ON THE CHARACTERIZATION OF EXPRESSIVE PERFORMANCE IN CLASSICAL MUSIC: FIRST RESULTS OF THE *CON ESPRESSIONE* GAME

Carlos Cancino-Chacón[1,4] **Silvan Peter**[2] **Shreyan Chowdhury**[2]
Anna Aljanaki[3] **Gerhard Widmer**[1,2]
[1] Austrian Research Institute for Artificial Intelligence, Vienna, Austria
[2] Institute of Computational Perception, Johannes Kepler University Linz, Austria
[3] Institute of Computer Science, University of Tartu, Estonia
[4] RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, University of Oslo, Norway
carlos.cancino@ofai.at, silvan.peter@jku.at

ABSTRACT

A piece of music can be expressively performed, or interpreted, in a variety of ways. With the help of an online questionnaire, the *Con Espressione* Game, we collected some 1,500 descriptions of expressive character relating to 45 performances of 9 excerpts from classical piano pieces, played by different famous pianists. More specifically, listeners were asked to describe, using freely chosen words (preferably: adjectives), how they perceive the expressive character of the different performances. In this paper, we offer a first account of this new data resource for expressive performance research, and provide an exploratory analysis, addressing three main questions: (1) how similarly do different listeners describe a performance of a piece? (2) what are the main dimensions (or axes) for expressive character emerging from this?; and (3) how do measurable parameters of a performance (e.g., tempo, dynamics) and mid- and high-level features that can be predicted by machine learning models (e.g., articulation, arousal) relate to these expressive dimensions? The dataset that we publish along with this paper was enriched by adding hand-corrected score-to-performance alignments, as well as descriptive audio features such as tempo and dynamics curves.

1. INTRODUCTION

In the Western classical music tradition, music exists at an interplay of creative intentions of composer, performers and listeners. Composers encode their ideas using written notation (i.e., musical scores), and performers bring these ideas to life, guided by the expression markings, performance traditions, and their own creative imagination. Each performance can sound very different from the next.

Much of the research on musical expression has focused on what pieces express through attributes of their musical structure [1–3]. There has been much focus on the expression of *emotion* in particular (e.g., [4–8]); however, a comprehensive description of the expressive character of a performance includes additional, not specifically emotion-related aspects. The aim of this research is to find the dimensions of musical expression that can be attributed to a performance, as perceived and described in natural language by listeners.

Within the classical music tradition, there is already a practice of assigning verbal descriptors to aspects of musical expression. For example, instructions related to expressive character are sometimes marked on the score by the composer with a fixed set of (mostly) Italian terms (e.g., *Allegro*, *dolce*). Many of those terms describe emotions, but they describe a wide range of other aspects as well, including movement analogies and metaphors (e.g., free, flowing) [9].

Using an online questionnaire, the *Con Espressione* Game (CEG), which is part of the research project of the same name [10], we collected verbal descriptors of expressive performances. In this paper we present first results on three main questions: (1) can we observe any consistency in the way listeners perceive and describe *expressive character*?; (2) can we identify main descriptive dimensions along which these characterizations can be organized?; and (3) how are these dimensions related to measurable qualities (or parameters) from audio recordings, or to performance information extracted from these?

The rest of this paper is structured as follows: Section 2 points to some related work. Section 3 describes the CEG and the collected data. We continue the paper with individual sections addressing the three main questions we want to study: Section 4 presents results on how similar/consistent the characterizations of the participants are, Section 5 focuses on analyzing the main descriptive dimensions. Section 6 presents results relating performance and audio features to the expressive dimensions. Section 7 discusses the results of these experiments. Finally, Section 8 concludes the paper.

2. RELATED WORK

In this section we provide a few pointers to related work on research on music emotion and expressive performance, and on characterizing musical expression. A full overview is beyond the scope of this paper.

From a musicological and philosophical viewpoint, the study of musical expression has focused on music as in composition [1,2], although there is a trend in recent times to include and recognize both the role of the performer and the role of the listener. Some recent papers have focused on developing mid-level features that capture qualities of musical recordings that relate to perceptual aspects of the music [11, 12] . For an overview of literature on emotion in music, we refer the reader to [3] and [13].

Computational models of expressive performance are a means to study principles of performance in quantitative terms. In Western classical music, much of this work has focused on establishing relationships between structural aspects of a musical score and quantifiable aspects of a performance such as expressive timing and dynamics. At the intersection of research on music emotion and computational models of performance is the study of the relation between performance parameters (timing, dynamics) and emotion [4–8]. For an overview of computational modeling of music performance, we refer the reader to [14].

