Predicting the similarity between expressive performances of music from measurements of tempo and dynamics

Renee Timmers[a)]
Austrian Research Institute for Artificial Intelligence, Freyung 6/VI, 1010 Vienna, Austria

(Received 27 October 2003; revised 19 October 2004; accepted 26 October 2004)

Measurements of tempo and dynamics from audio files or MIDI data are frequently used to get insight into a performer's contribution to music. The measured variations in tempo and dynamics are often represented in different formats by different authors. Few systematic comparisons have been made between these representations. Moreover, it is unknown what data representation comes closest to subjective perception. The reported study tests the perceptual validity of existing data representations by comparing their ability to explain the subjective similarity between pairs of performances. In two experiments, 40 participants rated the similarity between performances of a Chopin prelude and a Mozart sonata. Models based on different representations of the tempo and dynamics of the performances were fitted to these similarity ratings. The results favor other data representations of performances than generally used, and imply that comparisons between performances are made perceptually in a different way than often assumed. For example, the best fit was obtained with models based on absolute tempo and absolute tempo times loudness, while conventional models based on normalized variations, or on correlations between tempo profiles and loudness profiles, did not explain the similarity ratings well. [DOI: 10.1121/1.1835504]

PACS numbers: 43.75.St, 43.75.Cd [SEM] Pages: 391–399

I. INTRODUCTION

Measurements of performances, with an emphasis on piano performances, have become an important means of understanding musical expression. For example, measurements have shown the extensive use of subtle variations in tempo, timing, articulation, and dynamics, as well as the consistency of such variations in repeated performances (both already observed by Seashore, 1938), and the controllability of the variations (e.g., Kendall and Carterette, 1990; Palmer 1989, 1996). Comparisons between variations have given some insights into the diversity of interpretations. Repp (1990, 1992a) found, for example, that the diversity tends to be greater among professional musicians than among piano students, and that the diversity tends to be smaller at the phrase level than below the phrase level.

Though measurements provide a detailed account of what is physically happening in a performance, the concern central to this paper is to what extent they reflect the psychological reality of performers and listeners. The validity for performers and listeners of measurable variations has been suggested by several studies. For example, in the work of Sundberg and colleagues (Sundberg, Friberg, and Frydén, 1989; Friberg *et al.*, 1991), the remarks of a professional musician on how to improve a synthesized performance without expressive variations were translated into concrete formulations of rules for the variations of tempo, timing, dynamics, and intonation. In consequent studies, the quality of the synthesized performances with variations was judged to be high compared to the quality of performances without variations (see Thompson *et al.*, 1989).

However, several studies have also shown that the perception of time intervals does not directly correspond to their physical properties. For example, Repp (1992b, 1998) demonstrated a dependency of the perceptual length of a timing perturbation on its position within the phrase structure of music, and Nakajima (1987) found a systematic overestimation of short, empty duration intervals by a constant interval. This means that the overestimation is relatively large for shorter durations.

An additional problem: it is unknown which of the many quantitative representations of tempo and dynamics comes closest to the perceptual representation. The variety of representations includes the representation of timing as either duration or tempo variation (Friberg and Sundberg, 1999); the use of different time scales such as the note level, beat level, bar level, or phrase level (Bengtsson and Gabrielsson, 1983); the use of normalization, which means that the variations are expressed relative to the mean (see, e.g., Gabrielsson, 1987, 1988) instead of in absolute values such as milliseconds or beats per minute (see Repp, 1992a, 1992b; Langner and Goebl, 2003); the use of derivatives of tempo and dynamics rather than absolute values in order to represent the control of tempo change (Kronman and Sundberg, 1987), or the perception of the ''change of change'' in dynamics (Gjerdingen, 1988). Exploration of some of these representations for tempo showed a considerable effect of the time scale and the representation unit on the characteristics of the measured data (Timmers and Honing, 2002).

The main aim of the reported study is to test how well measured data represent perceptually salient characteristics of performances and what data representation comes closest

[a)]Current affiliation: Department of Music, King's College, London, UK. Electronic mail: renee.timmers@kcl.ac.uk

to perception. In comparing representations, it focuses on the size of the time span of the representation (local versus global), the unit of the representation (absolute versus normalized), and the appropriate derivative for the representation (i.e., absolute, first, or second derivative). In addition, the relative salience of tempo and dynamics is tested, as well as the validity of a compound measure that consists of the interaction between tempo and loudness. Tempo times loudness can be seen as a measure of integrated energy (cf. Todd, 1992; Zanon and Widmer, 2003). No comparison is made between the validity of tempo and that of duration. Instead, only tempo is used.

Thus this study addresses the questions of (1) whether the perception of performance deals more with global or local features; (2) whether variations are perceived as changes in absolute values or in relative values (i.e., relative to the average); (3) whether listeners pay attention to absolute levels or to changes therein, or even to changes within the variations; and finally, it addresses (4) whether listeners perceive tempo and loudness as separate dimensions or integrate the two into one compound feature.

The validity of the different data representations is tested by comparing their ability to explain the subjective similarity between pairs of performances. This is done in two experiments that have the same aim and musical material, but differ in experimental procedure. Experiment 2 is a replication of experiment 1 with a stricter experimental procedure. In both experiments, 20 participants listen to pairs of performances of a Chopin prelude and a Mozart sonata and rate the similarity between the performances. These performances are fragments from CD recordings of famous pianists. The tempo and loudness of the performances are measured at the beat level from the audio recordings. The distance in tempo and loudness between pairs of performances is then calculated using the different representations of tempo and loudness. Finally, these distance measures are input to separate multiple regression analyses and to stepwise regression analyses in an attempt to explain the similarity ratings. The degree to which each representation accounts for the variance in the similarity ratings is interpreted as a measure of its ability to capture salient characteristics of the performances.

