmonic content of the signal.
Spectral Centroid	The weighted mean frequency in the signal, with frequency magnitudes as the weights.
Spectral Flatness	A measure to quantify how much noise-like a sound is, as opposed to being tone-like.
Spectral Bandwidth	The second order bandwidth of the spectrum.
Spectral Rolloff	The frequency under which 85% of the total energy of the spectrum is contained.
Spectral Complexity	The number of peaks in the input spectrum.
Tempo (BPM)	Tempo estimate from audio in beats per minute.

Table 2: Low-level Features

Inter Onset Interval	The time interval between consecutive notes per beat.
Duration	Two features describing the empirical mean and standard deviation of the notated duration per beat in the snippet.
Onset Density	The number of note onsets per beat. A chord constitutes a single onset.
Pitch Density	The number of unique notes per beat.
Mode	Binary feature denoting major/minor modality, computed using the Krumhansl-Schmuckler key finding algorithm [14] (to reflect the fact that the dominant key over the segment may be different from the given key signature).
Key Strength	This feature represents how much does the tonality by the "Mode" feature fit the snippet.

Table 3: Score Features

3.3 Mid-level Features

Mid-level features, described in [15], are perceptual musical features that are intuitively understandable to the average listener. They seem well-suited to bridge the "semantic gap" between low-level audio features and high-level descriptors such as emotion and have been shown to be useful in explainable music emotion recognition [16]. We learn these features from the Mid-level Dataset [15] using a receptive-field regularised residual neural network (RF-ResNet) model [17]. Since we intend to use this model to extract features from solo piano recordings (a genre that is not covered by the original training data), we use a domain-adaptive training approach as described in [18]. We use an input audio length of 30 seconds, padded or cropped as required. As these features cannot be strictly defined, Table 4 lists a rough description adapted from the questions in [15] that were shown to the annotators of the dataset to help them rate the audio clips.

3.4 DEAMResNet Emotion Features

To compare the mid-level features with another deep neural network based feature extractor, we train a model with

Melodiousness	How singable is this music?
Articulation	Overall impression of articulation in terms of staccato or legato playing style. Higher means more staccato.
Rhythmic Stability	How easy is it to march-along with the music?
Rhythmic Complexity	How difficult is it to follow the music by tapping? Rhythmic layers and different meters correlate with higher complexity.
Dissonance	Noisier timbre or presence of dissonant intervals (tritones, seconds, etc.)
Tonal Stability	How clear or apparent the tonic and key are.
Minorness	Relates to the perceived tonality. More "minor-sounding" music will have higher minorness.

Table 4: Mid-level Features

the same architecture (RF-ResNet) and training strategy on the DEAM dataset [19] to predict arousal and valence from spectrogram inputs. Since this model is trained to predict arousal and valence, it is expected to learn representations suitable for this task. As with the mid-level model, we perform domain adaptation while training this model also.

Features are extracted from the penultimate layer of the model, which gives us 512 features. Since these are too many features to use for our dataset containing only 288 data points, we perform dimensionality reduction using PCA (Principal Component Analysis), to obtain 9 components explaining at least 98% of the variance. These 9 features are named as `pca_x` with x being the principal component number.

4. FEATURE EVALUATION EXPERIMENTS

In this section, we evaluate the four feature sets. The aim of this section is to answer the following questions:

1. How well can each feature set fit the arousal and valence ratings? How do these feature sets compare to the ones used by B&S? (Sections 4.1 and 4.2)

2. In each feature set, which features are the most important? (Section 4.3)

3. Which feature set best explains variation of arousal and valence *between pieces*? (Section 4.4)

4. Which feature set best explains variation of arousal and valence *between different performances of the same piece*? (Section 4.5)

We use ordinary least squares fitting on the dataset in question and calculate the regression metrics. The metrics we report are adjusted coefficient of determination ($\tilde{R}^2$), root mean squared error between true and predicted values (RMSE), and Pearson's correlation coefficient between true and predicted values (Corr).

