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

Renee Timmers
Austrian Research Institute for Artificial Intelligence, Vienna

Address of correspondence:
Renee Timmers, Austrian Research Institute for Artificial Intelligence, Freyung 6/VI, A-1010 Vienna, Austria. tel: +43-1-5336112-24, fax: +43-1-5336112-77, e-mail: renee74@xs4all.nl

Running title: Subjective and objective similarity between performances

Abstract

Measurements of expression in music performances have highlighted the subtle variations in tempo, timing and dynamics that musicians apply when performing music. Comparisons between the variations have given a first insight into the diversity of interpretations of music. Correlation is an often-used method for such comparisons. This study investigated the perceptual validity of such measurements and, more specifically, of such objective comparisons. 20 participants rated the similarity between 5 performances of a fragment from a Chopin Prelude and between 6 performances of two fragments from a Mozart Sonata. Variations in tempo and dynamics were measured from the audio recordings of the performances. These measurements were input to different models that predicted the perceived similarity between performances. Overall, the models could predict a fair amount of the similarity ratings. Tempo was especially important for the prediction, more important than loudness, followed by the interaction between tempo and loudness. Models based on absolute measures were stronger than models based on normalized measures. These results were independent of the musical background of participants. The implications for future research include a reconsideration of correlation for the comparison of performances and a revaluation of absolute local measures. In addition, the study suggests directions for the further investigation of the perception of music performance.

Introduction

Measurements of performances and especially of piano performances have become an important means to improve the understanding of musical expression and interpretation. A performer offers an interpretation of the composed music and this interpretation is expressed or enclosed in his or her rendition of the score. By delaying or accentuating notes, s/he may put stress to certain notes or chords, while by performing a movement legato, slow and with small dynamical variations, s/he may give it for example a sad connotation (Gabrielsson & Juslin, 1996).

Measurements have shown the extensive use of subtle variations as well as the consistency of such variations in repeated performances (both already observed by Seashore, 1938), and the controllability of such variations (e.g. Kendall & Carterette, 1990; Palmer 1989; 1996). Comparisons between variations have given a first insight 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, an important concern is to what extent it reflects the psychological reality of performers and listeners. The ecological validity for performers and listeners of measurable variations has been suggested by several studies. First of all, in several studies, performers were asked to perform music without expression, with normal expression and with exaggerated expression (e.g. Kendall & Carterette, 1990; Penel & Drake, 1999). In the exaggerated expression, variations enlarge, while the variations diminish in the performance without expression. The presence of some residue of the variations in the performance without expression suggested some uncontrollability. Secondly, analysis-by-synthesis studies provide evidence for the importance and validity of expressive variations. Sundberg, Friberg and Frydén (1989) developed a rule system for performances by teaching a computer to perform music with appropriate expression (see also Friberg et al, 1991). In other words, all remarks from a professional musician (Frydén) to improve the performance were translated into concrete formulations of rules for the variation of tempo, timing, dynamics or intonation. Similarly, the quality of synthesised performances has been judged to be high compared to the quality of deadpan performances (Thompson et al, 1989).

On the other hand, several studies have shown that the perception of time intervals do not directly correspond to their physical properties. For example, Repp (1992b, 1998) tested the detection of "timing perturbations" and found a dependency on position within the phrase structure of music. This dependency showed that to some extent perception deviates from measured variations. And Nakajima found a systematic over-estimation of short empty duration intervals by a constant interval, which means that the relative over-estimation increases with decreasing durations (Nakajima, 1987). For longer time-spans, time estimation did seem to be in direct proportion to physical time (Clarke & Krumhansl, 1990). The communication of performed dynamic levels has been investigated by Nakamura (1987). He found that the communication from performer to listener of intended dynamics level was fairly good, even when evidence in measured intensity was absent or minor. A complication of the perception of dynamics and loudness accents is the interaction of loudness with duration and with musical structure. The subjective intensity of short tones tends to increase with duration and tones that are structurally accented are perceived as louder

than structurally unaccented tones (for a discussion of different accent types see e.g. Jones, 1987; Lerdahl & Jackendoff, 1983; Parncutt, 1994).

This study also tested the perceptual validity of measurements of performance variations, but took a different perspective. The aim was to test to what extent measured characteristics of the tempo and loudness of a performance capture the main attributes of the performance and in particular to what extent calculated differences between these measurements are able to predict the perceived distance between performances. As such, it took a more global perspective than previous studies. Commercially available recordings were used and the variations in tempo and loudness analysed. Twenty listeners rated the similarity between pairs of performances. These similarity ratings were predicted on the basis of differences in measured characteristics of tempo and loudness in order to test their weight and importance for the characterisation of the performance. If tempo and loudness are most salient characteristics, the similarity prediction should well succeed. If however other attributes are equally important or if tempo and loudness are less directly perceived as the measurements assume, the similarity prediction should be much less successful.

The test of the presence of a gap between measurement and perception was an implicit test for research that analyses local tempo and local loudness to characterise a performance and to capture the main attributes of a performer's style (see e.g. Zanon, & Widmer, 2003). The current test seems crucial for the interpretation of such studies. Moreover by comparing the predictive power of different models that have their origin in music performance research, the perceptual value of the different methods was examined.

Method

Musical material

Five performances of Chopin's Prelude op. 28 no. 17 were used and six performances of the first movement of Mozart's Sonata K.281. The five performances of the Chopin Prelude were by Argerich, Harasiewicz, Kissin, Pollini, and Rubinstein (to be referred to as p1, p2, p3, p4, and p5, respectively) and the six performances of the Mozart Sonata were by Barenboim, Batik, Gould, Pires, Schiff, and Uchida (to be referred to as p1, p2, p3, p4, p5, and p6 respectively). The opening bars of the two pieces were 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). Reference to these fragments will be Ch, M1, and M2.