II. METHOD EXPERIMENT 1

A. Musical material

Five performances of Chopin's Prelude Op. 28, No. 17 are used, as well as six performances of the first movement of Mozart's Sonata KV281. The five performances of the Chopin prelude are by Argerich, Harasiewicz, Kissin, Pollini, and Rubinstein (to be referred to as p1, p2, p3, p4, and p5, respectively)[1] and the six performances of the Mozart sonata are by Barenboim, Batik, Gould, Pires, Schiff, and Uchida (to be referred to as p1, p2, p3, p4, p5, and p6, respectively).[2] The opening bars of the two pieces are used in the experiment (mm. 1–10 for Chopin, and mm. 1–4 for Mozart) as well as six bars from the development section of the Mozart sonata (mm. 22–27 with upbeat). To refer to these fragments, the abbreviations Ch, M1, and M2 are used. These three

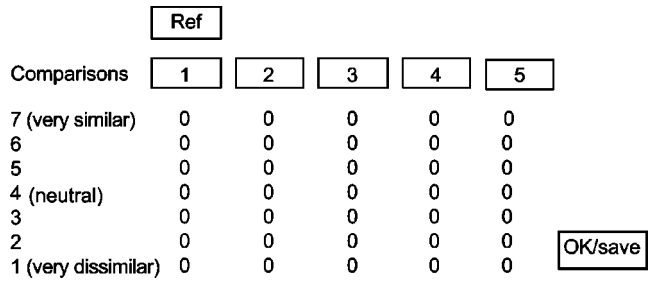

FIG. 1. User interface for the similarity rating in experiment 1.

fragments were chosen because they are expected to differ in the degree to which tempo and dynamics play a role. The importance of these parameters may be especially large in the Chopin prelude, which consists mainly of chords in a repeated eighth-note rhythm. It may be less important in the opening bars of the Mozart sonata, which contains other expressive features such as ornaments and arpeggios. It may again be important in the bars from the development section of the Mozart sonata, which contains leaps in sixteenth notes.

B. Participants

Seven women and 13 men participated in experiment 1. Their age varied from 26 to 45. Fifteen participants were experienced musicians who had had 10 or more years of musical training. Five participants were nonmusicians, who had no more than 3 years of instrumental lessons. Among the musicians were six pianists and nine nonpianists.

C. Procedure

The participants were tested on an individual basis. Half of the participants first listened to the Chopin performances and then to the Mozart performances, while the order was reversed for the other half. The order of the Mozart fragments was always M1 followed by M2.

The presentation of performance pairs was semirandom. To facilitate the similarity judgments, the presentation of performance pairs was grouped into blocks that contained one reference performance and four or five comparison performances. For the Chopin prelude, the participants made four comparisons per block, since the total number of performances was five. The fifth performance was the reference performance. For the Mozart sonata, the participants made five comparisons per block, since the total number of performances was six. Figure 1 shows the organization of one block for the Chopin prelude. In this way, the participant was confronted with all performances basically at once, which provided a stable frame of reference for the similarity ratings. In each subsequent block, a different performance became the reference. The order of these references was randomized over participants as well as the order of the comparison performances within a block.

The participants sat in front of a Macintosh iBook computer and saw the user interface depicted in Fig. 1 on the screen. The interface contained play buttons for the reference performance and the comparison performances. They listened alternatively to the reference performance and a comparison performance via headphones, and rated the similarity

between the two on a scale from 1 to 7 by pressing one of the radio buttons. One meant very dissimilar, while 7 meant very similar. They could listen to each performance as often as they wished and could correct the ratings until they pressed the ok/save button. This would bring up the following block of performances, which consisted of the same performances in a different order of comparison and with a different reference performance. The session ended when all performances had been the reference performance once. This resulted in 20 comparisons for the Chopin fragment and 30 comparisons for the Mozart fragments. Each of the 10 performance pairs of the Chopin fragment and 15 performance pairs of the Mozart fragments were rated twice: once with one of the performances as reference and the once with the other performance as reference.

After the experiment, the participants filled out a questionnaire about their rating strategy. They were asked to describe on what bases they made the similarity judgments and to what aspects of the performances they paid most attention. This last question was answered by giving a rating to a list of parameters on a scale from 0–3. Zero meant no attention, while 3 meant most attention. The parameters are listed in Table III. The total duration of the experiment was around 1 h.

D. Apparatus

The experimental data were collected using POCO (Honing, 1990), running on an Apple iBook under Macintosh OS 9.2. A special POCO module was designed containing a user interface (see Fig. 1), playback of audio files, and recording the responses into a log file. The sound files used in the judgments were CD-quality stereo audio files (sampled at 44.1 kHz). Sony's dynamic stereo professional headphones MDR-7506 were used.

E. Similarity predictions

Central to this study is the prediction of the subjective judgment of similarity between performances on the basis of different representations of their tempo and dynamics. The measurement of tempo and loudness from the audio recordings was performed using algorithms developed by members of the Music and AI group at the Austrian Research Institute for Artificial Intelligence. A beat-tracking algorithm was used to locate the beat within the audio file (see Dixon, 2001). The output of the beat-tracking procedure was hand-corrected using an interface especially designed for this purpose (Dixon and Goebl, 2001). The location of the beat was defined to coincide with the onset of the corresponding melody note. Local tempo in beats per minute was calculated for each interbeat interval. To get a measure of local loudness, the localized beats were used as well. The maximum amplitude level was selected around each beat, which spans from the halfway point of the previous interbeat interval to the halfway point of the following interbeat interval (see also Langner and Goebl, 2003). This level in dB was recalculated into sones, which is an approximation of the subjective perception of the loudness of tones (see Pampalk *et al.*, 2002).