4.1 Evaluation on B&S Data

As a starting point, we take the data used by B&S in Experiment 3 of their paper – Gulda's performances rated on valence and arousal We perform regression with our feature sets and compare with the values obtained by B&S

using their features Attack Rate, Pitch Height, and Mode. The results are summarised in Table 5.

	Arousal			Valence		
	$\tilde{R}^2$	RMSE	Corr	$\tilde{R}^2$	RMSE	Corr
Mid-level	0.84	0.36	0.93	**0.79**	0.42	0.91
DEAMResNet	**0.91**	0.27	0.96	0.69	0.50	0.86
Low-level	0.86	0.29	0.96	0.67	0.45	0.89
Score	0.31	0.74	0.67	0.61	0.55	0.83
B&S (exp 3)	0.48	-	-	0.75	-	-

Table 5: Regression on Gulda data from B&S [11].

We can see that all three audio-based features perform considerably well for both arousal and valence to motivate further analysis.

4.2 Evaluation on Our Dataset

Next, we perform regression on our complete dataset (comprising of 288 unique recordings – 48 pieces × 6 pianists). The results summary can be seen in Table 6a. Here again, we observe that while DEAMResNet Emotion features perform best on arousal and Score features perform best on valence, Mid-level features show a balanced performance across both the emotion dimensions.

To evaluate generalizability, we perform cross-validation with three different kinds of splits – piece-wise (all 6 performances of a piece are test samples in a fold, for a total of 48 folds), pianist-wise (all 48 pieces of a pianist are test samples in a fold, for a total of 6 folds), and leave-one-out (one recording is the test sample per fold, for a total of 288 folds). This is summarized in Table 6b.

Feature Set	Arousal			Valence		
	$\tilde{R}^2$	RMSE	Corr	$\tilde{R}^2$	RMSE	Corr
Mid-level	0.68	0.56	0.83	0.63	0.60	0.80
DEAMResNet	**0.70**	0.54	0.84	0.42	0.72	0.69
Low-level	0.62	0.59	0.81	0.41	0.74	0.67
Score	0.41	0.75	0.65	**0.75**	0.49	0.87

(a) Regression metrics with our data

Feature Set	Piece-wise		Pianist-wise		LOO	
	A	V	A	V	A	V
Mid-level	**0.68**	0.63	**0.68**	0.64	**0.69**	0.65
DEAMResNet	0.67	0.37	0.61	0.41	0.68	0.43
Low-level	0.54	0.20	−0.11	−0.05	0.57	0.30
Score	0.08	**0.67**	0.39	**0.75**	0.37	**0.74**

(b) $\tilde{R}^2$ for different cross-validation splits. A: Arousal, V: Valence, LOO: Leave-One-Out

Table 6: Evaluation results on our data

We see that Mid-level features show good generalization for arousal and are robust to different kinds of splits. They also show balanced performance between arousal and valence for all splits. The good performance of the Score features on the valence dimension (V), here and in the previous experiment, is mostly due to the *Mode* feature; there is a substantial correlation in the annotations between major/minor mode and positive/negative valence.

4.3 Feature Importance within Feature Sets

We use the absolute value of the t-statistic of a feature as the importance measure. T-statistic is defined as the estimated weight scaled with its standard error. We focus on the audio-based feature sets here, as in most realistic applications scenarios, the score information will not be available (and, being constant across different performances, will not be able to distinguish performance aspects). We perform a regression using all audio-based features (numbering 39 in total) and compare the t-values in Figure 2.

We see that the top-4 and top-2 features in arousal and valence, respectively, are Mid-level features. These features also make obvious musical sense – modality is often correlated with valence (positive or negative emotional quality), and tempo, rhythm, and articulation with arousal (intensity or energy of the emotion).