These performances are part of a database of recordings from which local tempo and loudness at the beat level have been analysed (see Widmer, 2003). A Mozart Sonata and a Chopin Prelude were chosen, because several performances of these pieces by famous pianists were available, and because performances of both Mozart Sonata's and Chopin Preludes are often used in performance research. In addition, the expectation was that the prediction would work better for the Chopin Prelude than for the Mozart Sonata. While the Chopin Prelude has a stable 8^{th} note rhythm and a constantly full texture of chords, the Mozart Sonata contains ornaments, trills and arpeggiated chords, the interpretation of which might be more important for the characterization of the performances than the tempo and dynamics.

Participants

Twenty participants rated the similarity between pairs of performances. The participants had different levels of musical training. Five participants did not have had any musical training. Six participants were professional musicians, among them four pianists and two non-pianists (string players). Five participants were amateur musicians (all non-pianists), but were nevertheless full-time busy with music. The other four participants were amateur musicians, among them two pianists and two non-pianists.

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 of the participants. The order of the Mozart fragments was always M1 followed by M2. The comparisons between performances were grouped into blocks that contained one reference performance and four or five comparison performances, depending on the total number of performances of a fragment. This grouping of performances was done to familiarize the participants with the variety of performances present in the test and so provide a framework for the similarity ratings. The order of the comparisons and the reference performances was randomised over participants.

The participants sat behind a computer and saw the user interface depicted in Figure 1 on the screen. The interface contained play buttons for one reference performance and for four or five 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. 1 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 that consisted of the same performances in a different order of comparison performances and with a different reference performance. The reference performance was not present among the comparison performances. 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: one time with one of the performances as reference and the other time with the other performance as reference.

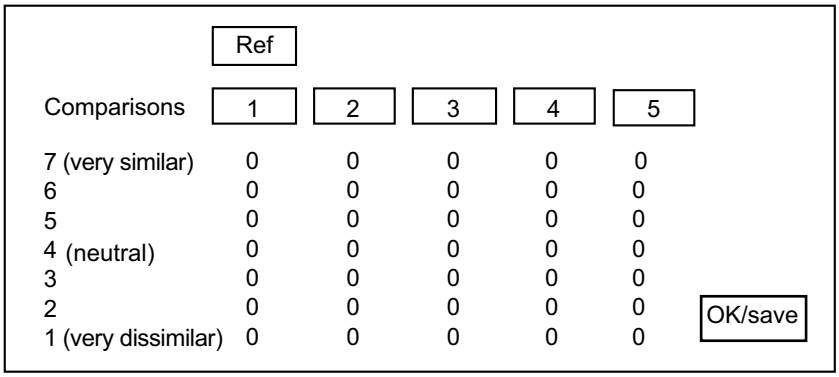


Figure 1: User interface for the similarity rating study

Similarity Predictions

Central to this study was the prediction of the subjective judgement of similarity between performances on the basis of objective measurements of the performances. Two approaches were taken for this prediction. The first approach compared the predictive power of different models that were all based on a loudness and a tempo component. The difference between the models concerned the calculation of the difference in tempo and loudness between two performances. The second approach defined for each participant a separate optimal model that explained the maximum of variance. This optimal model could consist of any combination of different loudness and/or tempo components as long as the contribution of each parameter to the explanation was significant. The description given next only concerns the first

approach. In the "Result" section, the procedure for the optimal model is further explained.

The measurements were done 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 & Goebl, 2001). Local tempo was calculated for each inter-beat-interval. To get a measure of local loudness, the localised beats were used as well. Around each beat, the maximum amplitude level was selected (see also Langner & Goebl, 2003). This level in dB was recalculated into sone (see Pampalk et al., 2002).

Appendix 1 shows local tempo and local loudness at the 8^{th} note level for the Chopin fragment, and at the quarter note level for the Mozart fragments. The 8^{th} note level is quite a fine-grained level for the Chopin piece that is in 6/8, which is the reason for the measured tempo of this fragment to be rather high. The dynamics of the M1 show a clear drop in the middle of the fragment, which is due to a dotted quarter note pause in the music. The beat coinciding with the pause is left out of the calculations on the loudness component, because it would bias the calculations towards greater similarity.

Local tempo and local loudness were input to the similarity prediction models. Regression models were used that consisted of a tempo and a loudness component and predicted the distance between two performances as a linear function of the absolute difference in local tempo and local loudness between the performances. Different models used different measures of the difference in local tempo and local loudness.

To clarify, the general form of the regression model assumes that the similarity ratings can be predicted from a weighted summation of the difference in tempo (D_t) and the difference in loudness (D_l).

$$s = a + b * D_t + c * D_l \qquad (1)$$

The equations for the different measures of the difference in local tempo and in local loudness are listed below. For brevity, the measures are only presented for the tempo component. Similar calculations were applied to the loudness component.