As mentioned in the Introduction, the different representations of tempo and dynamics to be compared vary in time scale (local or global), unit (absolute or normalized), and derivation (absolute or derivative). To predict the subjective distance between two performances, the difference in tempo and loudness between two performances was calculated using different methods for different representations. Equations (1)–(9) give the calculation method for each representation. For brevity, the calculations are only presented for tempo. Similar calculations were applied to the loudness of each beat. Capital T is a vector that consists of a tempo indication for each interbeat interval ($T=[t_1,t_2,\ldots,t_n]$, in which t stands for local tempo at interbeat interval n).

In the equations below, the subscript of T refers to one of the two performances of a performance pair that are compared. The horizontal line above an expression indicates averaging. The vertical lines at both sides of an expression indicate that the absolute value is taken. The superscripts in change, and change of change, indicate the first and second derivative, respectively.

More specifically, the first distinction is between a global and a local representation. The global representation of tempo calculates the average tempo of a performance. The distance in tempo between two performances is then calculated by taking the absolute difference in average tempo of performances 1 and 2; see Eq. (1)

$$\text{Difference in global tempo} \quad |\overline{T_1}-\overline{T_2}|. \tag{1}$$

The distance in local tempo between two performances is calculated by calculating the absolute difference in local tempo for each beat of the two performances and taking the average of these absolute differences; see Eq. (2)

$$\text{Difference in local tempo} \quad |\overline{T_1-T_2}|. \tag{2}$$

The second distinction is between absolute tempo and loudness [such as in Eqs. (1) and (2)], and relative tempo and loudness. Relative tempo (or loudness) is calculated by dividing the absolute value by the average tempo (or loudness) of the performance. The difference in relative tempo variation is given by Eq. (3)

$$\text{Difference in relative tempo variation} \quad \left|\overline{\frac{T_1}{\overline{T_1}}-\frac{T_2}{\overline{T_2}}}\right|. \tag{3}$$

The third distinction is between absolute tempo and loudness and the changes within the tempo and loudness values. The derivative captures the change in tempo (or loudness) over time. In other words, it measures the amount of acceleration and deceleration, and the amount of crescendo and decrescendo. The second derivative captures the change within the change in tempo (or loudness) over time. This means that it measures changes in the direction and amount of acceleration/deceleration, and crescendo/decrescendo. The distance between two performances based on differences in tempo change and the change of tempo change is given by Eqs. (4) and (5), respectively

$$\text{Difference in tempo change} \quad |\overline{T_1'-T_2'}|, \tag{4}$$

Difference in change of tempo change $|\overline{T_1'' - T_2''}|$. (5)

The derivative is defined as the difference in local tempo (or loudness) between successive beats, so T' is $[t_2 - t_1, t_3 - t_2, \cdots, t_n - t_{n-1}]$, in which t stands for local tempo at the nth interbeat interval. In this way, the calculation is not strictly a derivative over time, but an event-based calculation of change.

In the fourth distinction, tempo and dynamics are either treated as separate variables as above in Eqs. (1)–(5), or treated as a compound variable as in Eq. (6) and Eq. (7). The compound variable integrates tempo and loudness per beat through multiplication. The difference in this measure between two performances is calculated on a global level using Eq. (6) and a local level using Eq. (7)

Difference in global tempo times loudness

$$|\overline{T_1^* L_1} - \overline{T_2^* L_2}|, \tag{6}$$

Difference in local tempo times loudness

$$|\overline{T_1^* L_1 - T_2^* L_2}|. \tag{7}$$

Finally, two measures that are regularly used in performance research are added to be more complete. The first is the standard deviation of the variation (see, e.g., Timmers *et al.*, 2000; Timmers, 2003; and Zanon and Widmer, 2003). The distance between two performances in tempo and dynamics is with this measure assumed to be the absolute difference between the standard deviation of local tempo or loudness of performance 1 and 2; see Eq. (8)

Difference in amount of tempo variation

$$|\mathrm{std}(T_1) - \mathrm{std}(T_2)|. \tag{8}$$

The final measure calculates the correlation between the local tempo (and loudness) profile of the two performances; see Eq. (9). This is the most frequently used method to assess the similarity between the timing and dynamics profiles of different performances (Clarke, 1993; Repp, 1994, 2000; Timmers *et al.*, 2000)

Correlation between tempo profiles $\mathrm{corr}(T_1, T_2)$. (9)

To compare the predictive power of the different representations, two methods are used. The first approach compares the predictive power of the nine measures by running separate multiple regression analyses for each measure. All of these models consist of one tempo and one loudness component, except for the models based on the compound measures as in Eq. (5) and Eq. (6), which have only one component. The regression models with two components have the format shown in Eq. (10). The regression model consists of two components (the difference measures D_t and D_l), an intercept (a) and weights (b and c). The regression models for the compound measures have only one component (D_{t*l}), an intercept (a), and one weight (b)

$$s = a + b*D_t + c*D_l. \tag{10}$$

The second approach takes all difference measures based on tempo, loudness, and the interaction between them as input of a stepwise regression analysis. The stepwise regression analysis adds components to the analysis in order of explained variance. It adds components as long as their addition to the explained variance is significant. An additional restriction is that only components are included for which the effect is in the predicted direction: an increase in difference (and a decrease in correlation) should lead to a decrease in similarity rating. In this way, the components are sorted in order of explained variance. These analyses were done using JMP 4.0.

