J Intell Inf Syst (2018) 51:415–437
https://doi.org/10.1007/s10844-018-0510-y

Musical performance analysis in terms of emotions it evokes

Jacek Grekow[1]

Received: 6 December 2017 / Revised: 8 May 2018 / Accepted: 21 May 2018 /
Published online: 16 June 2018

Abstract Finding pieces with a similar emotional distribution throughout the same composition was the aim of this work. A comparative analysis of musical performances by using emotion tracking was proposed. A dimensional approach of dynamic music emotion recognition was used in the analysis. Music data annotation and regressor training were done. Values of arousal and valence, predicted by regressors, were used to compare performances. The obtained results confirm the validity of the assumption that tracking and analyzing the values of arousal and valence over time in different performances of the same composition can be used to indicate their similarities. Detailed results of analyzing different performances of Prelude No.1 by Frédéric Chopin were presented. They enabled to find the most similar performances to the performance by Arthur Rubinstein, for example. The author found which performances of the same composition were closer to each other and which were quite distant in terms of the shaping of arousal and valence over time. The presented method gives access to knowledge on the shaping of emotions by a performer, which had previously been available only to music professionals.

Keywords Emotion tracking · Musical performances · Similarity

1 Introduction

The question "Which pianist plays Chopin like Rubinstein in terms of the emotions?" is one we can ask ourselves when we are listening to many performances of the same composition

✉ Jacek Grekow
j.grekow@pb.edu.pl

[1] Faculty of Computer Science, Bialystok University of Technology, Wiejska 45A, Bialystok 15-351, Poland

and one of the performers is the renown pianist Artur Rubinstein, who we like. The same composition, based on one musical notation, can be performed differently, with each performance differing in the emotional content. Performing a piece written by a composer, performer, musician, artist gives it its own shape – interpretation. We can enjoy some performances more than others.

In this paper, the stages of building a computer system that enables to find pieces with a similar emotional distribution throughout the same composition were presented. The building stages include such issues as music data annotation, building of the regressors, aligning of different renditions, visualization of emotions over time, and result analysis.

Such a computer solution could be an extension of systems searching musical compositions on Internet databases, which more and more often add an option of selecting emotions to the basic search parameters, such as title, composer, genre, etc. A system that compares the details of different performances could compare performances more precisely and quickly than people. The question "Which performance of the same composition is similar or different?" is often asked during music competitions, such as The International Frédéric Chopin Competition. Musicologists studying the interpretation of compositions could also be interested in such a system.

2 Related work

For comparative analysis of musical performances, the dimensional approach of dynamic music emotion recognition was used. Emotion recognition was treated as a regression problem. A 2D emotion model proposed by Russell (1980) was used, where the dimensions are represented by arousal and valence. It had been used in many works used for music emotion recognition (Schmidt et al. 2010; Yang et al. 2008).

Dynamic music emotion recognition analyzes changes in emotions over time. Methods for detecting emotion using a sliding window are presented in Grekow (2012, 2016), Korhonen et al. (2005), Lu et al. (2006), Schmidt et al. (2010), Yang et al. (2008).

Comparisons of multiple performances of the same piece often focused on piano performances (Goebl et al. 2004; Sapp 2007). Tempo and loudness information were the most popular characteristics used for performance analysis. They were used to calculate correlations between performances in (Sapp 2007; 2008). In the study (Goebl et al. 2004), tempo and loudness derived from audio recordings were segmented into musical phrases, and then clustering was used to find individual features of the pianists' performances.

Four selected computational models of expressive music performance were reviewed in Widmer and Goebl (2004). In addition, research on formal characterization of individual performance style, like performance trajectories and performance alphabets, was presented.

A method to compare orchestra performances by examining a visual spectrogram characteristic was proposed in Liem and Hanjalic (2015). Principal component analysis on synchronized performance fragments was applied to localize areas of cross-performance variation in time and frequency.

A connection between music performances and emotion was presented in Bresin and Friberg (2000), where a computer program (Director Musices) was used to produce performances with varying emotional expression. The program used a set of rules characteristic for each emotion (fear, anger, happiness, sadness, solemnity, tenderness), which were used to modify such parameters of MIDI files as tempo, sound level, articulation, tone onsets and delays.

This paper presents a comparative analysis of musical performances by using emotion tracking. Use of emotions for comparisons is a novel approach, not found in the work of other authors.

3 System for comparative analysis of musical performances

The proposed system for comparative analysis of musical performances using emotion tracking is shown in Fig. 1. It is composed of collected music training data, segmentation, feature extraction, regressors, aligning, and a result presentation module.

The input data are different performances of the same composition, which underwent segmentation. After the feature extraction process, the prediction of arousal and valence occurs for subsequent segments, and previously trained regressors are used for prediction. In the next phase, the valence and arousal values are aligned in the aligning module, which causes that the same musical fragments of different performances are compared. The obtained results are sent to the result presentation module, where the course of arousal and valence over time is presented, scape plots are constructed, and parameters indicating the most alike compositions are calculated.

4 Building prediction models

4.1 Music data for regressor training

In our approach, emotion recognition was treated as a regression problem. The data set for regressor training consisted of 324 6-second fragments of different genres of music:

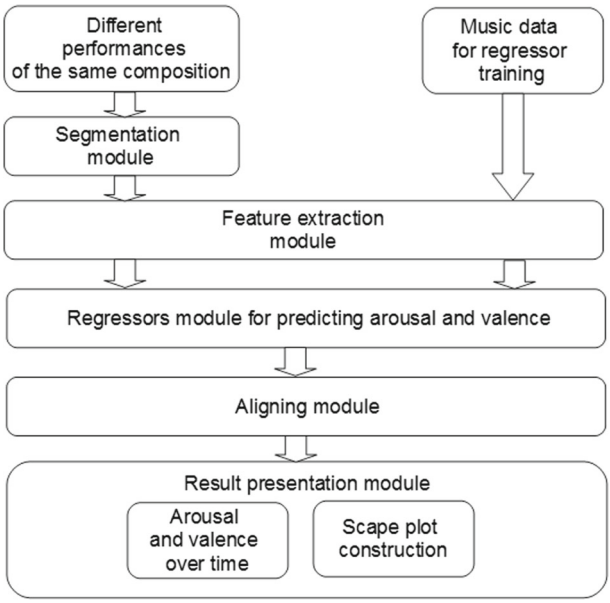

Fig. 1 System construction for comparative analysis of musical performances by using emotion tracking

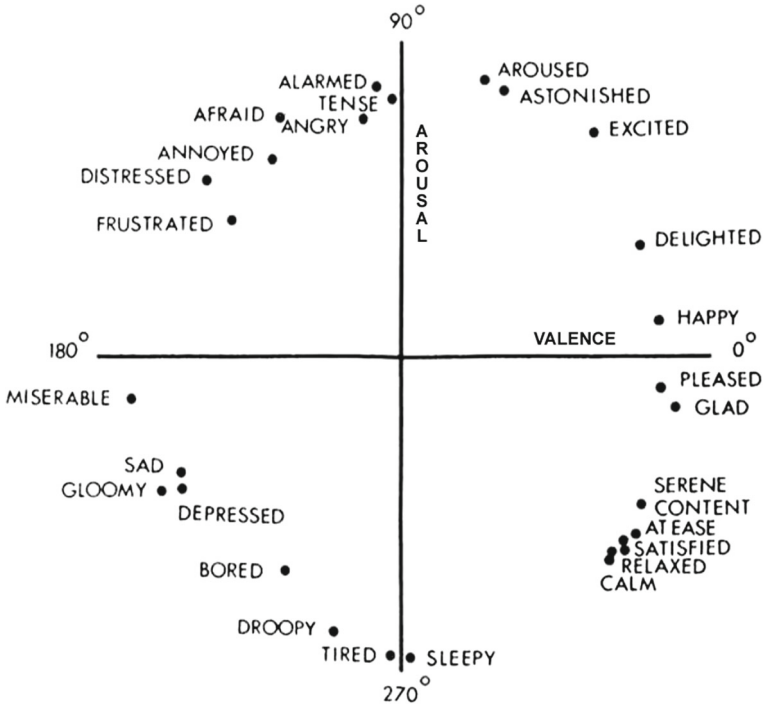

Fig. 2 Russell's circumplex model (Russell 1980)

classical, jazz, blues, country, disco, hip-hop, metal, pop, reggae, and rock. The tracks were all 22050 Hz mono 16-bit audio files in .wav format. The training data were taken from the generally accessible data collection project MARSYAS.[1] The author selected samples and shortened them to the first 6 seconds.