4.4 Explaining Piece-wise Variation

We observe from the annotation data (see Figure 1) that the distribution of emotion ratings of each piece is distinct (here, the 6 performances for each piece form the "distribution" of the piece). In essence, the mean value of arousal or valence depends on the piece in question. Therefore, to take into account this factor of variation, we use linear mixed models [20] to model arousal and valence.

In this linear mixed effect model, the piece id is considered as a "random effect" intercept, which models part of the residual remaining unexplained by the features we are evaluating ("fixed effects"). A feature set that models piece-wise variation better than another set would naturally have a lesser residual variation to be explained by the random effect. We therefore look at which feature set has the least fraction of residual variance explained by the random effect of piece id, defined as:

$$E_{\text{random}} = \frac{\text{Var}_{\text{random}}}{\text{Var}_{\text{random}} + \text{Var}_{\text{residual}}} \quad (1)$$

where $\text{Var}_{\text{random}}$ is the variance of the random effect intercept and $\text{Var}_{\text{residual}}$ is the variance of the residual that remains after mixed effects modeling.

Feature Set	Arousal	Valence
Mid-level	0.50	0.86
DEAMResNet	**0.47**	0.89
Low-level	0.66	0.90
Score	0.63	**0.68**

Table 7: Fraction of residual variance explained by the random effect of "piece id".

We see from Table 7 that the DEAMResNet emotion features best explain piece-wise variation in arousal, followed closely by Mid-level features. For valence, the performance of all three audio-based features are close, with Mid-level features performing the best, however, score features outperform them with a large margin. This is again due to the relationship between mode and valence, and mode covarying tightly with the piece ids.
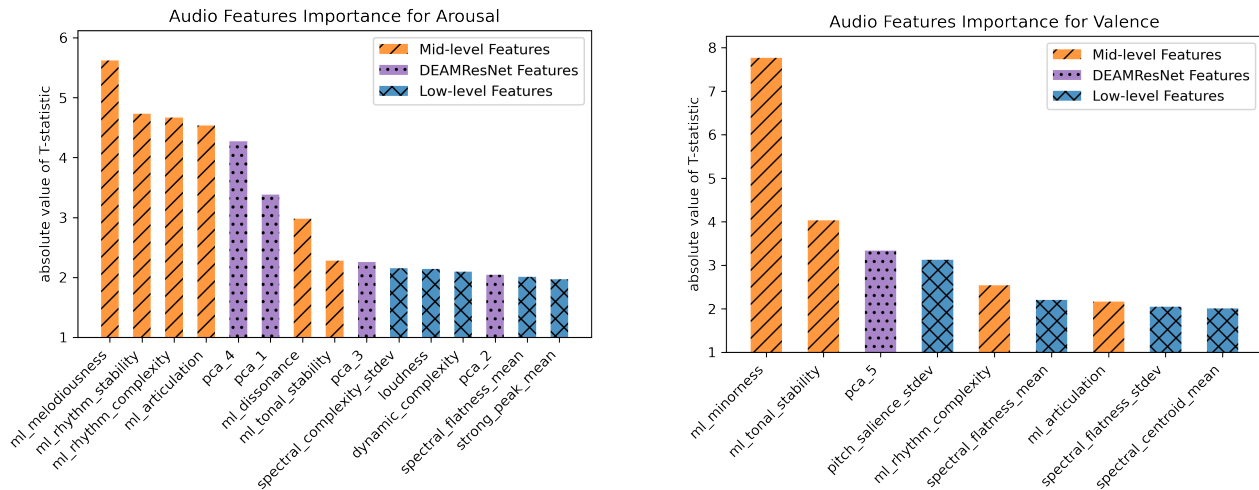
Figure 2: Feature importance for audio features using T-statistic. Only features with p<0.05 are shown.

4.5 Explaining Performance-wise Variation

Evaluation of performance-wise variation modelling cannot be done with the mixed effects approach as in the previous section because the means (of arousal or valence across all pieces) are nearly identical for each pianist.