III. RESULTS EXPERIMENT 1

The presentation of the results is divided into three parts: First, the results of the similarity ratings are presented, followed by the results of the prediction of the similarity ratings by the different models and the stepwise regression analysis. Third, the results of the questionnaire are presented and related to the results of the similarity rating study.

A. Similarity ratings

To test if the effect of performance pair is systematic over participants and presentations, a repeated measures ANOVA was run in SPSS10 with pair and order as independent within-subject variables and the similarity rating as dependent variable. A separate ANOVA was run for each fragment.

For M1 and Ch, the main effect of pair is the only significant effect. This is also the case when the analysis is corrected for violations of sphericity using the Greenhouse–Geisser correction [$F(9,11) = 25.7$, $p < 0.001$ for Ch; $F(14,6) = 19.4$, $p < 0.001$ for M1].[3] For M2, however, all effects are significant using the same correction for violations of sphericity. The effect of pair is the strongest effect, followed by the main effect of order [$F(14,6) = 21.7$, $p < 0.001$ for the main effect of pair, $F(1,19) = 10.8$, $p = 0.004$ for the main effect of order, and $F(14,6) = 3.2$, $p = 0.004$ for the interaction effect]. The main effect of order is hard to explain. Like the interaction between pair and order, it might suggest that for some similarity ratings of M2 the similarity rating depended on which of the performances was the reference performance. The blocking of stimuli may have caused this context effect. In experiment 2, this issue is resolved by presenting all performance pairs sequentially.

Although the interaction between pair and order was significant for M2, the size of the effect was rather small. In fact, the average difference in ratings of a pair in the two orders remained under 1.15 points (pair 4–5 of Ch). To get a robust similarity rating that is independent of context, the following analyses use the average of the ratings of pairs in the two orders. It therefore has 10 data points per participant for Ch, and 15 data points per participant for M1 and M2.

B. Prediction of similarity ratings

The first comparison between the explanatory power of the different representations is done by fitting the nine measures to the similarity ratings using separate multiple regres-

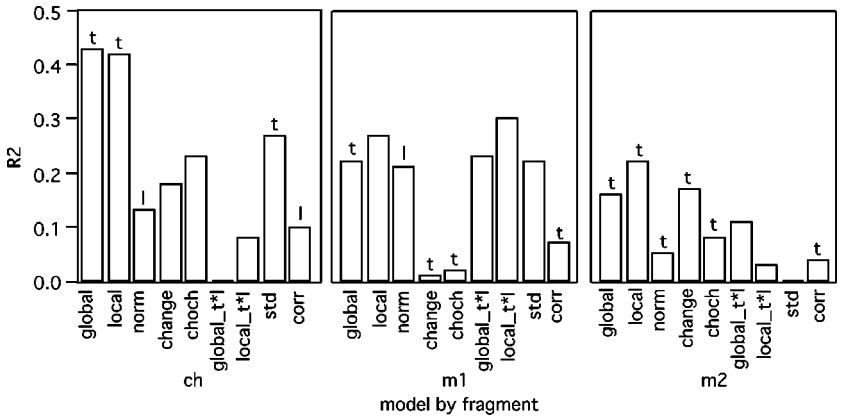

FIG. 2. Explained variance (R^2) of the fits of each model to the similarity ratings of all participants, separated per fragment for experiment 1. The models included are based on global, local, and normalized (norm) tempo and loudness; change and change of change (choch) of tempo and loudness; global and local tempo times loudness (t^*l), the standard deviation of (std) and the correlation (corr) between tempo and loudness profiles. Letters indicate the component that accounted for 2/3 or more of the explained variance (t for tempo and l for loudness).

sion models for each measure. All of these models consist of one tempo and one loudness component, except for the models based on the compound measures that have only one component. Figure 2 shows the explained variance for each of the regression models. The letter on top of each line indicates which parameter (t for tempo and l for loudness) contributes most to the explanation, which means that it explains at least 2/3 of the total explained variance. If no letter is indicated, both tempo and loudness contribute roughly equally to the explanation. Global and local tempo times loudness do not have a letter, because they always include both variables.

Figure 2 shows that the explained variance is higher for the Chopin fragment than for the Mozart fragments, and that tempo is the more important factor for the explanation. It further shows that (1) the regression analysis based on local tempo and loudness does better than based on global tempo and loudness for the Mozart fragments. (2) To take relative values instead of absolute values does not improve the explained variance. This is suggested by the low explained variance of the regression analysis that uses normalization (abbreviation norm). (3) To take the derivative of tempo and loudness is not an improvement, given the low explained variance for models based on change and change of change (abbreviation choch). (4) To treat tempo and loudness as separate values is better than to treat them as one compound variable for Ch and M2, but not for M1. (5) The regression analysis based on the standard deviation of tempo and loudness does quite well for Ch and M1, but is not best. (6) Correlation does not explain the similarity ratings well.

The second comparison between the relative strength of the different representations takes all separate difference measures based on tempo and loudness as potential input of a stepwise regression model. This includes the measures given by Eqs. (1)–(9) and the parallel measures for loudness. The components are added stepwise in order of explained variance as long as the addition in explained variance is significant ($p<0.05$) and the direction of the effect of the component is as predicted, which means that an increase in the measured difference between two performances leads to a decrease in the similarity rating.