Data annotation of perceived emotion was done by five music experts with a university musical education. Each annotator annotated all records in the dataset, which has a positive effect on the quality of the received data (Aljanaki et al. 2016). During annotation of music samples, the two-dimensional arousal-valence (A-V) model to measure emotions in music (Russell 1980) was used. The model (Fig. 2) consists of two independent dimensions of valence (horizontal axis) and arousal (vertical axis). Each person making annotations, after listening to a music sample, had to specify values on the arousal and valence axes in a range from -10 to 10 with step 1. On the arousal axis, a value of -10 meant low while 10 high arousal. On the valence axis, -10 meant negative while 10 positive valence. The data collected from the five music experts was averaged.

The amount of examples in the quarters on the A-V emotion plane is presented in Table 1. Pearson correlation coefficient was calculated to check if valence and arousal dimensions are correlated in our music data. The obtained value $r = -0.03$ indicates that arousal and valence values are not correlated, and the music data are a good spread in the quarters on the A-V emotion plane. This is an important element according to the conclusions formulated in Aljanaki et al. (2016).

[1] http://marsyas.info/downloads/datasets.html

Table 1 Amount of examples in quarters on the A-V emotion plane

Quarter Abbreviation	Quarter Arousal-Valence	Amount of examples
Q1	high-high	93
Q2	high-low	70
Q3	low-low	80
Q4	low-high	81

The previously prepared, labeled by A-V values, music data set served as input data for the tool used for feature extraction. For feature extraction, a tool for audio analysis Essentia (Bogdanov et al. 2013) was used. The obtained by Essentia length of feature vector was 530 features.

4.2 Regressor training

Regressors for predicting arousal and valence using the WEKA package (Witten and Frank 2005) were built. For training and testing, the following regression algorithms were used: SMOreg, REPTree, M5P. Before constructing regressors, arousal and valence annotations were scaled between $[-0.5, 0.5]$.

The performance of regression was evaluated using the tenfold cross validation technique (CV-10). The whole data set was randomly divided into ten parts, nine of them for training and the remaining one for testing. The learning procedure was executed a total of 10 times on different training sets. Finally, the 10 error estimates were averaged to yield an overall error estimate.

The highest values for determination coefficient (R^2) were obtained using SMOreg (implementation of the support vector machine for regression). After applying attribute selection (attribute evaluator: Wrapper Subset Evaluator, search method: Best First), $R^2 = 0.79$, for arousal and $R^2 = 0.58$ for valence were obtained. Mean absolute error reached values $MAE = 0.09$ for arousal and $MAE = 0.10$ for valence. Predicting arousal is a much easier task for regressors than valence and the values predicted for arousal are more precise.

More detailed results from the conducted experiments during the building of the regressors used in this work have been presented in the article (Grekow 2017), where the usefulness of audio features during emotion detection in music files was presented. Different features sets were used to test the performance of the built regression models intended to detect arousal and valence.

5 Aligning of different renditions

Our task was to compare different performances of the same composition using emotional distribution. Because the musical performances are played at varying tempos, with various accelerations and decelerations, an alignment of audio recordings is necessary to compare two performances. This enables comparing the same fragments of different renditions. Without doing an alignment to compare performances second to second, we would be comparing fragments of varying content. Just adjusting the time of different renditions, for example through stretching or compression of time, will not synchronize the performances in terms of music content. Only an exact alignment of the recordings, note by note, guarantees that we are comparing the same fragments.

MATCH (Dixon and Widmer 2005), a toolkit for accurate automatic alignment of different renditions of the same piece of music, was used. MATCH is based on a dynamic time warping algorithm (DTW), which is a technique for aligning time series and has been well known and used in the speech recognition community (Rabiner and Juang 1993). Frames of audio input data are represented by positive spectral difference vectors, which emphasize note onsets in the alignment process. They are used in the DTW algorithm's match cost function, which uses an Euclidean metric. The path returned by the DTW algorithm, as result of alignment of two audio files, is used to find the location of the same musical fragment in both files.

Figure 3 present, in waveform images, the beginnings of three different renditions of the same composition (Prelude in C major, Op.28, No.1 by Frédéric Chopin) before and after alignment. Before alignment (Fig. 3a), the compositions are placed one after the other and the vertical line indicates the time from the beginning of the composition, but these are different fragments in terms of music content. After alignment (Fig. 3b), the vertical line indicates the same fragment in different performances. We notice the varying locations of the same motif from the beginning of the composition depending on the rendition, which is connected to the differing tempos played by different performers. The top first recording is a reference recording and the remaining pieces are compared to it. To present the waveform images of audio files and to visualize the alignment results, Sonic Visualizer (Cannam et al. 2010) with installed MATCH Vamp Plugin was used.

6 Analyzed performances

The collection of analyzed performances consisted of the following compositions by Frédéric Chopin (1810-1849):

- Prelude in C major, Op.28, No.1;
- Prelude in D major, Op.28, No.5;
- Prelude in F minor, Op.28, No.18;
- Prelude in C minor, Op.28, No.20 (the first 8 bars).

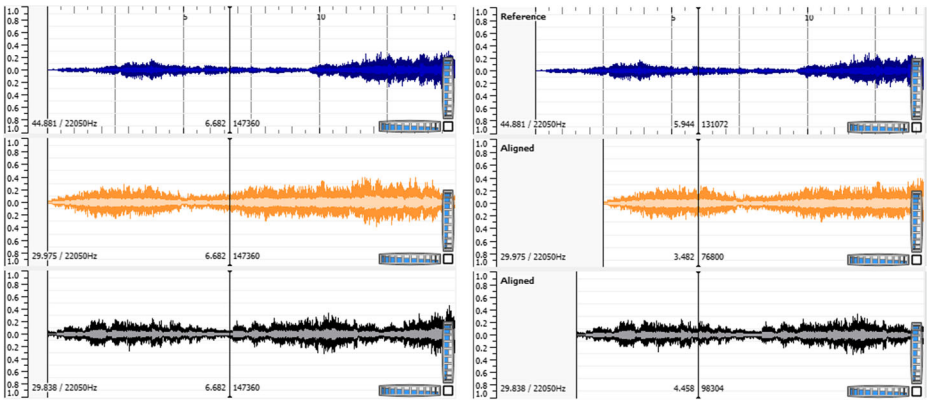

Fig. 3 Waveform images of three different music performances of Prelude in C major, Op.28, No.1 by Frédéric Chopin before alignment (**a**) and after alignment (**b**)

All the analyzed Chopin performances were audio recordings played by 5 famous pianists:

- Artur Rubinstein recorded in 1946;
- Emil Gilels recorded in 1953;
- Grigory Sokolov recorded in 1990;
- Martha Argerich recorded in 1997;
- Rafał Blechacz recorded in 2007.

Four musical works, with 5 different performances of each, were analyzed. Detailed results are available on the web.[2]

7 Results

The best obtained regressors (Section 4.2) were used for predicting arousal and valence of the musical performances, which were divided into 6-second segments with a 3/4 overlap. For each segment, features were extracted and regressors for arousal and valence were used. As a result, arousal and valence values for every 1.5 seconds of a musical piece were obtained.