Therefore, we look at one piece at a time and compute the fraction of variance unexplained (FVU) and Pearsons's correlation coefficient (Corr) between predicted and true values across performances for each such test piece. This is done as leave-one-piece-out cross-validation, and aggregated by taking the means. The p-values of the correlation coefficients are counted and we report the percentage of pieces for which $p < 0.1$. With only 6 performances per piece, a significance level of $p < 0.05$ is obtained for only a handful of pieces. Since score-features based predictions are exactly equal for all performances of a piece, these metrics are not meaningful, and hence the Score feature set is not included here.

Feature Set	Arousal		Valence	
	FVU	Corr (p<0.1)	FVU	Corr (p<0.1)
Mid-level	**0.31**	**0.58** (47.9%)	**0.36**	0.42 (27.0%)
DEAMResNet	0.32	0.54 (43.8%)	0.61	**0.47** (37.5%)
Low-level	0.43	0.56 (54.2%)	0.75	0.38 (22.9%)

Table 8: Evaluation metrics for performance-wise variation. FVU: Fraction of Variance Unexplained. Corr: Pearson's correlation coefficient.

Again, Mid-level features come out at the top in most measures. To illustrate the modelling of performance-wise variation, we select a few example pieces that have a high variation of emotion between performances and plot them together with the predicted values using mid-level features in Figure 3. The predicted emotion dimensions follow the ratings closely, even for performances that deviate from the average (e.g. the arousals of Gulda's performance of Prelude in A major and Tureck's performance of Fugue in E minor.)
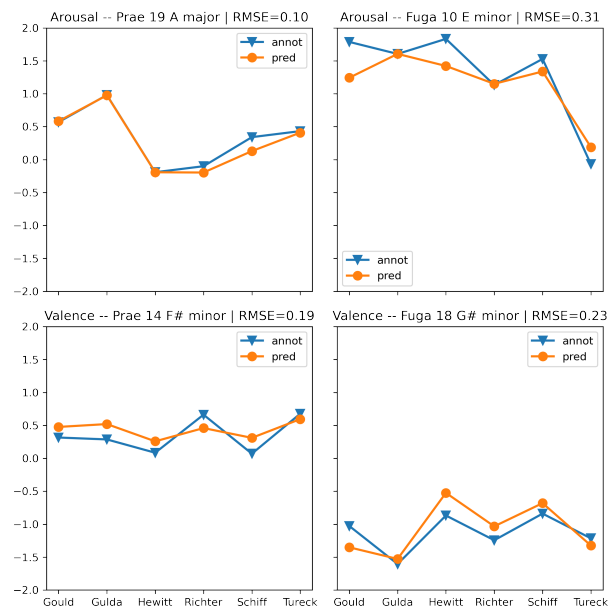

Figure 3: Some example pieces with high emotion variability between performances which are modeled particularly well using mid-level features.

5. PROBING FURTHER

We now describe two additional experiments designed to further probe the predictive power of our feature sets.

5.1 Predicting Emotion of Outlier Performances

Figure 4 shows two examples of pieces where one performance has a vastly different emotional character than the others – in the first example, Gould even produces a negative valence effect (mostly through tempo and articulation) in the E-flat major prelude, which the others play in a much more flowing fashion. A challenge for any model would thus be to predict the emotion of such idiosyncratic performances, not having seen them during training.