Global	$\left\|\overline{T_1} - \overline{T_2}\right\|$		(2)
Local	$\overline{\left\|T_1 - T_2\right\|}$	Absolute and Normalized	(3)
Std deviation	$\left\|std(T_1) - std(T_2)\right\|$	Absolute and Normalized	(4)
Change	$\overline{\left\|T_1' - T_2'\right\|}$	Absolute and Normalized	(5)
Change of change	$\overline{\left\|T_1'' - T_2''\right\|}$	Absolute and Normalized	(6)
Correlation	$Corr(T_1, T_2)$		(7)

Capital *T* refers to the local tempo at each beat of a performance, while the subscript refers to one of the two performances of a performance pair that are compared. The accents in change and change of change indicate the first and second derivative of tempo, respectively. The subtraction of the tempo (and loudness) of two performances is done per beat for all models, except for the global model for which the subtraction is done on the average tempo of the two performances. In other words, the local models first subtract the tempo (and loudness) per beat and than average this calculated difference, while the global model first calculates the average and then subtracts.

The models based on local tempo and loudness, the standard deviation of tempo and loudness, tempo and loudness change and tempo and loudness change of change had two versions: an absolute and a normalized version. The input to the first were the absolute values of tempo and loudness, while the input to the second were the normalized values of tempo and loudness. Normalized values express the tempo and loudness levels as ratios of or deviations from average tempo or loudness. This

procedure is often used in empirical performance research (see e.g. Gabrielsson, 1987; 1988).

All measures have a precedent in performance research. Bengtsson and Gabrielsson (1983) differentiate between the abstract mean tempo, the main tempo, local tempo and the beat rate. The first and last are used here (although we refer to the beat rate as local tempo), which are the measures most directly derived from performance timing data. The main tempo and local tempo are tempi that hold for a longer or shorter period. The main tempo differs from the mean tempo when the performance contains large ritenuti at the end of pieces or sections (see also Repp, 1994a). Because the fragments that are used in the present study are phrases taken from the start and the middle of the piece, we did not find this extra measure necessary. To take the average tempo of short time periods (local tempo in Bengtsson and Gabrielsson's terminology) might be a sensible alternative that for brevity was not considered here. The models of this study take the two extremes of global and beat-level measures and test the validity of these extremes.

The standard deviation is used to measure the amount of rubato present in a performance, which could be characteristic to certain performers (see e.g. Timmers et al, 2000; Timmers, 2003; and Zanon & Widmer, 2003). Acceleration is used in models of motion and movement and is taken as the first derivative of tempo by Kronman and Sundberg (1987). Gjerdingen (1988) introduces the concept of change of change in his discussion of representations of pitch and intensity trajectories in vocal performance and argues that change of change or the speed of change in dynamics and pitch are strong cues for perception. Finally, correlation is an often-used measure to compare performances and to get an indication of the similarity or dissimilarity in interpretation between two performances (Clarke, 1993; Repp, 1994b; 2000; Timmers et al., 2000).

Besides models based on a separate loudness and tempo component, two models were used that consisted of a combined tempo and loudness component and test the predictive ability of the differences between the interaction between tempo and loudness of two performances. The first model takes the absolute difference in the average tempo times loudness per beat, while the second takes the average of the absolute difference in tempo times loudness per beat.

Global T*L $$\left|\overline{T_1 * L_1} - \overline{T_2 * L_2}\right| \quad (8)$$

Local T*L $$\overline{\left|T_1 * L_1 - T_2 * L_2\right|} \quad (9)$$

The twelve regression models based on these eight absolute measures and four normalized measures were fitted per fragment to the similarity ratings of individual participants and to the average similarity rating of all musicians and all non-musicians.

Results

The average dissimilarity rating for the two presentations of each performance pair of each fragment was input to a multidimensional scaling analysis in order to get an impression of the distances between the performances as judged by the participants. The dissimilarity rating was simply defined as the inverse of the similarity rating. The multidimensional scaling analysis of SPSS10 was used that bases the scaling on an asymmetrical dissimilarity interval input. Figure 2 shows the two-dimensional solution for each fragment together with an indication of stress and explained variance (R^2). For the Chopin fragment, the performances were polarized on the second dimension (p1 and p5 – Argerich and Rubinstein – versus the others) and well spread on the first dimension. P4 (Pollini) is most close to all the others. For the Mozart fragments, the distances between the performances were similar for the two fragments with the exception of p2 (Batik) who changes position. There was a clustering of performances around p5 (Schiff), whose performance was relatively close to the performances of p1 and p6 (Barenboim and Uchida). P3 (Gould) and to some extent p4 (Pires) were more at a distance to the other performances.

The interpretation of the basis of these relative distances between performances was left to the prediction of the similarity ratings by the different models. A preliminary interpretation of these dimensions suggested however the importance of global tempo, which correlates highly with dimension 1 of Chopin and M1, global tempo times loudness, which correlates highly with dimension 1 of M2, and the standard deviation of loudness, which correlates highly with dimension 2 of M2. (These correlations were all above 0.84).