7.1 Performances and arousal over time

Due to the limited nature of the research, the analysis and presentation of the results was restricted to 5 different performances of Prelude in C major, Op.28, No.1 by Frédéric Chopin.

Observation of the course of arousal in the performances (Fig. 4) shows that the performance by G. Sokolov had significantly lower values than the remaining pieces. The reason for this is that the performance was played at a slower pace and lower sound intensity. Between samples 15 and 20, there was a clear rise in arousal for all performances, but the performance by E. Gilels achieved maximum (Arousal = 0.3). Also, the performance by E. Gilels is the most dynamically aroused in this fragment. It is not always easy to detect on the graph which performances are similar. However, we can notice a convergence of lines between A. Rubinstein and G. Sokolov (samples 30-50), and between R. Blechacz and E. Gilels (samples 37-50).

To compare the musical performances, the Pearson correlation coefficient r was used, which was calculated for each pair of performances (Table 2). Each performance is represented by a sequence of arousal values. The correlation coefficient ranges from 1 to -1. Value 1 means perfectly correlated sequences, 0 there is no correlation, and -1 that the sequences are perfectly correlated negatively. A correlation coefficient between two of the same performances is at maximum equal to 1 and was not taken into account, i.e. $r(X, X) = 1$.

Pearson's correlation coefficient r between two sequences $(x_1, x_2, ..., x_n)$ and $(y_1, y_2, ..., y_n)$ of the same length n is defined in (1).

$$r(x, y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{1}$$

[2] http://aragorn.pb.bialystok.pl/grekowj/HomePage/Performances

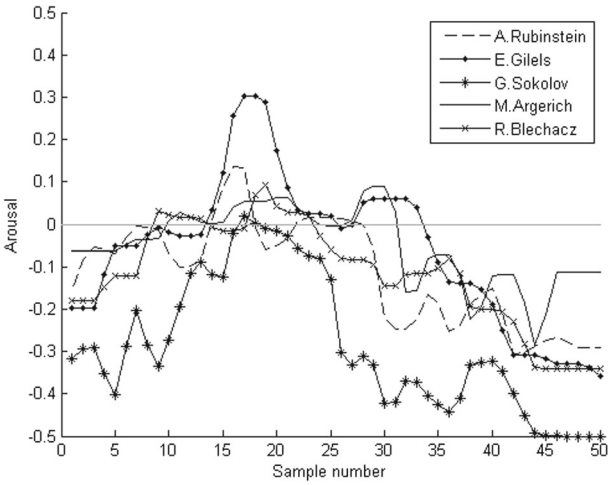

Fig. 4 Arousal over time for 5 performances of Prelude in C major, Op.28, No.1 by Frédéric Chopin

where x_i and y_i are values of elements in sequences, $\overline{x}$ and $\overline{y}$ are mean values of each sequence.

Comparing the musical performances in terms of arousal (Table 2), we see that the performance by A. Rubinstein was most similar to G. Sokolov ($r = 0.80$). G. Sokolov was most similar to A. Rubinstein ($r = 0.80$) and R. Blechacz ($r = 0.83$). The performances by R. Blechacz were similar to those by E. Gilels ($r = 0.87$) and G. Sokolov ($r = 0.83$). M. Argerich's performance is the least similar to the rest, although it is closest to E. Gilels ($r = 0.74$) and R. Blechacz ($r = 0.75$). From all 5 performances in terms of arousal, the pieces by G. Sokolov and M. Argerich were the most different ($r = 0.67$).

7.2 Performances and valence over time

Observation of the course of valence in the performances (Fig. 5) shows that there was a similar decrease in this value for samples 15-20 in all performances. There is a similar line shape, i.e. good correlation, between R. Blechacz and G. Sokolov (samples 5-20), and between M. Argerich and E. Gilels (samples 35-50).

Table 3 presents correlation coefficients for valence calculated for each pair of performances. In terms of valence distribution, G. Sokolov's performance was similar to A.

Table 2 Correlation coefficient r for arousal calculated for each pair of performances of Prelude in C major, Op.28, No.1 by F. Chopin

	A. Rubinstein	E. Gilels	G. Sokolov	M. Argerich	R. Blechacz
A. Rubinstein	1.00	0.74	**0.80**	0.69	0.75
E. Gilels	0.74	1.00	0.78	0.74	**0.87**
G. Sokolov	**0.80**	0.78	1.00	0.67	**0.83**
M. Argerich	0.69	0.74	0.67	1.00	0.75
R. Blechacz	0.75	**0.87**	**0.83**	0.75	1.00

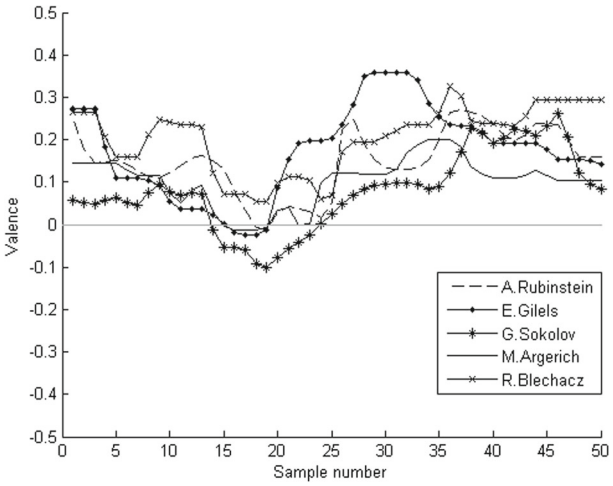

Fig. 5 Valence over time for 5 performances of Prelude in C major, Op.28, No.1 by Frédéric Chopin

Rubinstein ($r = 0.82$) and R. Blechacz ($r = 0.81$). The performance by E. Gilels was similar to M. Argerich ($r = 0.70$), and less similar to G. Sokolov, A. Rubinstein and R. Blechacz (r = 0.41-0.42).

7.3 Performances and arousal-valence over time

Another possibility of comparing performances is to take arousal and valence into account simultaneously. To compare performances described by two sequences of values (arousal and valence), these sequences should be joined. In order to join the two sequences of arousal (2) and valence (3) values of one performance, standard deviation and the mean of the two sequences of data should be equivalent. It was decided to leave arousal values without change and convert valence values (5); although we could have converted arousal and left valence without a change and this would not have affected the correlation results. During joining, the sequences of each feature are interleaved (4).

Sequences for arousal and valence:

$$A = (a_1, a_2, a_3, ..., a_n) \tag{2}$$

$$V = (v_1, v_2, v_3, ..., v_n) \tag{3}$$

Table 3 Correlation coefficient r for valence calculated for each pair of performances of Prelude in C major, Op.28, No.1 by F. Chopin

	A. Rubinstein	E. Gilels	G. Sokolov	M. Argerich	R. Blechacz
A. Rubinstein	1.00	0.41	**0.82**	0.65	0.79
E. Gilels	0.41	1.00	0.41	0.70	0.42
G. Sokolov	**0.82**	0.41	1.00	0.61	**0.81**
M. Argerich	0.65	0.70	0.61	1.00	0.70
R. Blechacz	0.79	0.42	**0.81**	0.70	1.00

Table 4 Correlation coefficient r for joined arousal and valence calculated for each pair of performances of Prelude in C major, Op.28, No.1 by F. Chopin

	A. Rubinstein	E. Gilels	G. Sokolov	M. Argerich	R. Blechacz
A. Rubinstein	1.00	0.58	**0.81**	0.67	0.77
E. Gilels	0.58	1.00	0.60	0.72	0.64
G. Sokolov	**0.81**	0.60	1.00	0.64	**0.82**
M. Argerich	0.67	0.72	0.64	1.00	0.72
R. Blechacz	0.77	0.64	**0.82**	0.72	1.00

Result of joining arousal and valence sequence into one sequence AV:

$$AV = (a_1, vnew_1, a_2, vnew_2, a_3, vnew_3, ..., a_n, vnew_n) \tag{4}$$

Formula for calculating new valence sequence values:

$$vnew_n = \overline{a} + sd_a \frac{v_n - \overline{v}}{sd_v} \tag{5}$$

where sd_a stands for standard deviation of sequence A, sd_v - standard deviation of sequence V, $\overline{a}$ - mean value of sequence A, $\overline{v}$ - mean value of sequence V.