We therefore create a test set by picking out the outlier performance for each piece in arousal-valence space
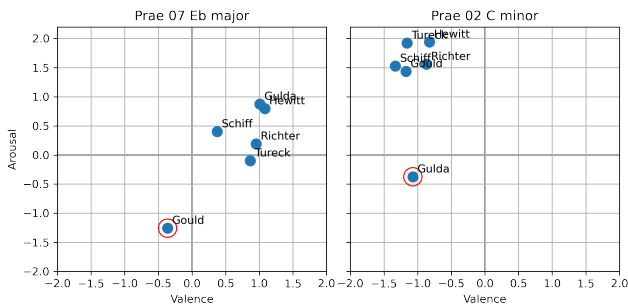
Figure 4: Two examples of outlier performances: Prelude # 7 in Db major, outlier is Gould (left); Prelude #2 in C minor (Gulda, right).

using the elliptic envelope method [21]. This gives us a split of 240 training and 48 test samples (the outliers). We train a linear regression model using each of our feature sets and report the performance on the outlier test set in Figure 5. We see again that Mid-level features outperform the others, for both emotion dimensions. We take this as another piece of evidence for the ability of the mid-level features to capture performance-specific aspects. The surprisingly good performance of score features for valence can be attributed to the fact that for most pieces, the outlier points are separated mostly in the arousal dimension – the spread of valence is rather small (though not always: see the Gould case in Fig. 4) – and the score feature "mode" is an important predictor of valence (see earlier sections).
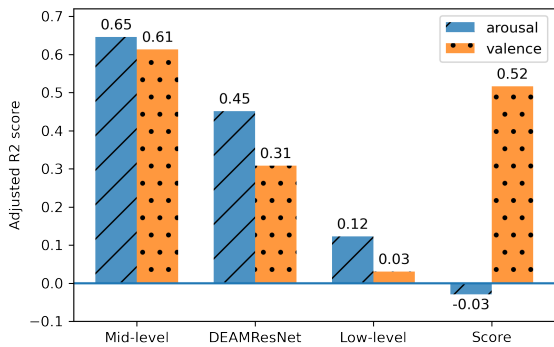

Figure 5: $\tilde{R}^2$ scores on outlier performances. The outliers were selected using elliptic envelope on the rated arousal-valence space. Out of 48 pieces, the number of times each pianist was an outlier are Gould: 13, Gulda: 10, Tureck: 10, Schiff: 5, Hewitt: 5, Richter: 5.

5.2 Predicting Discrete Emotions

Finally, we evaluate how the feature sets perform in a discrete emotion classification task, which might be relevant in a music recommendation setting, for instance. The emotion ratings are converted to classes simply by reducing them to quadrants in the arousal-valence space. In the literature, these are often associated with the basic emotion labels *happy*, *relaxed*, *sad*, and *angry* (in clockwise fashion, starting at upper right). We then train logistic regres-

sion models using our feature sets and report the leave-one-out cross-validation accuracy in Figure 6. We observe that in this task, all feature sets perform more-or-less equally well, again with a slight advantage for the Mid-level features. Note that the random choice baseline accuracy is 0.25. An emotion classification model based on mid-level perceptual features might be attractive for performance-emotion-aware music recommendation, being able to offer the mid-level concepts not only as explanations but also as 'handles' to search for performances with certain characteristics.
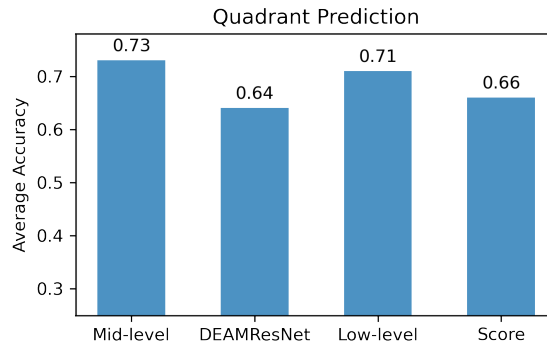

Figure 6: Discrete emotion classification.

6. CONCLUSION

In this work, we evaluated four feature sets – mid-level perceptual features, pre-trained emotion features, low-level audio features, and score-based features on their ability to model and predict emotion in terms of arousal and valence. Specific focus was given on the three audio-based features and their modelling power over performance-wise variation of emotion. Mid-level features emerge as the most robust and important among these.