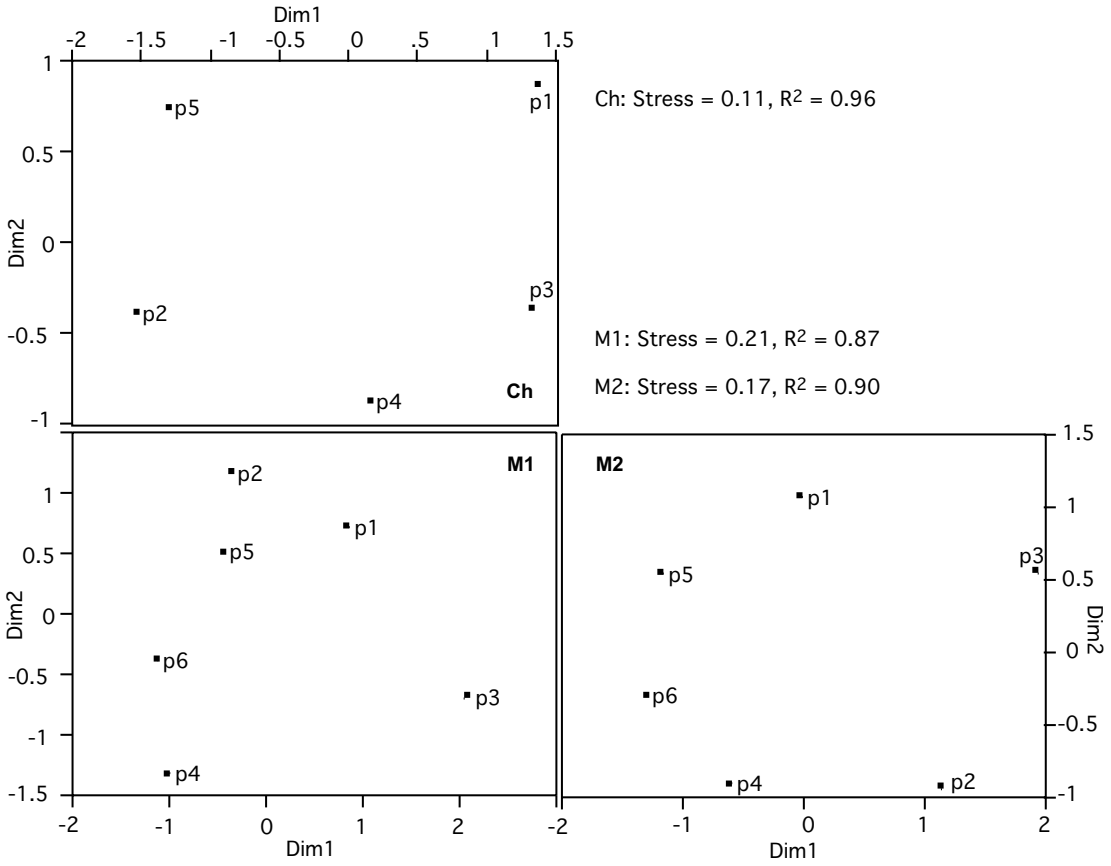


Figure 2: Solution of a two-dimensional scaling analysis on the average dissimilarity rating of pairs of performances of the three fragments.

To test if the effect of performance pair was systematic over participants and presentations, a repeated measures ANOVA was run with pair, order and the interaction between pair and order as independent variable and the similarity ratings as dependent variable. For M1 and Ch, the main effect of pair was the only significant factor, also when the analysis was corrected for violations of sphericity ($\varepsilon < .001$). For M2 however, all effects were significant using the Greenhouse-Geisser correction for violations of sphericity, though the effect of pair remained the strongest effect (F (5.4, 102) = 122, $\varepsilon < .001$ for the main effect of pair, F (1, 19) = 10.8, $\varepsilon = .004$ for the main effect of order, and F (6.8, 129) = 7.6, $\varepsilon = .004$ for the interaction effect). The main effect of order was unexpected and hard to explain. The effect was not due to a real order effect, since the order of presentations was randomised. Instead it seemed due to the specific pairs of which the first order (the pianist with the lower number as reference) was more often rated relatively high than the second order (the pianist with

the higher number as reference). The interaction between pair and order was not surprising however and is further discussed in the section on "Inconsistency".

Although the main effect of performance pair was significant, the agreement between participants was not very high. The average correlation between the ratings of different participants was 0.44 for Chopin, 0.38 for M1 and 0.43 for M2, which is about significant for a correlation between 20 or 30 data points respectively ($p =< 0.05$).

More important than the average ratings for the performance pairs were the results of the similarity prediction by the different models. Figure 3 shows the average R^2 (averaged over participants) of the twelve models that were used to explain the similarity ratings of individual participants. The letter on top of each line indicates which parameter (t for tempo and l for loudness) contributed most often significantly to the explanation. If no letter is indicated, both tempo and loudness contributed equally often to the explanation or none of them reached significance. The latter was the case for the standard deviation of M2 and the normalized standard deviation for Ch. Global and local tempo times loudness do not have a letter, because they always include both variables.

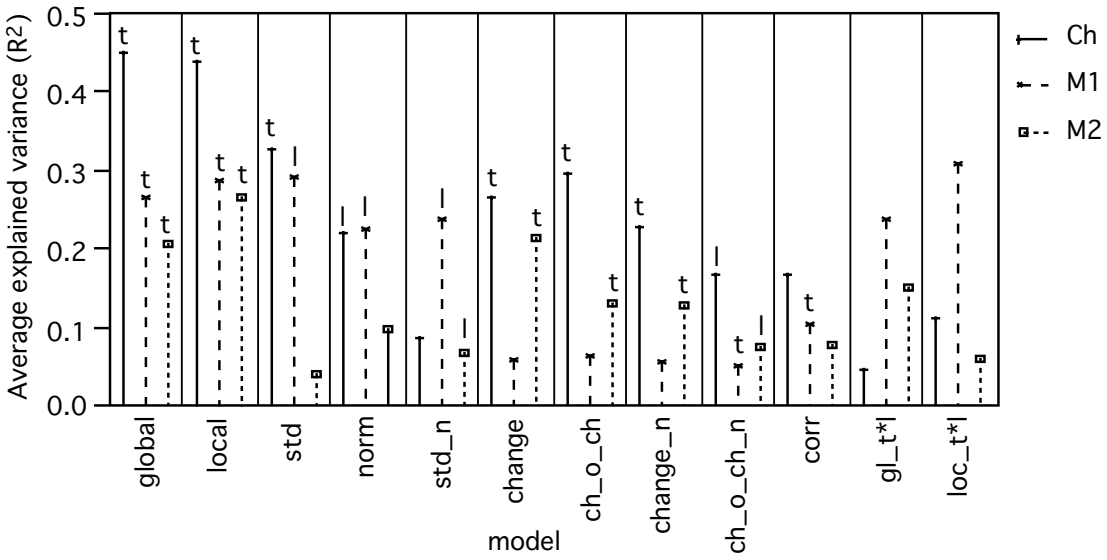


Figure 3: Average explained variance (R^2) of the fits of each model to the similarity ratings of individual participants. Separate bars are plotted for the three fragments. Letters indicate the component (t for tempo and l for loudness) that contributed most often significantly to the fit.