Table 4 presents the correlation coefficients r for joined arousal and valence, calculated for each pair of performances. We see that the most similar performances are by R. Blechacz and G. Sokolov ($r = 0.82$) and the most different by A. Rubinstein end E. Gilels ($r = 0.58$). It can be stated that in terms of arousal and valence we have two groups of performances. The first group consists of performances by R. Blechacz, G. Sokolov and A. Rubinstein, and the second group by E. Gilels and M. Argerich.

7.4 Visualization of similarity using dendrograms

On the basis of correlation coefficients presented in Table 4, hierarchical clustering was done. The dendrogram obtained from clustering is presented in Fig. 6a. As can be seen, the

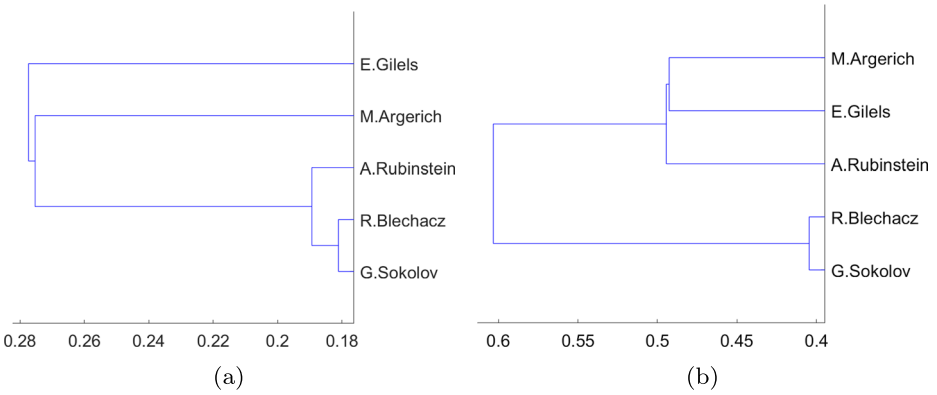

Fig. 6 Dendrogram for 5 performances of Prelude in C major, Op.28, No.1 (**a**) and Prelude in C minor, Op.28, No.20 (**b**)

performances by G. Sokolov, R. Blechacz, and A. Rubinstein were the most correlated. The performance by E. Gilels was the farthest from this group.

A slightly different correlation distribution can be seen on the dendrogram for Prelude in C minor, Op.28, No.20 (Fig. 6b). We can see that similarities between various performers are not always the same for different compositions. In this case, the performance by A. Rubinstein is closer to the performance by M. Argerich and E. Gilels.

7.5 Arousal-valence trajectory

Figure 7a presents the trajectories of two different performances (E. Gilels, G. Sokolov) of Prelude No.1 by Frédéric Chopin on the A-V emotion plane ($r = 0.60$, Table 4). A square marker on the trajectory indicates the beginning of a piece.

The trajectories illustrate how the artist moved in the 4 quarters on the Arousal-Valence emotion plane. Both performances begin and end in quarter Q4 (arousal low - valence high). The course of the middle part of these two performances varies. G. Sokolov's trajectory moves through quarter Q3 (arousal low - valence low), and E. Gilels' through quarters Q2 (arousal high - valence low) and Q1 (arousal high - valence high).

Figure 7b presents the trajectories of performances by R. Blechacz and G. Sokolov, which are more similar than trajectories in Fig. 7a. Bigger similarity also expresses a greater correlation coefficient between two trajectories ($r = 0.82$, Table 4).

7.6 Scape plot

Comparing musical performances by using correlation coefficient r calculated for the whole length of a composition is only a general analysis of similarities between recordings. In order to analyze similarities between the performances in greater detail, scape plotting was used.

Scape plot is a plotting method that allows presenting analysis results for segments of varying lengths on one image. Their advantage is that they enable to see the entire structure of a composition. Any type of analysis can be used during scape plot construction. In our case, analysis consisted of assessing correlations between arousal and valence values of different musical performances.

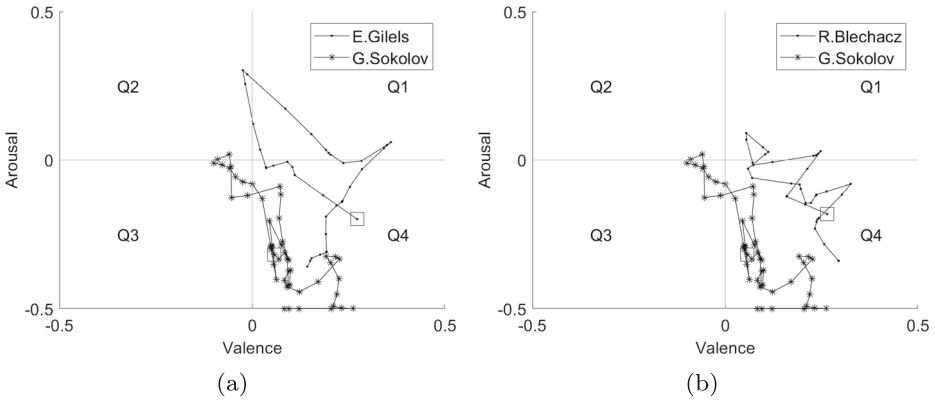

Fig. 7 Performances of Prelude in C major, Op.28, No.1 by Frédéric Chopin on the A-V emotion plane

The scape plotting method was designed by Craig Sapp for structural analysis of harmony in musical scores (Sapp 2001). It has also been applied to timbral analysis (Segnini and Sapp 2006) and for visualization of the tonal content of polyphonic audio signals (Gómez and Bonada 2005). The scape plots were used to present similarities between tempo and loudness features extracted from recordings of the same musical composition (Sapp 2007, 2008). Scape plots are also used for visualizing repetitive structures of music recordings (Mủller and Jiang 2012).

In this paper, a scape plot is used to present calculated correlations between analogous segments of the examined recordings. A comparison of sequences of emotional features (arousal and valence) in different performances of the same composition is a novel application for scape plots.

Figure 8 presents a method for creating a scape plot to analyze a sample composition consisting of 5 elements: a, b, c, d, e. These elements are first examined separately, and then grouped by sequential pairs: ab, bc, cd, de. Next, 3-element sequences are created: abc, bcd, cde; followed by 4-element sequences: $abcd, bcde$; and finally one sequence consisting of the entire composition: $abcd$e. The obtained sequences are arranged on a plane in the form of a triangle, where at the base are the analysis results of examining the shortest sequences and at the top of the triangle are the results of analyzing the entire length of the composition. In a scape plot, the horizontal axis represents time in the composition, while the vertical axis represents the length of the analyze d sequence.

Figure 9 shows a sample result of creating a scale plot using arousal sequences for 5 different performances. First, a reference performance is selected for the scape plot; in this case, it was performance AAA. Then, for each cell in the scape plot, the arousal correlation between the reference performance and all other performances is calculated. Finally, the winner, i.e. the performance with the highest correlation value, is denoted using a color for each cell. Additionally, a percentage content is calculated for each winning performance.

On the provided example, reference performance AAA is most correlated with performance BBB when comparing the entire length of the composition, as indicated by the top element of BBB in the triangle. This is additionally confirmed by the percentage of wins, i.e. the area occupied, by performance BBB (59%). Performance CCC is in second place with 27% wins. The placement of the wins for this performance, bottom left of the scape plot, indicates a similarity with the reference performance in the first half of the composition. Performances DDD and EEE showed little similarity (4% and 10% wins) with the reference performance.