Particularly relevant to the current study are adjective lists for describing musical expressiveness, including the seminal work by Hevner [15] and its more recent updates [16, 17]. Hevner identified several clusters of adjectives that describe musical expression. Many of these clusters relate to emotions, and include aspects such as *'happy'* and *'sad'*. Schaerlaeken et al. [18] conducted a large study to investigate the use of metaphors for describing Western classical music. They propose the Geneva Musical Metaphors Scale (GEMMES), which comprises 5 metaphorical scales including aspects like *flow, movement, force, interior* and *wandering*. The focus of their study was on characterization of different pieces of music, rather than on description of the character of different expressive performances. A related study [9] presents a perception-based clustering of expressive musical terms (i.e., performance directives such as *catabile* and *leggiero*), relating these terms to locations in Russell's arousal–valence space [19]. Murari et al. [20] study of listeners' characterization of music using non-verbal sensory scales.

3. THE *CON ESPRESSIONE* GAME

We collected our data through a Web-based questionnaire ("CEG"), where participants listened to several performances of each of 9 classical piano pieces. They were then asked to describe the performance in free text (preferably adjectives, as many as they liked), concentrating on the performative aspects and not on the piece itself. The target phenomenon thus is what we would call the *expressive character* of performances. Users could also select their favorite performance of each piece. Upfront, a few general questions were asked regarding their level of musical education, how often they listen to classical music, and whether they play the piano. The order of pieces and performances was randomized. Users could stop at any time.

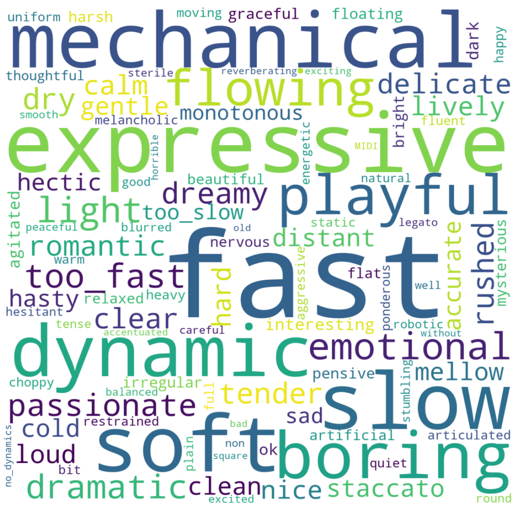

Figure 1. A word cloud of the terms in the dataset.

3.1 Pieces and Responses

The CEG contains 45 performances of 9 excerpts (see Table 2).[1] The length of the excerpts was between 27 seconds and 188 seconds. The CEG was launched on the 3rd of April 2018. The questionnaire was filled out by 194 participants, out of which 88% had some kind of music education – on average 11.7 years; 179 participants answered in English, 12 in German and three in each of Russian, Spanish and Italian. On average, participants listened to the performances of 4.5 out of 9 pieces, 27 participants listened to all the 45 excerpts.

3.2 Meta-data and Annotations

We enriched the dataset by adding score-to-performance alignments for all performances in the dataset, as well as descriptive performance features. The alignments were produced by manually annotating the position of the beats in the audio files using Sonic Visualiser [21]. The scores were encoded manually and exported in MusicXML format following publicly available editions.[2] From these alignments we compute performance features such as tempo and loudness curves (see Section 6.1). The dataset is available online.[3]

4. HOW SIMILARLY DO LISTENERS DESCRIBE A PERFORMANCE OF A PIECE?

Probably the first question that arises concerns the similarity of the descriptions in the dataset, i.e., whether there are

[1] `con-espressione.cp.jku.at`

[2] from `https://imslp.org`.

[3] The dataset and code necessary to reproduce the experiments reported in this paper can be found in `https://cpjku.github.io/con_espressione_game_ismir2020/`.