These results are based on twenty data points for the Chopin fragment and thirty data points per model for the Mozart fragments. In other words, the calculated differences in local tempo and loudness for each performance pair were fitted to the ratings of each pair in two orders. The main reason for this was that it provided a sufficient number of data points for the regression analysis and the variability with order remained to be taken into consideration.

Figure 3 shows that the explained variance was generally higher for the Chopin fragment than for the Mozart fragments. It also shows that the models based on absolute measures did better than the models based on normalized values. The model based on correlations did rather badly. The models based on the interaction between loudness and tempo did well for the two Mozart fragments, but not for the Chopin fragment.

It also shows that tempo was more often responsible for the explained variance than loudness, but for some models the pattern was reverse: Loudness was more often significant for the normalized models and the models based on the standard deviation for the Mozart fragments.

The problem of this presentation of the results is that the independent variables of the different models correlate with each other to some extent. The average of the absolute correlation between independent variables was 0.38 for the Chopin fragment and 0.32 for the Mozart fragments. Table 1 shows for the Chopin fragment the variables that had a correlation larger than 0.8. Therefore part of the explained variance by one model could have been due to the correlation with another model. For example, the local and global models confound to a considerable extent as well as the models based on change and change of change.

Table 1: *Independent variables of models of the Chopin fragment that have a correlation of 0.8 or more.*

Variable 1	Variable 2	r	Variable 1	Variable 2	r
Dtlocal*	Dt	0.96	Dtch_norm	Dtstd	0.85
Dtlocal	Dtstd	0.88	Dtch_norm	Dtchange	0.90
Dllocal	Dl	0.87	Dtch_norm	Dtch_o_ch	0.95
Dtchange	Dtstd	0.87	Dlch_norm	Dlchange	0.93
Dtchange	Dtstd_norm	0.90	Dlch_norm	Dlch_o_ch	0.86
Dtch_o_ch	Dtstd	0.91	Dlch_o_ch_norm	Dlchange	0.92
Dtch_o_ch	Dtstd_norm	0.84	Dlch_o_ch_norm	Dlch_o_ch	0.93
Dtch_o_ch	Dtchange	0.93	Dlch_o_ch_norm	Dlch_norm	0.87
Dlch_o_ch	Dlchange	0.97	Dt*l_local	Dt*l	0.85

* The name of the variables indicates that the variable concerns the difference (D) in tempo (t) or loudness (l) using the local measure or one of the other measures.

To overcome this problem, an optimal regression model for each participant was sought that would only consist of components that contribute significantly to the explanation and that together explain the largest part of the data (have the highest R^2). This optimal model could consist of any combination of tempo and loudness components of the different measures.

Figure 4 shows the sum of the explained variance by components that were once or more often part of an optimal model. The numbers on top of each bar indicate for how many participants the specific component was part of the optimal model. For most instances (36 out of 20 participants times three fragments), the optimal model consisted of only one component. In 19 instances, it consisted of two components and, for only 1 instance, it consisted of three components. For the remaining four instances, no significant model was found. When the optimal model consisted of several components the total explained variance was equally divided over each component in the calculation of the sum of the explained variance per component as plotted in Figure 4.

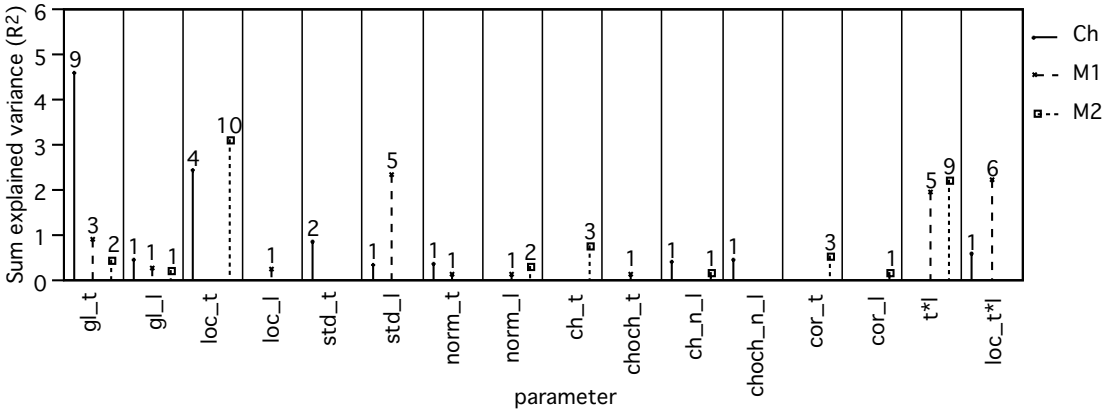


Figure 4: Sum of the explained variance (R^2) of parameters that were once or more times part of an optimal model for individual participants. Separate bars are plotted for the three fragments. Numbers indicate the number of participants for which the specific parameter was significant.