By using stepwise regression, the components are sorted in order of explained variance. The benefit of this method is that there is no overlap in explained variance, and the focus is only on those variables that are most responsible for the explanation of variance. This is in contrast to the previous presentation of the results, in which part of the explained variance by one model might be due to the correlation with another model.

Tables I and II show the results of the stepwise regression analyses for each fragment for the musicians and nonmusicians separately. The results for the musicians and nonmusicians are highly similar: Local or global tempo is the strongest component for Ch and M2, and global or local tempo times loudness is second, although not for the musicians' ratings of Ch. The reverse is true for M1; the main component is local tempo times loudness, while the second component for the musicians is local tempo. The components are more often local than global. The explained variance is larger for the Chopin fragment than for the Mozart fragments. It is larger for the nonmusicians than for the musicians.

Note that some of the models that did quite well in the separate multiple regression analyses do not occur in the stepwise regression analyses; they did not account for vari-

TABLE I. Results of the stepwise regression analysis for musicians ($N=15$). Parameters in order of entrance of the stepwise regression model; the total explained variances for each step, and the F and p value for the full model.

Fragment	Parameters	R^2	F value	p value
Ch	Global t	0.45	$F(1,148)=120$	<0.0001
M1	Local t^*l	0.27	$F(2,222)=46.3$	<0.0001
	Local t	0.29		
M2	Local t	0.23	$F(3,221)=43.4$	<0.0001
	Global t^*l	0.35		
	Change l	0.37		

TABLE II. Results of the stepwise regression analysis for nonmusicians ($N=5$). Parameters in order of entrance of the stepwise regression model; the total explained variances for each step, and the F and p value for the full model.

Fragment	Parameters	R^2	F value	p value
Ch	Local t	0.52	$F(2,47)=30.0$	<0.0001
	Global t^*l	0.56		
M1	Local t^*l	0.40	$F(1,73)=47.8$	<0.0001
M2	Local t	0.22	$F(2,72)=17.3$	<0.0001
	Global t^*l	0.32		

TABLE III. Sum of attention ratings for the Chopin and Mozart fragments expressed as percentage of the maximal sum of the ratings.

Parameter	Chopin (% of max)	Mozart (% of max)
Tempo	71	78
Loudness	36	42
Rubato	73	73
Dynamics	62	64
Articulation	56	78
Pedal	40	33
Phrasing	78	71
Interpretation	76	76
Character	76	78
Emotion	49	44

ance in addition to the variance accounted for by the strongest components. Note as well that change in loudness makes a small contribution to the explained variance of the musicians' ratings of M2.

C. Interviews

After the similarity rating experiment, the participants were asked to indicate to what aspects of the performances they had paid most attention. This was first done by free choice and secondly by giving an attention rating to ten variables (see Table III).

The aspects that were mentioned and the number of participants mentioning them in response to the free-choice question are tempo (11), articulation (7), character and style/overall impression (7), interpretation (6), rubato (5), dynamics (4), the quality of the pianist/the smoothness of playing (4), loudness (2), arpeggios, ornaments (2), the sound of the recording (1), phrasing (1), and perception of movement (1). The attention ratings show a similar pattern (Table III): most attention is paid to global tempo and rubato, less attention is paid to dynamics, and little to overall loudness. Articulation is an important factor for the Mozart fragments, but less so for the Chopin fragment, which may have been due to the larger use of pedal in the Chopin fragment. The interpretation of the music is important as well as the character and style of the performance.

The importance of tempo and tempo variation and the lesser importance of dynamic variation and overall loudness agree with the results of the experiment. The importance of articulation for the Mozart fragments may account for the lower explained variance for the Mozart fragments than for the Chopin fragment. The tendency of several participants to listen to the overall impression of a performance may indicate that they did not listen very analytically, which possibly may relate to the large contribution of the tempo times loudness measures. Differences in interpretation of the music such as phrasing may have been accounted for indirectly, though probably taking the relationship with musical structure into account would have improved the variance explained.

IV. EXPERIMENT 2

A second experiment was run with the same purpose as the first experiment. The only difference with the first experiment is the experimental procedure, which was changed to be in accordance with the general procedure for similarity rating studies. Instead of blocking stimuli into groups of comparison performances with a reference performance, the performance pairs were presented one after another and the rating was done for each pair sequentially, to avoid possible dependencies between the ratings of different pairs. The results of experiment 2 are not expected to be different from the results of experiment 1. Instead, experiment 2 is a replication of experiment 1 with a stricter experimental procedure.

V. METHOD OF EXPERIMENT 2

A. Musical material

The musical material was the same as in experiment 1, with the exception that pairs of performances were in experiment 2 combined into one audio file with a 2200-ms interval between the end of the first performance and the start of the second performance. A practice trial was added that used four performances of the second movement of Mozart KV332 piano sonata by Argerich, Gould, Pires and Schiff. These performances were taken from the same CDs as used in experiment 1.

B. Participants

Eleven women and nine men participated in experiment 2. Their age varied between 21 and 48. Fifteen participants were experienced musicians who had had 10 or more years of musical training. Five participants were nonmusicians, who had had no more than 3 years of instrumental lessons. Among the musicians were nine pianists and six nonpianists.