Composer	Piece	#	Pianists
Bach	Prelude No.1 in C, BWV 846 (WTC I)	7	Gieseking, Gould, Grimaud, Kempff, Richter, Stadtfeld, MIDI
Mozart	Piano Sonata K.545 C major, 2nd mvt.	5	Gould, Gulda, Pires, Uchida, MIDI
Beethoven	Piano Sonata Op.27 No.2 C# minor, 1st mvt.	6	Casadesus, Lazić, Lim, Gulda, Schiff, Schirmer
Schumann	Arabeske Op.18 C major (excerpt 1)	4	Rubinstein, Schiff, Vorraber, Horowitz
Schumann	Arabeske Op.18 C major (excerpt 2)	4	Rubinstein, Schiff, Vorraber, Horowitz
Schumann	Kreisleriana Op.16; 3. Sehr aufgeregt (ex. 1)	5	Argerich, Brendel, Horowitz, Vogt, Vorraber
Schumann	Kreisleriana Op.16; 3. Sehr aufgeregt (ex. 2)	5	Argerich, Brendel, Horowitz, Vogt, Vorraber
Liszt	Bagatelle sans tonalité, S.216a	4	Bavouzet, Brendel, Katsaris, Gardon
Brahms	4 Klavierstücke Op.119, 2. Intermezzo E minor	5	Angelich, Ax, Serkin, Kempff, Vogt

Table 1. Performances used in the *Con Espressione* Game.

commonalities in the way listeners describe and like performances. In this section we present a series of analyses that provide different perspectives on the data.

4.1 Frequency and Distribution of Terms

Users provided 1,515 individual descriptions for a total of 3,166 terms, of which 1,415 (approx. 45%) are unique. Figure 1 shows a word cloud of the terms appearing in the dataset. Taken all answers together, each performance was described using at least 47 and at most 114 terms. The average number of terms per piece varies between 60.3 for Liszt's *Bagatelle sans tonalité* and 98.4 for Bach's *Prelude in C*. Each performance is characterized by at least 44 and at most 91 unique terms. The average number of unique terms per piece varies between 51.0 for Liszt's Bagatelle and 78.4 for Bach's Prelude.

The terms *'dynamic'*, *'expressive', 'fast', 'like', 'loud', 'mechanical', 'slow'*, and *'soft'* appear in at least one performance of every piece. 420 terms appear more than once and, coincidentally, also for different performances. Only 40 terms appear in more than ten performances. The most frequently used terms *across performances* are *'dynamic', 'expressive'*, and *'soft'*, which appear in at least 22 performances. In increasing order of occurrence, the terms *'playful', 'boring', 'dynamic', 'mechanical', 'slow', 'soft', 'expressive'*, and *'fast'* are used at least 40 times, with this last term being the most used with 64 occurrences.

4.2 Complexity of the Descriptions

An interesting question is whether there is a relation between listeners' musical backgrounds and the complexity of their answers. In particular, we are interested in determining whether listeners with more musical experience with Western classical music describe performances of this kind of music in a more complex language. As a measure of the complexity of the descriptions we use the Dale-Chall readability score [22], a measure that takes into account the number and commonality of words (defined as words that would be familiar to American students) in a weighted sum. Intuitively, we can understand this measure as a way of combining the length (i.e., number of terms) of the answer and the use of specialized vocabulary (such as musical terms). A larger score means a more complex answer. To test whether listeners with more musical training describe performances in a more complex language, we conduct a linear regression (years of musical training vs. answer complexity). This analysis reveals a positive correlation that we assessed for statistical significance using a one-tailed Wald test (Pearson's $r = 0.27$, $W = 2.32$, $p = 0.01$, $R^2 = 0.07$), which suggests a small effect. We also test whether listeners that often *listen* to classical music describe performances in more complex terms, and find a small, non-significant correlation. These results present weak evidence supporting the idea that listeners with more musical experience describe the expressive character in more complex ways.

4.3 Listeners' Preferences

To determine how similar the preferences of the listeners are, i.e., if they like (or dislike) similar performances we compute a χ^2-test for each piece to determine if there are clear preferred performances, or if all performances of the same piece are equally liked. There are only three pieces that reject the null hypothesis (the frequency of preferred performances is flat) at a level $\alpha = 0.01$, namely the performance of Bach's piece by Richter; and the performances of the two excerpts of Schumann's Arabeske (in this case, the preferred performances were both by Vorraber). Two additional pieces reject the null hypothesis at level $\alpha = 0.05$: Schumann's Kreisleriana performed by Brendel and Mozart's piece by Pires. Both Gould's performances and the deadpan MIDI performances (for both Bach and Mozart) were the least preferred for this piece. It is important to emphasize that this analysis only indicates the presence of preferred performances; it does not identify what the preferred performance is.