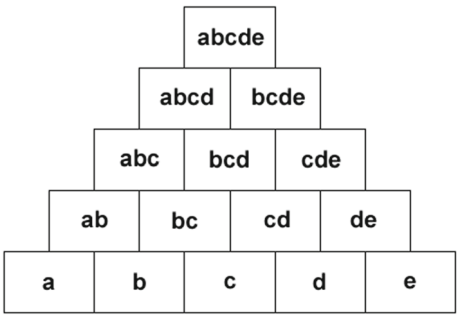

Fig. 8 Method for creating a scape plot from 5 elements

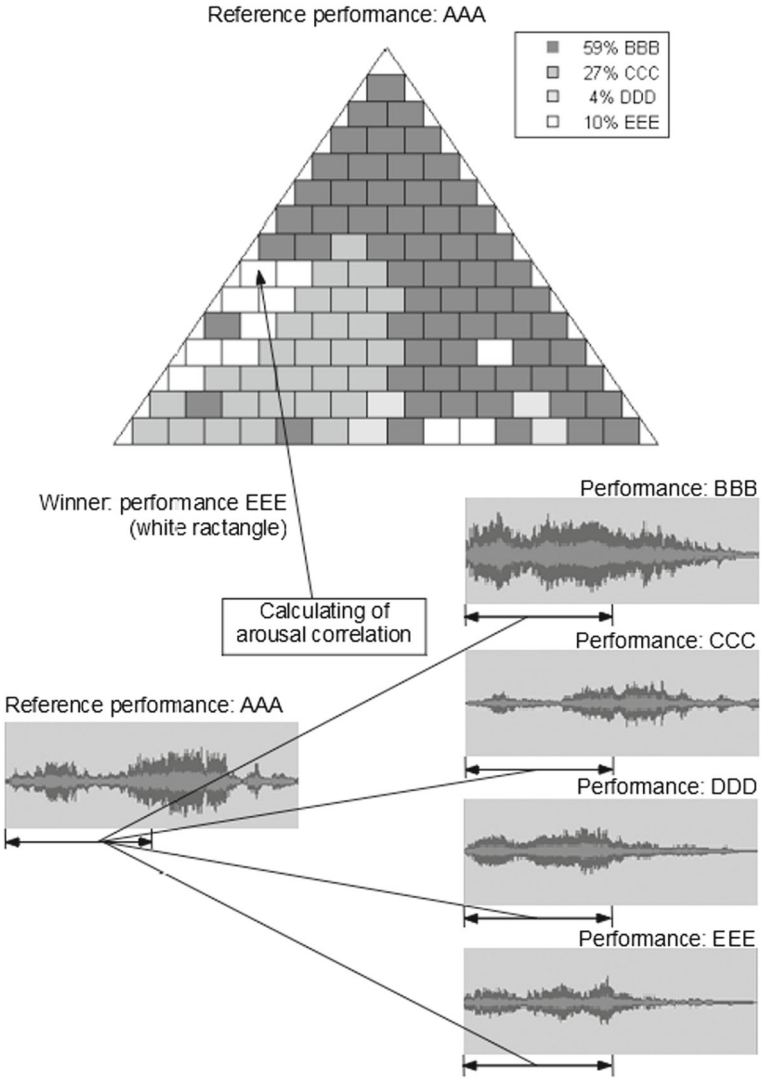

Fig. 9 Result of creating a scape plot by using arousal correlation between the reference performance and 4 other performances

7.7 Arousalscape

Figure 10 shows the Arousalscape, a scape plot generated for the arousal value sequence for 5 different performances of Prelude in C major, Op.28, No.1 by Frédéric Chopin. The performance by A. Rubinstein was selected as the reference performance.

Arousalscape illustrates which performances are the most correlated to A. Rubinstein, by different lengths of the examined sequences. In the lower levels of the triangle, it is difficult to choose the winner; but in the higher levels of the triangle, the winner is unequivocal: the performance by G. Sokolov. His area in the Arousalscape is the biggest and reaches a value of 59%. It is interesting that on the first half of Prelude, the performance by E. Gilels is the

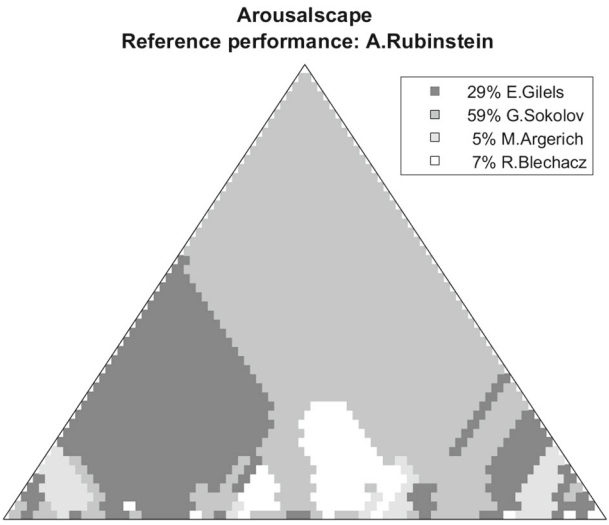

Fig. 10 Arousalscape for Prelude in C major, Op.28, No.1, reference performance: A. Rubinstein

most similar to A. Rubinstein. The remaining two performances (M. Argerich, R. Blechacz) were covered by the first two winners during comparison with the reference performance (A. Rubinstein).

7.8 Valencescape

Figure 11 shows the Valencescape, a scape plot generated for the valence value sequence for 5 different performances of Prelude in C major, Op.28, No.1 by Frédéric Chopin. The performance by A. Rubinstein was selected as the reference performance.

The situation here is not as unequivocal as for Arousalscape for the same composition (Section 7.7). Once again the winner was the performance by G. Sokolov, but its area on the Valencescape is smaller (49%). The triangle area is ruptured by the color white, which represents the performance by R. Blechacz (35% of the area). This means that this performance was also well correlated with the reference performance at many moments.

7.9 AVscape

Figure 12 shows the AVscape, a scape plot generated for sequences of joined arousal and valence values for 5 different performances of Prelude in C major, Op.28, No.1 by Frédéric Chopin. The performance by A. Rubinstein was selected as the reference performance. The definitive winner of the comparisons of correlation values with various lengths of sequences is the performance by G. Sokolov (67% of the area), with E. Gilels in second place (17% of the area), and R. Blechacz in third place (14% of the area). At the beginning of the composition, the reference performance by A. Rubinstein is similar to E. Gilels (dark colored left lower part of the triangle), and in the middle of the composition to R. Blechacz (white space in the middle lower part of the triangle).

When comparing 5 different performances of one composition, we can create 5 AVscapes, 5 Arousalscapes, and 5 Valencescapes. Each rendition is consecutively selected as

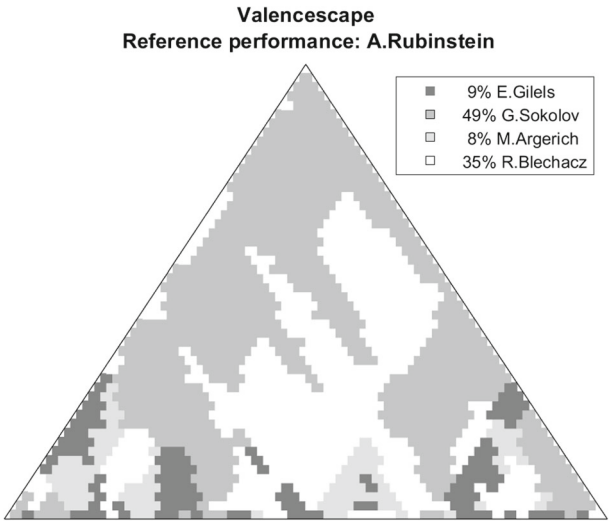

Fig. 11 Valencescape for Prelude in C major, Op.28, No.1, reference performance: A. Rubinstein

the reference performance and the degree of similarity is visualized in detail in comparison with the other 4 performances.