Table 2 lists the average explained variance for the three fragments and the two components that were most important for each fragment. In addition, it shows the components of the optimal models for the average rating of all musicians and the

average rating of all non-musicians, together with their respective explained variances.

Table 2: *Main parameters and average explained variance for the optimal models for individual participants as well as the explained variance and parameters of the optimal models for the average rating by musicians and the average rating by non-musicians.*

Fragment	Individual fit		Musicians		Non-musicians	
	Main par.	Av. R^2 (all par.)	Par.	R^2	Par.	R^2
Ch	gl_t loc_t	0.52	gl_t	0.87	loc_t	0.69
M1	std_l t*l_l	0.42	std_l t*l	0.68	std_l t*l	0.69
M2	loc_t t*l	0.40	loc_t t*l	0.63	loc_t t*l	0.68

Musicians and non-musicians

Among the participants were five without any musical training. The other participants were amateur or professional musicians. The five non-musicians were expected to show deviant results from the musicians: They were expected to show smaller explained variances and significance of models that imply a more simple evaluation of the performances than the models for the musicians would.

The first test of a systematic difference compared the optimal models for the average rating of the non-musicians and the musicians. The parameters and R^2 of these optimal models are given in Table 2 and show a close agreement. The results differ only for the Chopin fragment and in an opposite direction than expected: the global model was strongest for the musicians and the local model for the non-musicians, which suggests that the non-musicians paid closer attention to the details of the tempo variations than the musicians.

For a second test, participants were ordered according to their level of musical training. Professional pianists were assigned the highest number (5) while the five amateur musicians were assigned the lowest number (0). The other participants were assigned to four intermediate levels depending on the instrument they played (piano or another instrument) and the intensity of their musical occupations. Then the different measures were ordered in absolute and normalized measures, in global and local measures, and in direct and derived measures (tempo and loudness versus tempo and loudness change or change of change). The assumption was that 1) local measures were more difficult than global measures (more details need to be remembered), 2) normalized measures were more difficult than absolute measures (more abstract from

the physical input), and 3) measures based on change and change of change were more difficult than absolute measures (more abstract from the physical input). The measures that treated tempo and loudness as separate variables were considered equally difficult as the measures that combined tempo and loudness into one parameter, because, on the one hand, the separate treatment of the parameters may be considered more analytic, but, on the other hand, the focus on one parameter may also be considered to be simpler than the focus on two parameters.

The level of training was correlated with the R^2 of the optimal model, the levels of abstraction of the optimal model (local, normalized and derived), and the average inconsistency of the participants in their ratings of the performance pairs in the two presentations. None of these correlations were significant except for the correlation between training level and R^2 of the second Mozart fragment (r = -0.52, p < 0.02). Training and R^2 were significantly negatively correlated, which indicates that, for one fragment, the R^2 was relatively high for participants with less musical training, contrasting the expectations on the effect of musical training. Overall however, no systematic effect of musical training was found.

On average, the significant measures present in the optimal models were more often global than local for Chopin and M1 (4 out of 10 were local), but a little more often local than global for M2 (6.5 out of 10 were global), they were more often absolute than normalized (9 out of 10 were absolute), and more often direct than derived (9 out of 10 were direct).

Artefacts of the analysis method

The procedure of the selection of the optimal model brings three problems that should be considered to some detail here: 1) correlation between independent variables, 2) limited ability for more than one component to be significant, and 3) some measures contain more information about the performances than other measures and often include information of other measures as well.

In other words, the measure that keeps most information about the performance is most likely to explain a large part of the variance. This means that normalized measures have a disadvantage, because they lack information about the average tempo and loudness. It also means that the tempo times loudness measures may have the advantage of combining two parameters in one value. Finally it means that although certain measures may explain the largest part of the variance, other

effects could have contributed to the success of these measures. It seems important to consider the main results of this study in light of these difficulties.

Table 2 lists the parameters that were strong in explaining the similarity ratings. These were global tempo, local tempo, standard deviation of loudness, and local and global tempo times loudness. The local and standard deviation measures have the benefit of being correlated with the global measure, in other words they contain information about both global and local features. The question is whether two parameters would have done equally well or better, but had the disadvantage of being less easily significant. To test this, the explained variance of local tempo was compared to the explained variance of global tempo and normalized tempo and the explained variance of standard deviation of loudness was compared to the explained variance of global loudness and normalized standard deviation. Both measures withstood the test, though the difference between the two alternatives was small. Generally, local tempo and standard deviation explained more than global tempo & normalized tempo and global tempo & normalized standard deviation. This was true for 16 out of 19 cases[1].

The tempo times loudness measures have the benefit of combining two parameters. To test whether the interaction indeed performed better than the separation, the explained variance of the interaction was compared to the explained variance of the addition of the two components. For all participants that had tempo times loudness as their optimal model, the interaction explained considerably more than the addition of the separate measures. This was also the case for four out of six participants who had local tempo times loudness as their optimal model, though the differences in explained variance were smaller for the local measures.

Inconsistency

Although the optimal models were rather good at explaining the similarity ratings between performances, they were not perfect. One of the aspects that they did not account for was the effect of context of presentation. As explained before, participants rated each performance pair twice: once with one performance as reference and

[1] The 19 cases were four participants who had local tempo as their optimal model for the Chopin fragment, 10 participants who had local tempo as their optimal model for the second Mozart fragment and five participants who had the standard deviation of loudness as their optimal model for the first Mozart fragment.

another time with the other performance as reference. Inconsistency in similarity rating lowered the explained variance systematically as suggested by the significant negative correlation between inconsistency and explained variance per participant (r = -0.73 for Chopin, r = -0.72 for M1, though r = -0.22 for M2).