4.4 Semantic Similarity of the Descriptions

We use Li et al.'s [23] method for estimating the semantic similarity of short sentences, to compute the pairwise similarity between descriptions of performances. Similarity between individual terms is estimated as the semantic distance between words in WordNet [24], and the overall sentence similarity is weighted with corpus statistics. Intuitively, this method quantifies the overlap in terms (and very directly related synonyms of these terms) between the answers of the participants. The average similarity of the descriptions of a performance of a piece by the same pianist (i.e., *intra-performance*) is 16%, the average similarity of descriptions of a performance and the descriptions of other performances of the same piece by different pianists (i.e., *inter-performance*) is 15%; and the average similarity of descriptions of a performance and performances of other pieces (i.e., *inter-piece*, excluding performances of the same piece by other pianists) is 15%. To test whether the differences between these groups are significant, we performed a one-way ANOVA ($F(132, 2) = 9.8$, $p < 0.001$, $\eta^2 = 0.13$), which suggests a medium-small effect. We use t-tests with Bonferroni correction to test the pairwise differences (3 tests, $\alpha = .02$). These tests suggest that the average intra-performance similarity is larger than both the average inter-performance (one-tailed $t(88) = 3.4$, $p < 0.001$, Cohen's $d = .76$) and inter-piece similarity (one-tailed $t(88) = 3.6$, $p < 0.001$, Cohen's $d = .71$). The difference between the inter-performance and inter-piece similarities is not statistically significant (two-tailed $t(88) = 0.0$, $p = 0.99$, Cohen's $d = 0.0$). These results suggest that listeners describe a performance of a piece in more similar terms than they describe other performances of the same piece by different pianists. A possible explanation for the fact that there is little variation in how listeners describe performances of different pieces could be that listeners have a limited vocabulary with which to distinguish the difference in expressive character.

5. WHAT ARE THE MAIN DIMENSIONS FOR EXPRESSIVE CHARACTER?

Most of the terms in Hevner's adjective checklist [15], as well as the main five dimensions of expressive performance markings identified by Sulem et al. [9] and main factors of the GEMMES [18] are present in listeners' responses. However, the characterizations go beyond the aforementioned clusters in at least three aspects: many terms are non-emotional, technical, or disapproving. Non-emotional terms are e.g., those that are hard to unambiguously place in the arousal–valence space (unlike Hevner's or Sulem et al.'s clusters) such as *'clean'*, *'metallic'* or *'[t]his is Glenn Gould, obviously'*. Technical terms include terms that describe playing techniques such as *'legato'*, *'staccato'*, or more generally *'fast'*, *'loud'*, and *'mechanical'*. Disapproving terms include descriptions with negative connotations such as *'boring'*, *'sterile'*, or *'robotic'*.

Regarding automated analysis, characterization of expressive performance is not a common case in natural language processing (NLP). Also, the meanings of many terms in the context of expressive performance are slightly different from their common usage. Preliminary tests indicate that learned occurrence-based semantics on related or general topic corpora largely fail to represent more than superficial similarity for this dataset.

In order to identify the main dimensions of terms, we compute a principal component analysis (PCA) on the occurrence matrix of the dataset. The data is preprocessed in several steps: answers in other languages are translated to English and terms are stemmed. A term is omitted if (1) it shows up less than three times in the dataset (its contribution to the global variance is minimal) or (2) it is used for all interpretations of the same piece (its piecewise entropy is zero; for instance, a participant wrote *'i love mozart'* (sic) for all performances of the Mozart piece).

Table 2 shows the terms that have the strongest loading on the dimensions in the above PCA. Dimension 1 carries intuitive meaning: its extremes reach from *'hectic'* and *'agitated'* to *'gentle'* and *'calm'*. The other three dimensions are harder to connect to a clear semantic dimension. Note for instance the terms *'cold'* and *'warm'*; both influence dimension 4 strongly in the same direction. Figure 2 displays the terms used to describe Mozart's Sonata in the space spanned by the dimensions 1 and 2. The performances themselves are embedded in the space as the centroid of their respective terms. Three clusters emerge, with the deadpan MIDI and Glenn Gould clearly sticking out, and Mitsuko Uchida slightly more towards the *'calm'* and *'sad'* end of Dim.1 (an impression confirmed by listening).

6. HOW DO MEASURABLE/COMPUTABLE PERFORMANCE FEATURES RELATE TO THE EXPRESSIVE CHARACTER DIMENSIONS?

In this section we study whether there is a systematic relationship between the expressive character dimensions described in Section 5 and measurable or computed performance qualities that can be extracted directly from the audio files or from the score-to-performance alignments. In the rest of this article, we refer to these measurable or computed performance qualities as *performance features*.