Figure 13 presents the remaining four AVscapes for Prelude No.1, where as the reference performance the remaining performers – E. Gilels, G. Sokolov, M. Argerich, R. Blechacz – were selected. This allowed for a detailed analysis of similarities between the reference and other performances, at various sections of the composition: at the beginning, middle, end, or throughout the entire composition.

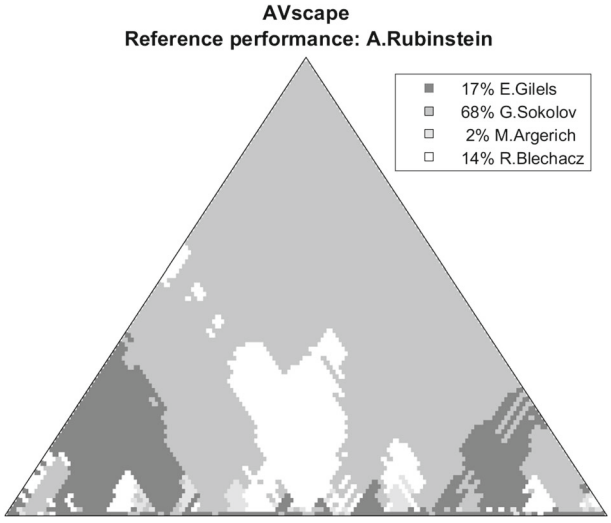

Fig. 12 AVscape for Prelude in C major, Op.28, No.1, reference performance: A. Rubinstein

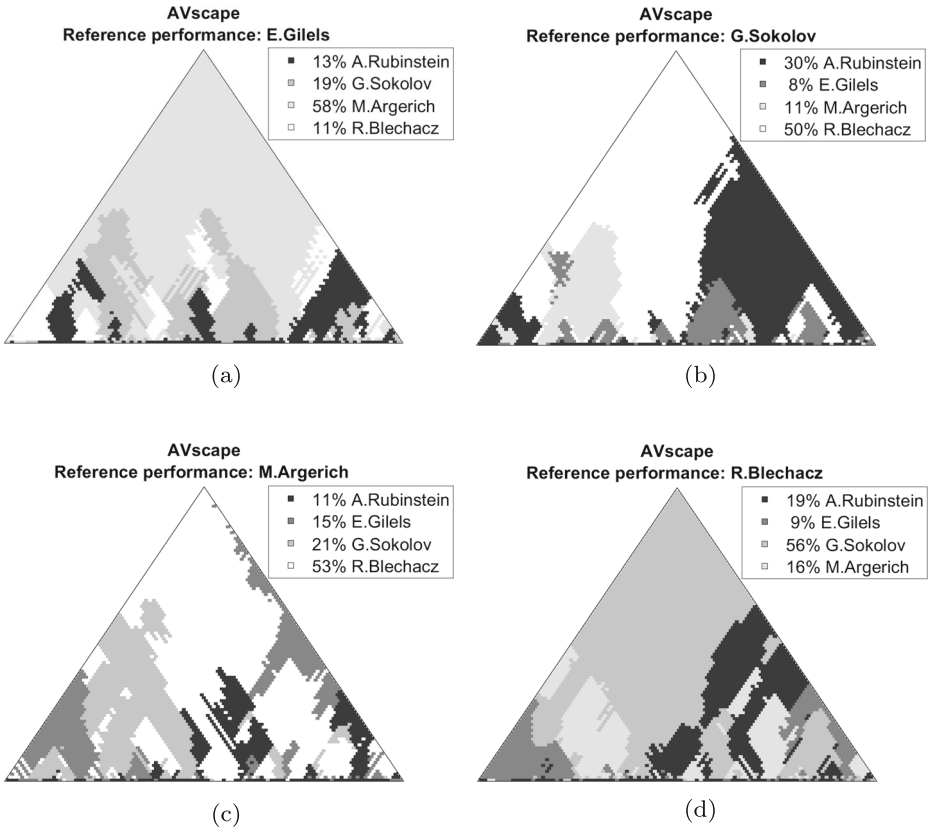

Fig. 13 AVscape for Prelude in C major, Op.28, No.1, reference performance: E. Gilels (**a**), G. Sokolov (**b**), M. Argerich (**c**), R. Blechacz (**d**)

8 Evaluation

8.1 Parameters describing the most similar performances

To find the most similar performance to the reference performance, you can use several indicators that result from the construction of the obtained scape plots (Sapp 2008):

- Score S_0 - the most general result indicating the most similar performance. The winner is the one with the best correlation for the entire sequence, entire length of time of the composition. On the scape plot, it is the top element of the triangle.
- Score S_1 - indicates the performance with the biggest area in the scape plot. The area of wins of a given performance shows its dominance at various lengths of analyzed sequences. The winner with the best correlation for the entire sequence (S_0) does not always have the largest area, or the largest number of wins on the scape plot.
- Score S_2 - the next best similar performance from the scape plot, calculated after removing the winner S_1. If two performances are very similar, then one will always win and cover the wins of the second. To calculate the S_2 score, a new scape plot is generated

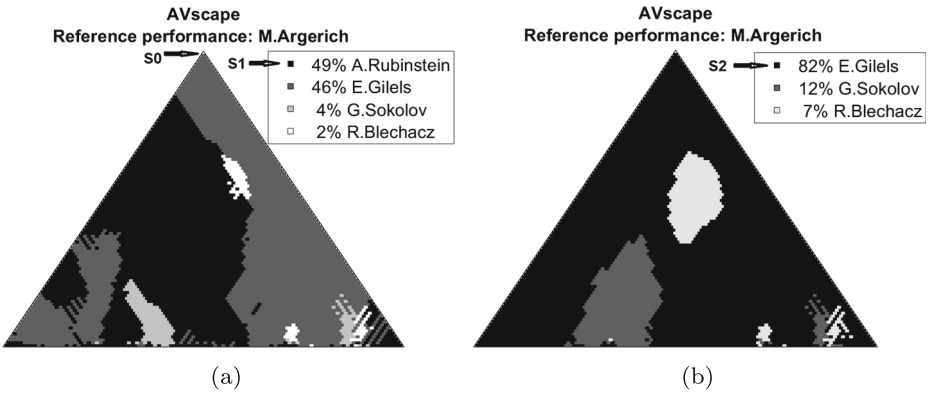

Fig. 14 AVscape for Prelude in C minor, Op.28, No.20, reference performance: M. Argerich (**a**) and AVscape with winner S_1 removed (**b**)

without the winner as indicated by S_1. It shows the performance with the biggest area in the newly created scape plot.

Figure 14a shows the AVscape for Prelude in C minor, Op.28, No.20, reference performance: M. Argerich. Score S_0 in this case is the performance by E. Gilels - the top cell of the triangle. S_1 indicates the performance with the biggest area and that is the performance by A. Rubinstein (49% of the area).

S_2 calculations are presented in Fig. 14b. The winner S_1 from Fig. 14a is removed and the performance with the biggest area (S_2) in the newly generated scape plot is the performance by E. Gilels (81% of the area).

8.2 Ground truth for performing similarity assessments between performances

Determining similarities between performers is not an easy task for a human. On the one hand, specialist musical knowledge and experience are required of the evaluators. On the other hand, comparison of even short one-minute performances can cause much difficulty even for experienced musicians.

A similarity matrix form [5 × 5] to collect expert opinions, presented in Table 5 was used, which is a symmetric matrix, with the main diagonal values equal to 10. The experts' task was to determine which performances of a given performer are the most similar and which are the least. Each expert had to determine these values in the form for each composition. Each value described the degree of similarity between performances. The filled values were subsequent natural numbers in the range of [1, 9], where 9 meant a very similar

Table 5 Form for similarity matrix between performances

	Perf. 1	Perf. 2	Perf. 3	Perf. 4	Perf. 5
Perf. 1	10				
Perf. 2		10			
Perf. 3			10		
Perf. 4				10	
Perf. 5					10

performance, and 1 very different. A value of 10 on the diagonal meant that the given performance is maximally similar. Information on the performer of a given rendition was kept from the evaluators.