The search for good features to model music emotion is a worthwhile objective since emotional effect is a very fundamental human response to music. Features that provide a better handle on content-based emotion recognition can have a significant impact on applications such as search and recommendation. Modelling emotion is also becoming increasingly relevant in generative music, allowing possibilities such as expressivity- or emotion-based snippet continuation and emotion-aware human-computer collaborative music.

From the experiments presented in this paper, it is clear that deep-learning-based feature extractors are strong competition to the audio features typically used for emotion recognition [3]. Here the importance of Mid-level features is even more pronounced – in addition to being able to model both arousal and valence well under different conditions, they also provide intuitive musical meaning to each feature, and have been previously used as the basis for explainable emotion recognition in [16].

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] A. Gabrielsson and P. N. Juslin, "Emotional expression in music performance: Between the performer's intention and the listener's experience," *Psychology of Music*, vol. 24, no. 1, pp. 68–91, 1996.

[2] J. Akkermans, R. Schapiro, D. Müllensiefen, K. Jakubowski, D. Shanahan, D. Baker, V. Busch, K. Lothwesen, P. Elvers, T. Fischinger *et al.*, "Decoding emotions in expressive music performances: A multi-lab replication and extension study," *Cognition and Emotion*, vol. 33, no. 6, pp. 1099–1118, 2019.

[3] R. Panda, R. M. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: a survey," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[4] M. Soleymani, A. Aljanaki, Y.-H. Yang, M. N. Caro, F. Eyben, K. Markov, B. W. Schuller, R. Veltkamp, F. Weninger, and F. Wiering, "Emotional analysis of music: A comparison of methods," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1161–1164.

[5] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010*, vol. 86, 2010, pp. 937–952.

[6] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 1–30, 2012.

[7] R. Orjesek, R. Jarina, M. Chmulik, and M. Kuba, "DNN based music emotion recognition from raw audio signal," in *29th International Conference Radioelektronika 2019 (RADIOELEKTRONIKA)*. IEEE, 2019, pp. 1–4.

[8] M. B. Er and I. B. Aydilek, "Music emotion recognition by using chroma spectrogram and deep visual features," *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1622–1634, 2019.

[9] N. HE and S. Ferguson, "Multi-view neural networks for raw audio-based music emotion recognition," in *2020 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2020, pp. 168–172.

[10] J. Grekow, "Musical performance analysis in terms of emotions it evokes," *Journal of Intelligent Information Systems*, vol. 51, no. 2, pp. 415–437, 2018.

[11] A. Battcock and M. Schutz, "Acoustically expressing affect," *Music Perception: An Interdisciplinary Journal*, vol. 37, no. 1, pp. 66–91, 2019.

[12] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, H. Boyer, O. Mayor *et al.*, "Essentia: An audio analysis library for music information retrieval." International Society for Music Information Retrieval (ISMIR), 2013, pp. 493–498.

[13] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[14] C. L. Krumhansl, *Cognitive foundations of musical pitch*. Oxford University Press, 2001.

[15] A. Aljanaki and M. 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*, 2018, pp. 615–621.

[16] S. Chowdhury, A. Vall, V. Haunschmid, and G. Widmer, "Towards Explainable Music Emotion Recognition: The Route via Mid-level Features," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, 2019.

[17] K. Koutini, H. Eghbal-Zadeh, M. Dorfer, and G. Widmer, "The Receptive Field as a Regularizer in Deep Convolutional Neural Networks for Acoustic Scene Classification," in *2019 27th European signal processing conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[18] S. Chowdhury and G. Widmer, "Towards explaining expressive qualities in piano recordings: Transfer of explanatory features via acoustic domain adaptation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 561–565.

[19] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a Benchmark for Emotional Analysis of Music," *PloS one*, vol. 12, no. 3, p. e0173392, 2017.

[20] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team, *nlme: Linear and Nonlinear Mixed Effects Models*, 2021, r package version 3.1-152. [Online]. Available: https://CRAN.R-project.org/package=nlme

[21] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.