6.1 Description of Performance Features

6.1.1 Performance Parameters

We consider two performance parameters, tempo and dynamics curves, to relate to the expressive dimensions described above. The tempo curves are extracted directly from the hand-corrected score-to-performance alignments by computing the inter-beat intervals. For computing loudness we use the loudness curve computed from the MIR Toolbox [25] using a perceptually weighted smoothing of the signal energy. For inter- and intra-piece comparisons, we calculate the average value, standard deviation, kurtosis and skewness of these curves. Average tempo/dynamics provides an indicator of how fast/loud a performance is, the standard deviation quantifies the tempo/loudness deviations, kurtosis provides a measure of how extreme these deviations are, and skewness indicates how asymmetric the tempo/loudness values are (e.g., whether the piece is regularly faster or slower than the average tempo).

6.1.2 Mid-level Features

Mid-level features are perceptual qualities of music such as articulation, rhythmic clarity and modality that describe overall properties of musical excerpts and are intuitively clear to listeners [11]. We extract the seven mid-level features described in [12], using the deep convolutional network architecture from [26] (the A2Mid variant, specifically). The 7 mid-level features are *melodiousness*, *articulation*, *rhythmic complexity*, *rhythmic stability*, *dissonance*, *tonal stability*, and *minorness* (see [12] for a detailed description of the features). We train our model on the mid-level features dataset [12], which contains 5000

Dimension 1				Dimension 2			
positive loading		negative loading		positive loading		negative loading	
hectic	0.17	sad	-0.20	rushed	0.22	hard	-0.19
staccato	0.15	gentle	-0.18	nervous	0.20	stumbling	-0.18
hasty	0.15	tender	-0.18	too fast	0.17	staccato	-0.17
agitated	0.14	calm	-0.16	bit	0.16	ponderous	-0.14
irregular	0.14	graceful	-0.16	hasty	0.15	monotonous	-0.13
Dimension 3				**Dimension 4**			
positive loading		negative loading		positive loading		negative loading	
monotonous	0.22	heavy	-0.14	ok	0.24	cold	-0.15
bad	0.17	graceful	-0.13	happy	0.21	warm	-0.14
warm	0.16	smooth	-0.12	joyful	0.19	floating	-0.14
peaceful	0.16	ponderous	-0.12	free	0.15	blurred	-0.14
beautiful	0.15	soaring	-0.10	breathy	0.14	mysterious	-0.13

Table 2. Terms with strongest loadings for each expressive character dimension.

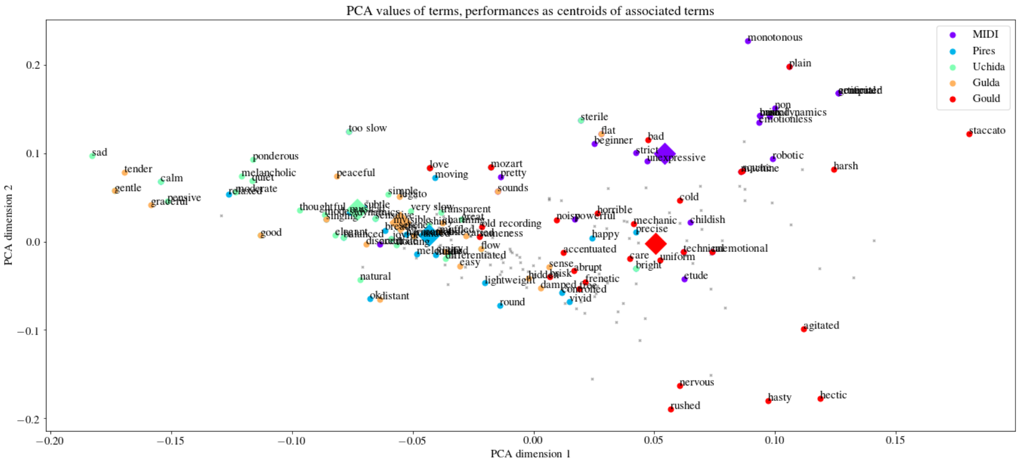

Figure 2. A visualization of the first two dimensions recovered by the PCA. Dots represent terms used in the CEG. Colored terms from characterizations of Mozart's Sonata K 545, grey terms from other pieces. The terms are colored according to the performance they characterize.

audio snippets of 15 seconds each, and use the trained model to predict the mid-level features of the piano performances without any fine-tuning, as there is too little data for a supervised fine-tuning step. To improve the validity of the transfer, we incorporate unsupervised domain adaptation [27] during the training phase. Since the pieces in the CEG are piano performances, we use a separate private collection of non-annotated piano music as the data source for domain adaptation. We observe more variation in the mid-level predictions between the performances while using a domain-adapted model than a non-domain-adapted one, which indicates that it is a useful step in the pipeline.