C. Procedure

The participants were tested on an individual basis. They sat in front of a Macintosh iBook computer and saw a simplified interface on the screen. The interface had one play button, one set of seven vertically aligned radio buttons to make the similarity rating, and an ok/save button. The participants read the instructions from paper. The instructions described the task of the participants and the procedure of the experiment. They were asked to listen to a pair of performances that would sound by pressing the play button, and to indicate the similarity between the performances on a scale from 1–7. One means that the performances are very different, while 7 means that they are very similar. The values in between could be used for fine tuning. After the rating of similarity, they pressed the ok/save button to continue with the next pair of performances.

The entire experiment consisted of four blocks: a block to practice and three blocks to rate the similarity between pairs of performances for each of the three fragments. Each participant rated all pairs of performances of a fragment once. The order of presentation of fragments and performance pair was randomized over subjects. The order of the

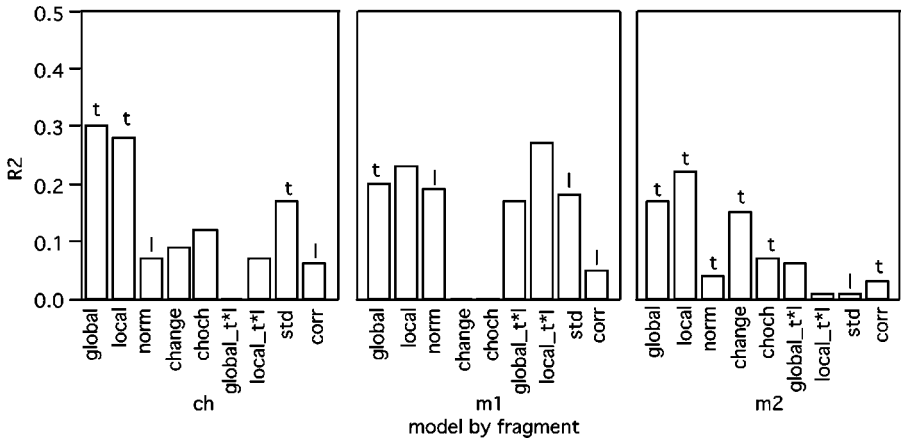

FIG. 3. Explained variance (R^2) of the fits of each model to the similarity ratings of all participants, separated per fragment for experiment 2. The models included are based on global, local, and normalized (norm) tempo and loudness; change and change of change (choch) of tempo and loudness; global and local tempo times loudness ($t*l$), the standard deviation of (std) and the correlation (corr) between tempo and loudness profiles. Letters indicate the component that accounted for 2/3 or more of the explained variance (t for tempo and l for loudness).

performances within a pair was counterbalanced over participants. The total duration of the experiment was around half an hour.

D. Apparatus

The same apparatus was used as in experiment 1 with the exception of the interface, which was simplified. The interface included only one play button and one rating scale with seven radio buttons.

E. Similarity predictions

The same analyses were done with the same predictions for the similarity ratings as in experiment 1.

VI. RESULTS EXPERIMENT 2

A. Similarity ratings

To test if the effect of performance pair was systematic over participants and order, a repeated measures ANOVA was run in SPSS10 with pair as within-subject variable, order as between-subjects variable, and the similarity rating as dependent variable. A separate ANOVA was run for each fragment.

For all three fragments, the main effect of pair was the only significant effect. This was also the case when the analysis was corrected for violations of sphericity using the Greenhouse–Geisser correction [$F(9,11)=13.5$, $p<0.001$ for Ch; $F(14,6)=11.1$, $p<0.001$ for M1; and $F(14,6)=11.3$, $p<0.001$ for M2].

B. Prediction of similarity ratings

As for experiment 1, the first comparison between the explanatory power of the different representations was made by fitting the nine measures to the similarity ratings using separate multiple regression models for each measure. Figure 3 shows the explained variance for each of the regression models.

Although the explained variance is a bit lower than in experiment 1, the results are highly similar overall. The models that do well and the ones that do not do well are the same as in experiment 1. Indeed, the correlation between the explained variances of the models as found in the two experiments is above 0.97 for each of the three fragments.

The results of the stepwise regression analyses show a similar high agreement with the results of experiment 1, especially for the musicians (see Tables IV and V). Again, local or global tempo, and local or global tempo times loudness are the two strongest components. In addition, the change of change or the correlation in loudness explains a small part of the variance for M2.

Experiment 2 generally replicates the results of experiment 1, with the exception that the explained variance was a bit higher in experiment 1 than in experiment 2. This difference can be attributed to the procedure of experiment 1, which provided the participants with a frame of reference for the similarity ratings by presenting all performances within each rating block. Besides this effect in quantity, there was no qualitative effect of the difference in procedure or difference in participant pool, which strengthens the generality of the results.

TABLE IV. Results of the stepwise regression analysis for musicians ($N=15$) for experiment 2. Parameters in order of entrance of the stepwise regression model; the total explained variances for each step, and the F and p value for the full model.

Fragment	Parameters	R^2	F value	p value
Ch	Global t	0.42	$F(1,148)=106$	<0.0001
M1	Local $t*l$	0.33	$F(2,222)=58.2$	<0.0001
	Local t	0.34		
M2	Local t	0.25	$F(3,221)=33.4$	<0.0001
	Global $t*l$	0.30		
	Change l	0.32		

TABLE V. Results of the stepwise regression analysis for nonmusicians ($N=5$) for experiment 2. Parameters in order of entrance of the stepwise regression model; the total explained variances for each step, and the F and p value for the full model.