Three music experts with a university music education participated in the experiment. It took approximately 30-40 minutes to compare 5 one-minute performances of one composition. This included multiple times of listening to a performance as well as finding the appropriate numerical values on the similarity matrix, which was not a trivial task.

After collecting data from the experts, the data were ranked. This way we eliminated different individual opinions on the similarity scale and maintained the sequence of degree of similarity. On different questionnaires, the most similar performances could have different values - the maximum value on a given questionnaire - but after ranking the maximum values will always be in first place.

Finally, the obtained values were averaged and rescaled to the range of $[1, 9]$. Thus, a similarity matrix obtained from experts was built. It constituted the ground truth for similarities for a given composition and was used to compare with the matrix of similarity between performances obtained by a computer system.

To check the agreement between three experts' opinions, Cronbachs α (Cronbach 1951) and average Spearman's ρ was calculated (Table 6). Spearman's ρ was calculated between each set of two individual opinions, an then the obtained values ware averaged.

The calculated parameter values confirm that the collected data from the experts are correlated. The positive avg. Spearman's ρ values (from 0.42 to 0.54) indicate a clear relation between the experts' responses. With regard to internal consistency represented by Cronbachs α, the obtained values are good and acceptable.

8.3 Evaluation parameters

To assess the built system, several parameters comparing the obtained results with data obtained from music experts were used.

8.3.1 Spearman's rank correlation coefficient

The first evaluation parameter used was Spearman's rank correlation coefficient (Spearman's ρ) (Chen and Popovich 2002), which is the Pearson correlation coefficient between ranked variables. Before calculating correlations, variables are converted to ranks. The rank correlation coefficient ranges from 1 to -1. A positive ρ indicates a positive relationship between the two variables, while a negative ρ expresses a negative relationship.

In our case, Spearman's ρ measures how much the similarity values provided by the computer system and experts have a similar rank. While calculating Spearman's ρ between the similarity matrix obtained from experts and the similarity matrix obtained from the

Table 6 Compositions and agreement between 3 experts' opinions

Composition	Cronbachs α	Avg. Spearman's ρ
Prelude No.1	0.76	0.51
Prelude No.5	0.73	0.48
Prelude No.18	0.77	0.54
Prelude No.20	0.68	0.42

computer system, only elements below the main diagonal from the matrix were taken into account. Matrixes are symmetric; diagonal and upper diagonal elements are irrelevant. The greater the obtained Spearman's rank correlation coefficient, the closer the system's results were to the experts' opinions.

Spearman's ρ was calculated for the results between the experts' opinions and three similarity matrices obtained from the system: arousal similarity matrix ρ_A, valence similarity matrix ρ_V, and arousal-valence similarity matrix ρ_{AV}.

8.3.2 Maximal similar number of hits

The next parameters evaluated the concordance of the indicators on the similarity matrix obtained from the experts and the similarity matrix obtained from the system. Indicators of the most similar performers according to the experts and the system were compared. First, from among the experts' opinions the most similar performance to the reference performance was found, and then checked if it was confirmed by the system. If the indicators from both sides were convergent, we had a hit. The comparisons were performed for all reference performances, and the result was a percentage of hits - MSH (maximal similar hits) defined in (6) and (7).

$$MSH = \frac{\sum_{i=1}^{n} H_i}{n} \times 100\% \tag{6}$$

$$H_i = \begin{cases} 1 \text{ if } MS_i(EX) = MS_i(CS) \\ 0 \text{ if } MS_i(EX) \neq MS_i(CS) \end{cases} \tag{7}$$

where EX is the similarity matrix obtained from experts, CS is the similarity matrix obtained from the computer system, $MS_i()$ is the most similar performance to the reference performance i, and n is the number of performances.

Calculating MSH, we can compare the similarity matrix obtained from the experts to the similarity matrix obtained from the system on the basis of different indicators: S_0 or S_1 (Section 8.1).

To check if the searched most similar performance indicated by the experts is in the top indications by the computer system, a variant of the previous parameter - $MSH2F$ (maximal similar hits 2 first) was introduced. The $MSH2F$ calculation checks if the most similar performance according to the experts is among the top 2 indicated by the system. In the case of comparison with the results obtained on the scape plot, the first 2 most similar performances are indicated by S_1 and S_2 ($MSH2F_{S_1S_2}$).

8.3.3 Evaluation results

The obtained results of the evaluation are presented in Table 7. The first columns present Spearman's ρ calculated for the results between the experts' opinions and three similarity matrices obtained from the system: arousal-valence similarity matrix ρ_{AV}, arousal similarity matrix ρ_A, and valence similarity matrix ρ_V. The positive Spearman's ρ values (avg. ρ_{AV} = 0.57) indicate a clear relation and accordance with the experts' opinions and the computer system's calculations.

MSH and $MSH2F$ were calculated between the experts' opinions and the arousal-valence similarity matrix. Analyzing the indicators for the most similar performance according to the experts as well as the system, the average accuracy of the applied method was 50% when using score S_0, and 55% score S_1. However, the higher values of avg. $MSH2F_{S_0}$ and avg. $MSH2F_{S_1S_2}$ (85%) indicate that the results provided by the experts are in the top results obtained from the system.

Table 7 Evaluation parameters for the analyzed compositions

Composition	ρ_{AV}	ρ_A	ρ_V	MSH_{S_0} %	$MSH2F_{S_0}$ %	MSH_{S_1} %	$MSH2F_{S_1S_2}$ %
Prelude No.1	0.52	0.60	0.52	40	80	40	80
Prelude No.5	0.72	0.30	0.61	40	80	40	60
Prelude No.18	0.50	0.69	0.30	80	80	80	100
Prelude No.20	0.55	0.88	0.43	40	100	60	100
Averages	0.57	0.62	0.46	50	85	55	85

9 Similarities between pianists

In the next experiment, the author decided to test if the similarities between pianists remain the same for different compositions. An additional 7 Preludes by Frederic Chopin were added to the set of analyzed compositions (Section 6). All the analyzed Chopin performances were audio recordings played by the same 5 famous pianists: A. Rubinstein, E. Gilels, G. Sokolov, M. Argerich, and R. Blechacz.

Table 8 presents a list of the most similar performances to a given performer for the set of analyzed preludes, which were obtained from the calculated arousal-valence similarity matrices. The winner's initials in the table refer to the name of the pianist that is the most similar to the performance.

The obtained results confirm that there are no clear similarities between performances by different pianists using the example of various compositions. In other words, we cannot state that all performances by a given pianist are similar to only one other pianist. There is a great diversity here. This could be explained by the fact that if a pianist was completely emotionally similar in his/her performances to only one other pianist he/she would not be an artistic personality standing out from other pianists.