6.1.3 High-level Features

As high-level emotion-related descriptors, we choose the common *arousal* and *valence* dimensions [19, 28] and aim to predict these from the audio recordings, to then relate them to the expressive character dimensions. We train a dynamic arousal-valence prediction network using the DEAM dataset [29]. We tested a VGG-like model, similar to the one described in Section 6.1.2, and we observed that when the network is pre-trained on the mid-level dataset and extended with a multi-layer GRU-RNN (Gated Recurrent Unit Recurrent Neural Network) that is trained on the DEAM dataset, we get the best results. To improve the results further, we use the recently released receptive-field regularized ResNet [30] for the pre-training phase, since it appears to give better results for short audio snippets than the VGG-like variant. The inputs to our network are Mel-spectrograms and we process 2-second segments of the spectrogram with 0.5-second hops. As in the case of the expressive performances, in order to compare the predicted arousal and valence curves for inter- and intra-piece comparisons, we compute average, standard deviation, kurtosis and skewness of these curves for each performance of each piece.

6.2 Analysis with Multiple Linear Regression

To study the relation between the performance parameters and mid- and high- level features to the expressive character dimensions described in Section 5 we use multiple linear regression (MLR) analysis. In this analysis, the *dependent variables* are each of the expressive character dimensions (Dimensions 1 to 4) and the *independent vari-*

Dimension 1		Dimension 2		Dimension 3		Dimension 4	
PP ($R^2 = 0.24$)		PP ($R^2 = 0.18$)		PP ($R^2 = 0.26$)		PP ($R^2 = 0.24$)	
loudness avg	0.51^{***}	loudness sk	0.45^{**}	loudness std	-0.53^{**}	beat period k	-0.34^{*}
						loudness std	-0.44^{*}
MF ($R^2 = 0.39$)		MF ($R^2 = 0.00$)		MF ($R^2 = 0.00$)		MF ($R^2 = 0.29$)	
rhythmic complexity	-0.74^{*}	minorness	0.15	articulation	-0.15	rhythmic complexity	0.52^{*}
tonal stability	-0.94^{**}					tonal stability	0.84^{***}
articulation	0.46^{*}						
HF ($R^2 = 0.22$)		HF ($R^2 = 0.00$)		HF ($R^2 = 0.36$)		HF ($R^2 = 0.09$)	
valence sk	0.48^{**}	valence avg	0.14	valence k	0.42^{**}	valence k	-0.33^{*}
				arousal avg	-1.24^{***}		
				valence std	0.27^{*}		
				valence avg	-0.82^{*}		

Table 3. Multiple Linear Regression Analysis. PP, MF and HF refer to performance parameters, mid- and high- level features, respectively. avg, std, k and sk denote average, standard deviation, kurtosis and skewness. The values are the regression coefficients (indicating the contribution of that feature to the model). R^2 is the adjusted coefficient of determination for the whole model. *, **, and *** indicate statistical significance at levels $\alpha = .05, .01$ and $\alpha < .001$, respectively.

ables are the performance features described above. We carry out $4 \times 3 = 12$ MLRs for each expressive character dimension (4 in total) and subset of performance features (expressive parameters, mid- and high-level features). Each of these regressions investigates whether each subset of performance features (expressive parameters, mid-/high-level) can significantly predict the position of the pieces in the expressive character dimensions. The position of each piece in the 4D expressive character space is computed as the centroid of all of its terms in this space. For each of these MLRs we perform a variable selection using the Zheng-Loh method [31]. The results are summarized in Table 3. The MLR results indicate that the expressive parameters are significant predictors of all 4 expressive character dimensions, with medium effect sizes (R^2). Mid-level features are only significant predictors of Dimensions 1 and 4. High-level features are only significant for Dimensions 1 and 3. Thus, Dimension 1 (the *'gentle'/'calm'* vs. *'hectic'/'agitated'* axis, see Section 5) seems systematically related to our performance features at all three levels, which further corroborates its significance.