Fragment	Parameters	R^2	F value	p value
Ch	Local $t*l$	0.11	$F(2,47)=4.8$	<0.02
	Global t	0.17		
M1	Global $t*l$	0.20	$F(1,73)=18.4$	<0.0001
M2	Local t	0.14	$F(3,71)=11.5$	<0.0001
	Global $t*l$	0.26		
	Corr l	0.33		

VII. GENERAL DISCUSSION

As outlined in the Introduction, the reported study addressed the questions of (1) whether the perception of performance deals more with the global or the local features; (2) whether variations are perceived as changes in absolute values or in relative values (i.e., relative to the average); (3) whether listeners pay attention to absolute levels or to changes therein, or even to changes within the variations; and (4) whether listeners perceive tempo and loudness as separate dimensions or integrate the two into one compound feature. The results of the two experiments show local models to be slightly stronger than global models, and models based on absolute values to be stronger than models based on normalized values. Models based on absolute values were also stronger than those based on derivatives. This strength of absolute representations concerned absolute tempo as well as the compound feature of tempo times loudness. Loudness was only sporadically significant as a separate feature, and when it was significant, it was in other representations than absolute loudness.

In addition, this study showed the limited ability of correlation to capture the similarity in tempo and dynamics of two performances as well as the medium strength of models based on the standard deviation of tempo and dynamics to do so.

The strength of the parameters changed with fragment in a similar way for the musicians and the nonmusicians and in the same way in the two experiments. This strongly suggests bottom-up processes driven by the specific characteristics of the musical stimuli.

Parts of these results are confirmations of previous literature, while other parts were less expected. For example, Repp (2000) also mentioned the limited ability of correlations to capture differences between patterns related to the means and the standard deviations of the variations. The current study stresses the importance of such aspects not covered by correlation as the extent of the variation and the absolute value of the measure. It demonstrated the salience of global tempo for the evaluation of performances, which is in line with studies on the emotion of music for which tempo is an important factor (e.g., Hevner, 1937) and studies on the reproduction of the absolute tempo of memorized music (Levitin and Cook, 1996). Nevertheless, the large role of local tempo in the similarity predictions opens an unexplored area of investigation. It implies that even in a comparison between performances, the absolute tempo of a time unit of one performer is compared to that of the other performer rather than the interpretation of the passage in terms of acceleration or deceleration. Similarly, the representation of performance variables as an integration of tempo and loudness has hardly been explored, although it is prominent in Todd's theory of expression (e.g., 1992). It seems important for future research to further investigate the relevance of these representations.

This study is not conclusive about the absolute strength of the representations described here. It only provides a relative ranking of the different measures. Probably the extraction of salient performance characteristics can still be improved upon and a more complete model might be defined to explain the distance between a pair of performances. The interview held with the participants highlighted some of the aspects that were missed by the models. For example, the models did not include articulation, timbre, fluency, or quality of the performance, and did not take the relation with the musical structure into account. In addition, it might have been that participants focused their attention on specific parts of the music rather than the entire fragment and that primacy and recency effects played a role.

Nevertheless, the measured differences in tempo and loudness were quite well able to predict the subjective distance between performances, and seem therefore reliable to represent a considerable part of the performance characteristics. The parameters most responsible for this explanation for both musicians and nonmusicians were local and global tempo, and local and global tempo times loudness. Local tempo and the interaction between tempo and loudness are not often used in performance research, and a shift in attention towards these representations of performances seems important for future research, also when comparing between different interpretations of music.

ACKNOWLEDGMENTS

This study was realized with financial support of the Mosart IHP-Network, HPRN-CT-2000-00115, the START program of the Austrian Federal Ministry for Education, Science and Culture (Grant No. Y99-INF), and a Talent stipendium of the Dutch Scientific Organization (NWO). In addition, the Austrian Research Institute for Artificial Intelligence acknowledges basic financial support by the Austrian Federal Ministry for Education, Science and Culture. I would like to thank Ric Ashley, Simon Dixon, Werner Goebl, Josef Linschinger, Elias Pampalk, Asmir Tobudic, and Gerhard Widmer for providing the performance data and the perfect environment to do this study and Henkjan Honing for his help with the interface and POCO. I also thank them and the reviewers for their helpful comments.

[1]The recordings of the Chopin Prelude were taken from the following CDs: Argerich, Philips Classics, 456 703-2 (Great Pianists), recording 10/1975, München Herkules-Saal (orig. Deutsche Grammophon); Harasiewicz, Philips Classics 442 268-2, recorded 1963; Kissin, BMG, 09026 63535 2, recorded 1999; Pollini, DGG 413 796-2, recorded 1975; Rubinstein, BMG, GD 60822, recorded 1946.

[2]The recordings of the Mozart Sonata were taken from the following CDs: Daniel Barenboim, Emi Classics, CDZ 7 67295 2, recorded 1984; Roland Batik, Gramola, 98701-705, recorded 1990; Glenn Gould, Sony Classical SM4K 52627, recorded 1967; Maria João Pires, DGG, 431 761-2, recorded 1991; András Schiff, ADD (Decca), 443 720-2, recorded 1980; Mitsuko Uchida, Philips Classics, 464 856-2, recorded 1987.

[3]The Greenhouse–Geisser correction for violations of sphericity was applied whenever appropriate. If the correction is applied, the uncorrected degrees of freedom are reported. The reported probability value is the probability following correction.

Bengtsson, I., and Gabrielsson, A. (**1983**). "Analysis and synthesis of musical rhythm," in *Studies of Music Performance*, edited by J. Sundberg (Royal Swedish Academy of Music, Stockholm), pp. 76–181.

Clarke, E. F. (**1993**). "Imitating and evaluating real and transformed musical performances," Music Percept. **10**, 317–341.