Table 8 List of the most similar performances by 5 pianists

	A. Rubinstein	E. Gilels	G. Sokolov	M. Argerich	R. Blechacz
Prelude No.1	G.S.	M.A.	R.B.	R.B.	G.S.
Prelude No.5	E.G.	R.B.	R.B.	R.B.	E.G.
Prelude No.10	M.A.	M.A.	R.B.	A.R.	A.R.
Prelude No.11	R.B.	R.B.	E.G.	A.R.	E.G.
Prelude No.12	M.A.	G.S.	E.G.	A.R.	A.R.
Prelude No.14	R.B.	G.S.	M.A.	G.S.	A.R.
Prelude No.16	G.S.	R.B.	M.A.	G.S.	M.A.
Prelude No.18	E.G.	M.A.	R.B.	E.G.	G.S.
Prelude No.19	R.B.	G.S.	E.G.	G.S.	A.R.
Prelude No.20	E.G.	M.A.	R.B.	E.G.	G.S.
Prelude No.23	E.G.	M.A.	M.A.	G.S.	E.G.
The most frequent perf.	E.G (4x)	M.A (5x)	R.B (4x)	G.S. (4x)	A.R (4x)

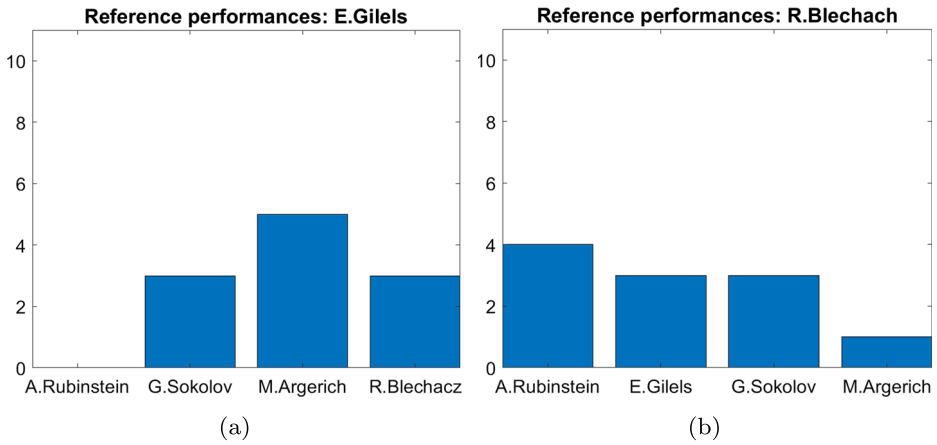

Fig. 15 Bar graph of the most similar performances to performances of E. Gilels (**a**) and to performances of R. Blechacz (**b**)

However, certain similarity tendencies can be noted in the most frequent occurrence of similarities. From the example bar graph of the most similar performances to the performances of E. Gilels (Fig. 15a), we can see that performances by E. Gilels are most often similar to performances by M. Argerich (5 x M.A.), and the least similar to A. Rubinstein (0 x A.R.). In the case of similarities to performances by R. Blechacz (Fig. 15b), we can note the most similarities to A. Rubinstein (4 x A.R) and the least to M. Argerich (1 x M.A). The presented results encourage undertaking further research to analyze similarities between the performances of various performers.

10 Conclusions

In this paper, the author attempted to answer the question if it is possible to find similar performances in terms of emotional content of the same composition. The presented method of comparative analysis of musical performances by using emotion tracking gave a positive response. Values of arousal and valence, predicted by regressors, were used to compare performances.

The author found which performances of the same composition were closer to each other and which were quite distant in terms of the shaping of arousal and valence over time. The applied approach comparing the obtained results with the opinions of music experts was evaluated. The obtained results confirm the validity of the assumption that tracking and analyzing the values of arousal and valence over time in different performances of the same composition can be used to indicate their similarities.

Comparison of performances by pianists over a greater set of compositions enables finding more detailed characteristics for a given performer. As it turns out, it is not so unequivocal, but rather complex, reflecting the pianist's artistic personality.

There can be multiple uses for the method of comparative analysis of musical performances using emotion tracking. Some of these include: searching databases in order to find similar or differing performances, analyzing and comparing pianists' performances during musical competitions, and supporting the learning process of young pianists.

Acknowledgements This research was realized as part of study no. S/WI/3/2013 and financed from Ministry of Science and Higher Education funds.

References

Aljanaki, A., Yang, Y.H., Soleymani, M. (2016). Emotion in music task: lessons learned. In *Working Notes Proceedings of the MediaEval 2016 Workshop*. Netherlands: Hilversum.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., Serra, X. (2013). ESSENTIA: an audio analysis library for music information retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference* (pp. 493–498), Curitiba.

Bresin, R., & Friberg, A. (2000). Emotional coloring of computer-controlled music performances. *Computer Music Journal*, *24*(4), 44–63.

Cannam, C., Landone, C., Sandler, M. (2010). Sonic visualiser: an open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference* (pp. 1467–1468), Firenze.

Chen, P.Y., & Popovich, P.M. (2002). *Correlation: parametric and nonparametric measures*. Sage: Thousand Oaks Calif.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Dixon, S., & Widmer, G. (2005). MATCH: a music alignment tool chest. In *ISMIR, 2005, 6th International Conference on Music Information Retrieval* (pp. 492–497). London: Proceedings.

Goebl, W., Pampalk, E., Widmer, G. (2004). Exploring expressive performance trajectories: six famous pianists play six chopin pieces. In *Proceedings of the 8th International Conference on Music Perception and Cognition (ICMPC'8)* (pp. 505–509), Evanston.

Gómez, E., & Bonada, J. (2005). Tonality visualization of polyphonic audio. In *Proceedings of the International Computer Music Conference*. Barcelona.

Grekow, J. (2012). Mood tracking of musical compositions. In Chen, L., Felfernig, A., Liu, J., Raś, Z.W. (Eds.) *Foundations of Intelligent Systems: 20th International Symposium, ISMIS 2012, Macau, China* (pp. 228–233). Berlin: Proceedings, Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34624-8_27.

Grekow, J. (2016). Computer Information Systems and Industrial Management: 15th IFIP TC8 International Conference, CISIM 2016, Vilnius, Lithuania. In Saeed, K., & Homenda, W. (Eds.) (pp. 697–706). Cham: Proceedings, Springer International Publishing. https://doi.org/10.1007/978-3-319-45378-1_60.

Grekow, J. (2017). Audio features dedicated to the detection of arousal and valence in music recordings. In *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 40–44), IEEE. https://doi.org/10.1109/INISTA.2017.8001129.

Korhonen, M.D., Clausi, D.A., Jernigan, M.E. (2005). Modeling emotional content of music using system identification. *Transactions on Systems Man and Cybernetics Part B*, *36*(3), 588–599.

Liem, C.C.S., & Hanjalic, A. (2015). Comparative analysis of orchestral performance recordings: an image-based approach. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015* (pp. 302–308), Málaga, Spain.

Lu, L., Liu, D., Zhang, H.J. (2006). Automatic mood detection and tracking of music audio signals. Trans Audio. *Speech and Language Proceedings*, *14*(1), 5–18.

Müller, M., & Jiang, N. (2012). A scape plot representation for visualizing repetitive structures of music recordings. In *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012* (pp. 97–102), Mosteiro S.Bento Da Vitória, Porto, Portugal.

Rabiner, L., & Juang, B.H. (1993). *Fundamentals of Speech Recognition*. Upper Saddle River: Prentice-Hall, Inc.

Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Sapp, C.S. (2001). Harmonic visualizations of tonal music. In *Proceedings of the 2001 International Computer Music Conference, ICMC 2001*, Havana, Cuba.

Sapp, C.S. (2007). Comparative analysis of multiple musical performances. In *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007* (pp. 497–500), Vienna, Austria.

Sapp, C.S. (2008). Hybrid numeric/rank similarity metrics for musical performance analysis. In *ISMIR 2008, 9th International Conference on Music Information Retrieval* (pp. 501–506). Philadelphia: Drexel University.

Schmidt, E.M., Turnbull, D., Kim, Y.E. (2010). Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the International Conference on Multimedia Information Retrieval, MIR '10* (pp. 267–274). New York: ACM.

Segnini, R., & Sapp, C. (2006). Scoregram: Displaying Gross Timbre Information from a Score (pp. 54–59). Berlin: Springer Berlin Heidelberg.

Widmer, G., & Goebl, W. (2004). Computational models of expressive music performance: the state of the art. *Journal of New Music Research*, *33*(3), 203–216.

Witten, I.H., & Frank, E. (2005). *Data Mining: practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.

Yang, Y.H., Lin, Y.C., Su, Y.F., Chen, H.H. (2008). A regression approach to music emotion recognition. Trans Audio. *Speech and Language Proceedings*, *16*(2), 448–457.