7. DISCUSSION

In Section 4.2 we observed a small positive relationship between the complexity of verbal descriptions and listeners' musical training. We expect that stronger evidence of a relationship would emerge if musical training were better controlled for (our sample had few listeners with < 5 years of training) and the complexity measure were further developed to account for specialized musical terms. Our analysis of listeners' preferred performances in Section 4.3 revealed that the deadpan performances and performances by Glenn Gould were least well-liked. Prior research has suggested that listeners prefer quantitatively average expressive performances [32], which might explain partially the lack of enthusiasm for Gould's idiosyncratic playing.

The results in Sections 4.4 and 5 suggest that listeners tend to describe performances of the same piece similarly, although there is some variability (e.g., a performance can be described both as *'beautiful'* or *'bad'* by different listeners; cf. both *'cold'* and *'warm'* being negatively correlated with Dimension 4). An important issue is that NLP methods for assessing similarity between the descriptions are not really suitable for analyzing performance descriptions, where each term is loaded with complex meaning [4] as well as many cross-domain mappings (e.g., metaphors).

The results in Section 6.2 reveal relationships between performance features and expressive dimensions that conform to musical intuition, with the effects being most pronounced for expressive character Dimension 1 (which is also the one that we find easiest to interpret, see table 2). For instance, the analysis suggests that louder performances or performances with large outliers in the valence curve would be perceived as more irregular and agitated, while softer performances or performances without large outliers in valence would be perceived as calm or graceful.

8. CONCLUSIONS AND FUTURE WORK

This paper has introduced the CEG dataset and presented some exploratory analysis addressing three main questions related to inter-listener agreement, main emerging description dimensions, and relations between user characterizations and measurable performance parameters.

Future work will focus on a more in-depth analysis of the question of semantic similarity. As discussed in Section 5, the description of expressive character includes many nuances that are not well suited to be analyzed with generic NLP methods, given how loaded with meaning certain terms are. We plan to investigate methods like *pile sorting* [33] with expert musicians to devise a meaningful semantic clustering of the terms. Furthermore, we plan to collect more human annotations (e.g., mid- and high-level features) as a basis for a more systematic comparison.

[4] For example, the performance of the Mozart piece by Austrian-trained Japanese pianist Mitsuko Uchida was described by a participant as 'Russian pianist'. To understand this description, it is necessary to have the concept of the Russian School of performance.

9. ACKNOWLEDGMENTS

This research has received support 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") and by the Research Council of Norway through its Centers of Excellence scheme, project number 262762 and the MIRAGE project, grant number 287152. We gratefully acknowledge the effort invested by our music expert, Hans Georg Nicklaus (Anton Bruckner Private University of Music, Linz) for helping with the selection of the different performances in the dataset. We thank Olivier Lartillot for sharing the Matlab code to compute the loudness features.

10. REFERENCES

[1] S. Davies, "Philosophical Perspectives on Music's Expressiveness," in *Music and emotion: Theory and research.*, P. N. Juslin and J. A. Sloboda, Eds. Oxford University Press, 2001, pp. 23–44.

[2] C. Rosen, *Music and Sentiment*. New Haven and London: Yale University Press, 2010.

[3] P. N. Juslin and J. A. Sloboda, "Music and emotion," in *The Psychology of Music*, 3rd ed., D. Deutsch, Ed. London, UK: Academic Press, 2013, ch. 15, pp. 583–645.

[4] P. N. Juslin, "Communicating emotion in music performance: a review and theoretical framework," in *Music and Emotion: Theory and Research*. Oxford University Press, 2001, pp. 309–337.

[5] ——, "Emotional communication in music performance: A functionalist perspective and some data," *Music Perception: An Interdisciplinary Journal*, vol. 14, no. 4, pp. 383–418, 1997.

[6] P. N. Juslin and R. Timmers, "Expression and communication of emotion in music performance," in *Handbook of music and emotion: Theory, research, applications*, P. N. Juslin and J. A. Sloboda, Eds. New York: Oxford University Press, 2010, pp. 453–489.

[7] P. Juslin, "Five facets of musical expression: a psychologist's perspective on music performance," *Psychology of Music*, vol. 31, no. 3, pp. 273–302, 2003.

[8] P. N. Juslin, A. Friberg, and R. Bresin, "Toward a computational model of expression in music performance: The GERM model," *Musicae Scientiae*, vol. 5, no. 1, pp. 63–122, 2001.