The repeated measures ANOVA reported above showed only for M2 a significant interaction between the effect of pair and order on the similarity ratings. Nevertheless, several participants showed considerable order effects for all fragments. This may have been due to a general inconsistency or change in attention to different aspects of the performance, but it may also sometimes have been a repeatable or systematic effect of context. Several participants were aware of the changing ratings that they gave or at least of the changing perspectives that they took when the reference performance changed. Pair p3-5 of M1 had the highest average inconsistency; the difference in similarity ratings between the two presentations was 1.85 on average. The similarity was judged to be higher when p3 was the reference than when p5 was the reference. This context effect is easily explained, when we consider that p3 (Gould) was the most different performance of all, while p5 (Schiff) was a more average or prototypical performance. The average similarity rating with other performances was 2.5 for p3 and 4.1 for p5, which were the lowest and highest average rating per pianist for M1. Given the limited rating scale, p3 was more different from p5 than the other performances, but the other performances were as distant or more distant to p3 as p5. Therefore the similarity ratings of p3 with p5 as reference were on average lower than of p5 with p3 as reference.

Interviews

After the similarity rating experiment, the participants filled in a questionnaire on their evaluation of the experiment and indicated in a table for different aspects of a performance to what extent they had paid attention to the phenomenon by assigning a number between 0 (no attention) and 3 (much attention) to each aspect.

The questionnaire asked four questions that concerned 1) general remarks about the test, 2) the difficulty of the test, 3) the clarity of the task, and 4) the factors that influenced the similarity judgements. The general remarks of the participants included the differences between the cues that were important for different fragments and for different reference performances, the fun of the test, the learning effect within the test, the tiredness towards the end, the length of the Chopin fragment and brevity of the first Mozart fragment, the large differences between the performances (some

musicians), but also the small difference between the performances that became easier to notice after a while (two non-musicians).

Most participants found the test a bit difficult. They especially encountered problems with the definition of similarity and with being consistent in judgement. The judgement depended on the way of listening, on the level of detail of listening, and the importance of the various aspects characteristic of a performance. Nevertheless the task was found clear and participants were able to indicate factors that had influenced their judgments. The factors that were mentioned and the number of participants that mentioned them were the following: Tempo (11), loudness (2), the sound of the recording (1), phrasing (1), rubato (5), dynamics (4), articulation (7), the quality of the pianist / the smoothness of playing (4), character & style / overall impression (7), interpretation (6), arpeggios, ornaments (2), perception of movement (1).

The attention ratings show a similar pattern (see Table 3): Most attention was paid to global tempo and rubato and less attention was paid to dynamics and little to overall loudness. Articulation was an important factor for the Mozart fragments, but less for the Chopin fragment. The interpretation of the music was important as well as the character and style of the performance.

Table 3: *Sum of attention ratings for the Chopin and Mozart fragments expressed as percentage of the maximal sum of the ratings (18*3: number of subjects that filled in the questionnaire times maximal attention rating).*

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

The importance of tempo, tempo variation and dynamic variation agrees 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. While, speculatively, the overall impression such as character and

style may be related to the larger contribution of the global measures and 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 an interaction with musical structure would have improved the variance explained.

Summary and discussion

The aim of this study was to test to what extent measured characteristics of the tempo and loudness of a performance capture the main attributes of the performance and in particular to what extent calculated differences between these measurements are able to predict the perceived distance between performances. In addition, the explanatory power of different models was compared.

The optimal models that explained the largest amount of variance for individual participants consisted mostly of the parameters: global and local tempo for Chopin, standard deviation of loudness and local tempo times loudness for M1, and local tempo and global tempo times loudness for M2. The optimal models for the average ratings of musicians and non-musicians consisted of the same parameters with this difference that global and not local tempo times loudness was most explanatory for the average ratings of M1. The results showed that generally the same parameters were important for non-musicians as for musicians, the explained variance was overall the same for musicians as for non-musicians, and global as well as local absolute and direct measures were most often significant for both groups. Of these measures, tempo was most often significant, followed by tempo times loudness. Loudness also contributed significantly to the explanation of the Mozart fragments mainly in the form of the standard deviation of and normalized loudness. In other words, loudness seemed to be evaluated in a more abstract manner than tempo.

One of the reasons for the lack of difference between musicians and non-musicians could have been that many musicians did the task with relative ease and speed, while the non-musicians did the task with larger concentration and effort. A different reason could have been that the tests used in this paper were not suited to find the specific differences in perception of performances between musicians and non-musicians that still may be expected to exist.

The results withstood the tests that were run to check if the results were due to artefacts of the used method. The recognized artefacts included the correlation between independent variables, a limited ability for more than one parameter to

become significant in an optimal model, and a difference in the amount of information that the measures contained about the performances. Although these aspects might still have played some role, the main results of the study did survive the tests for artefacts. For example, the interaction between tempo and loudness explained considerably more of the ratings of the Mozart fragments than the sum of the separate components, and local tempo explained a little more than global tempo and normalized tempo together.

The interviews held with the participants highlighted some of the aspects that were missed by the models: First of all, the models did not account for inconsistency of or changing perspective on the rating of similarity by the participants. Secondly, the models did not include articulation, timbre, fluency or quality of the performance, and did not take the interpretation of musical structure into account.