Dixon, S. (**2001**). "Automatic extraction of tempo and beat from expressive performances," J. New Music Res. **30**, 39–58.

Dixon, S., and Goebl, W. (**2001**). ''Pinpointing the Beat: Tapping to Expressive Performances,'' in *Proceedings of the 7th International Conference on Music Perception and Cognition* (ICMPC7, Sydney, Australia), pp. 617–620.

Friberg, A., and Sundberg, J. (**1999**). ''Does music performance allude to locomotion? A model of final *ritardandi* derived from measurements of stopping runners,'' J. Acoust. Soc. Am. **105**, 1469–1484.

Friberg, A., Frydén, L., Bodin, L. G., and Sundberg, J. (**1991**). ''Performance rules for computer-controlled contemporary keyboard music,'' Comput. Music J. **15**, 49–55.

Gabrielsson, A. (**1987**). ''Once again: The theme from Mozart's Piano Sonata in A major: A comparison of five performances,'' in *Action and Perception in Rhythm and Music*, edited by A. Gabrielsson (Royal Swedish Academy of Music, Stockholm), pp. 81–103.

Gabrielsson, A. (**1988**). ''Timing in music performance and its relation to music experience,'' in *Generative Processes in Music. The Psychology of Performance, Improvisation, and Composition*, edited by J. Sloboda (Clarendon, Oxford), pp. 27–51.

Gjerdingen, R. O. (**1988**). ''Shape and motion in the microstructure of song,'' Music Percept. **6**, 35–64.

Hevner, K. (**1937**). ''The affective value of pitch and tempo in music,'' Am. J. Psychol. **49**, 621–630.

Honing, H. (**1990**). ''POCO, An Environment for Analysing, Modifying and Generating Expression in Music,'' in *Proceedings of the 1990 International Computer Music Association* (CMA, San Francisco), pp. 364–368.

Kendall, R. A., and Carterette, E. C. (**1990**). ''The communication of musical expression,'' Music Percept. **8**, 129–164.

Kronman, U., and Sundberg, J. (**1987**). ''Is the musical ritard an allusion to physical motion?'' in *Action and Perception in Rhythm and Music*, edited by A. Gabrielsson (Royal Swedish Academy of Music, Stockholm).

Langner, J., and Goebl, W. (**2003**). ''Visualizing expressive performance in tempo-loudness space,'' Comput. Music J. **27**, 69–83.

Levitin, D. J., and Cook, P. R. (**1996**). ''Memory for musical tempo: Additional evidence that auditory memory is absolute,'' Percept. Psychophys. **58**, 927–935.

Nakajima, Y. (**1987**). ''A model of empty duration perception,'' Percept. **16**, 485–520.

Palmer, C. (**1989**). ''Mapping musical thought to musical performance,'' J. Exp. Psychol. Hum. Percept. Perform. **15**, 331–346.

Palmer, C. (**1996**). ''On the assignment of structure in music performance,'' Music Percept. **14**, 23–56.

Pampalk, E., Rauber, A., and Merkl, D. (**2002**). ''Content-based organization and visualization of music archives,'' in *Proceedings of the 10th ACM International Conference on Multimedia* (MM'02, Juan-les-Pins, France), pp. 570–579.

Repp, B. H. (**1990**). ''Patterns of expressive timing in performances of a Beethoven minuet by 19 famous pianists,'' J. Acoust. Soc. Am. **88**, 622–641.

Repp, B. H. (**1992a**). ''Diversity and commonality in music performance—An analysis of timing microstructure in Schumann's Traumerei,'' J. Acoust. Soc. Am. **92**, 2546–2568.

Repp, B. H. (**1992b**). ''Probing the cognitive representation of musical time: Structural constraints on the perception of timing perturbations,'' Cognition **44**, 241–281.

Repp, B. H. (**1994**). ''Relational invariance of expressive microstructure across global tempo changes in music performance: An exploratory study,'' Psychol. Res. **56**, 269–284.

Repp, B. H. (**1998**). ''Obligatory 'expectations' of expressive timing induced by perception of musical structure,'' Psychol. Res. **61**, 33–43.

Repp, B. H. (**2000**). ''Pattern typicality and dimensional interactions in pianists' imitation of expressive timing and dynamics,'' Music Percept. **18**, 173–211.

Seashore, C. E. (**1938**). *Psychology of Music* (Dover, New York).

Sundberg, J., Friberg, A., and Frydén, L. (**1989**). ''Rules for automated performances of ensemble music,'' Contemp. Music Rev. **3**, 89–109.

Thompson, W. F., Sundberg, J., Friberg, A., and Frydén, L. (**1989**). ''The use of rules for expression in the performance of melodies,'' Psychol. Music **17**, 63–82.

Timmers, R. (**2003**). ''On the contextual appropriateness of expression,'' Music Percept. **20**, 225–240.

Timmers, R., and Honing, H. (**2002**). ''On music performance, theories, measurement and diversity,'' Cognitive Processing (International Quarterly of Cognitive Sciences) **1–2**, 1–19.

Timmers, R., Ashley, R., Desain, P., and Heijink, H. (**2000**). ''The influence of musical context on tempo rubato,'' J. New Music Res. **29**, 131–158.

Todd, N. P. (**1992**). ''The dynamics of dynamics: A model of musical expression,'' J. Acoust. Soc. Am. **91**, 3540–3550.

Zanon, P., and Widmer, G. (**2003**). ''Learning to recognize famous pianists with machine learning techniques,'' in *Proceedings of the Third Stockholm Music Acoustics Conference* (SMAC03, Stockholm, Sweden).