[9] A. Sulem, E. Bodner, and N. Amir, "Perception-based classification of expressive musical terms: Toward a parameterization of musical expressiveness," *Music Perception: An Interdisciplinary Journal*, vol. 37, no. 2, pp. 147–164, 2019.

[10] G. Widmer, "Getting Closer to the Essence of Music: The *Con Espressione* Manifesto," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 2, pp. 1–13, Jan. 2017.

[11] A. Friberg, E. Schoonderwaldt, A. Hedblad, M. Fabiani, and A. Elowsson, "Using listener-based perceptual features as intermediate representations in music information retrieval," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1951–1963, 2014.

[12] 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.

[13] 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.

[14] C. Cancino-Chacón, M. Grachten, W. Goebl, and G. Widmer, "Computational Models of Expressive Music Performance: A Comprehensive and Critical Review," *Frontiers in Digital Humanities*, vol. 5, p. 25, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fdigh.2018.00025

[15] K. Hevner, "Experimental Studies of the Elements of Expression in Music," *The American Journal of Psychology*, vol. 48, no. 2, pp. 246–268, April 1936.

[16] P. R. Farnsworth, "A Study of the Hevner Adjective List," *The Journal of Aesthetics and Art Criticism*, vol. 13, no. 1, pp. 97–103, September 1954.

[17] E. Schubert, "Update of the Hevner Adjective Checklist," *Perceptual and Motor Skills*, vol. 96, pp. 1117–1122, 2003.

[18] S. Schaerlaeken, D. Glowinski, M.-A. Rappaz, and D. Grandjean, ""Hearing Music as ...": Metaphors Evoked by the Sound of Classical Music," *Psychomusicology: Music, Mind and Brain*, vol. 29, no. 2-3, pp. 100–116, 2019.

[19] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[20] M. Murari, A. Rodà, S. Canazza, G. D. Poli, and O. D. Pos, "Is vivaldi smooth and takete? non-verbal sensory scales for describing music qualities," *Journal of New Music Research*, vol. 44, no. 4, pp. 359–372, 2015. [Online]. Available: https://doi.org/10.1080/09298215.2015.1101475

[21] C. Cannam, C. Landone, and M. Sandler, "Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the ACM Multimedia 2010 International Conference*, Firenze, Italy, October 2010, pp. 1467–1468.

[22] E. Dale and J. S. Chall, “A Formula for Predicting Readability,” *Education Research Bulletin*, vol. 27, no. 1, pp. 11–20+28, January 1948.

[23] Y. Li, D. McLean, Z. Bandar, J. D. O’Shea, and K. Crockett, “Sentence Similarity Based on Semantic Nets and Corpus Statistics,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 8, pp. 1138–1150, August 2006.

[24] G. A. Miller, “WordNet: A Lexical Database for English,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[25] Lartillot, Olivier and Toiviainen, Petri, “A Matlab toolbox for musical feature extraction from audio,” in *Proceedings of the International Conference on Digital Audio Effects (DAFx-07)*, Bordeadux, France, 2007.

[26] 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)*, Delft, The Netherlands, 2019.

[27] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proceedings of the 32th International Conference on Machine Learning (ICML 2015)*, Lille, France, 2015.

[28] N. H. Frijda, *The emotions*. Cambridge, UK: Cambridge University Press, 1986.

[29] A. Aljanaki, Y.-H. Yang, and M. Soleymani, “Developing a benchmark for emotional analysis of music,” *PloS one*, vol. 12, no. 3, 2017.

[30] K. Koutini, H. Eghbal-zadeh, and G. Widmer, “Receptive-field-regularized cnn variants for acoustic scene classification,” in *Proceedings of the Accepted at Detection and Classification of Acoustic Scenes and Events 2019 (DCASE Workshop 2019)*, New York, NY, USA, 2019.

[31] X. Zheng and W.-Y. Loh, “A consistent variable selection criterion for linear models with high-dimensional covariates,” *Statistica Sinica*, vol. 7, no. 2, pp. 311–325, 1997. [Online]. Available: http://www.jstor.org/stable/24306081

[32] B. Repp, “The aesthetic quality of a quantitatively average music performance: Two preliminary experiments,” *Music Perception*, vol. 14, no. 4, pp. 419–444, 1997.

[33] R. Trotter and J. M. Potter, “Pile sorts, a cognitive anthropological model of drug and aids risks for navajo teenagers: Assessment of a new evaluation tool,” *Drugs & Society*, vol. 7, no. 3-4, pp. 23–39, 1993.