Nevertheless, some of the more important aspects of a performance to which the participants said to attend to were included in the models such as global tempo, rubato, and dynamics. Indeed the variables that were included could account for a fair amount of the data given the average R^2s of 0.44 for the explanation of the ratings of individual participants, and 0.73 and 0.69 for the explanation of the average ratings of the musicians and the non-musicians, respectively. These values may even be increased when the perception of duration intervals and loudness levels as discussed in the introduction are taken into account.

Important for future research seems to be the low reliability of correlation to predict the subjective distance between performances and the importance of absolute measures of tempo and of tempo times loudness. It is up to future research to confirm the difference in evaluation of loudness and tempo with the perception of loudness being more abstract than tempo and to consider the conditions for which the interaction between tempo and loudness is stronger than the separate evaluation of the two dimensions. It is also up to future research to further investigate the representation of performance variations by listeners, while taking into account the alternatives that have been explored in this paper.

To conclude, in this study, measured differences in tempo and loudness were fairly well able to predict the subjective distance between performances and seemed therefore reliable to represent a considerable part of the performance characteristics. The majority of the variance was explained for the Chopin fragment and a little less for the Mozart fragments. The parameters most responsible for this explanation for

both musicians and non-musicians were global and local tempo, global and local tempo times loudness, and the standard deviation of loudness. Interviews with the participants suggested that full explanation would have needed among other things the inclusion of articulation and a relation to musical structure.

Acknowledgements

This study was realized with financial support of the Mosart IHP-Network, HPRN-CT-2000-00115 and the START program of the Austrian Federal Ministry for Education, Science and Culture (Grant no. Y99-INF). 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 Simon Dixon, Werner Goebl, Josef Linschinger, Elias Pampalk, Asmir Tobudic and Gerhard Widmer to provide the perfect performance data and the perfect environment to do this study. I also thank them for their helpful comments.

References

Bengtsson, I., & Gabrielsson, A. (1983). Analysis and synthesis of musical rhythm. In J. Sundberg (ed.) *Studies of music performance.* Stockholm: Royal Swedish Academy of Music, pp. 76-181.

Clarke, E. F., & Krumhansl, C. L. (1990). Perceiving musical time. *Music Perception, 7,* 213-252.

Clarke, E. F. (1993). Imitating and evaluating real and transformed musical performances. *Music Perception, 10* (3), 317-341.

Dixon, S. (2001a). Automatic Extraction of Tempo and Beat from Expressive Performances. *Journal of New Music Research, 30* (1), pp 39-58.

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

Friberg, A., Frydén, L., Bodin, L. G., & Sundberg J. (1991). Performance Rules for Computer-Controlled Contemporary Keyboard Music. *Computer Music Journal, 15* (2), 49-55.

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

Gabrielsson, A. (1988). Timing in music performance and its relation to music experience. In J. Sloboda (ed.), *Generative processes in music. The psychology of performance, improvisation, and composition.* Oxford: Clarendon Press, pp. 27-51.

Gabrielsson, A., & Juslin, P. N. (1996). Emotional expression in music performance: Between the performer's intention and the listener's experience. *Psychology of Music, 24* (1), 68-91.

Gjerdingen, R. O. (1988). Shape and motion in the microstructure of song. *Music Pereption, 6* (1), 35-64.

Jones, M. R. (1987). Dynamic pattern structure in music: Recent theory and research. *Perception & psychophysics, 41* (6), 631-634.

Kendall, R. A., & Carterette, E. C. (1990). The Communication of Musical Expression. *Music Perception, 8* (2), 129-164.

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

Langner, J., & Goebl, W. (2003). Visualizing expressive performance in tempo-loudness space. *Computer Music Journal*, in press.

Lerdahl, F. & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, Massachusetts: MIT Press.

Nakajima, Y. (1987). A model of empty duration perception. *Perception, 16,* 485-520.

Nakamura, T. (1987). The communication of dynamics between musicians and listeners through musical performance. *Perception & psychophysics, 41,* 525-533.

Palmer, C. (1989). Mapping musical thought to musical performance. *Journal of Experimental Psychology: Human Perception and Performance, 15,* (12), 331-346.

Palmer, C. (1996). On the assignment of structure in music performance. *Music Perception, 14* (1), 23-56.

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

Penel, A. & Drake, C. (1999). Seeking "one" explanation for expressive timing. In S. W. Yi (Ed.), *Music, Mind & Science.* Seoul: Seoul University Press, pp. 271-297.

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception, 11* (4), 409-464.

Repp, B. H. (1990). Patterns of expressive timing in performances of a Beethoven minuet by 19 famous pianists. *Journal of the Acoustical Society of America, 88,* 622-641.

Repp, B. H. (1992a). Diversity and commonality in music performance - an analysis of timing microstructure in Schumann's Traumerei. *Journal of the Acoustical Society of America, 92* (5), 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. (1994a). On determining the basic tempo of an expressive music performance. *Psychology of Music, 22,* 157-167.

Repp, B. H. (1994b). Relational invariance of expressive microstructure across global tempo changes in music performance: An exploratory study. *Psychological research, 56,* 269-284.

Repp, B. H. (1998). Obligatory 'expectations' of expressive timing induced by perception of musical structure. *Psychological Research, 61* (1)*,* 33-43.

Repp, B. H. (2000). Pattern typicality and dimensional interactions in pianists' imitation of expressive timing and dynamics. *Music Perception, 18* (2), 173-211.

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

Sundberg, J., Friberg, A., & Frydén, L. (1989). Rules for Automated Performances of Ensemble Music. *Contemporary Music Review, 3,* 89-109.

Sundberg, J., Friberg, A. & Frydén, L. (1991). Threshold and preference Quantities of Rules for Music Performance, *Music Perception, 9* (1), pp. 71-92.

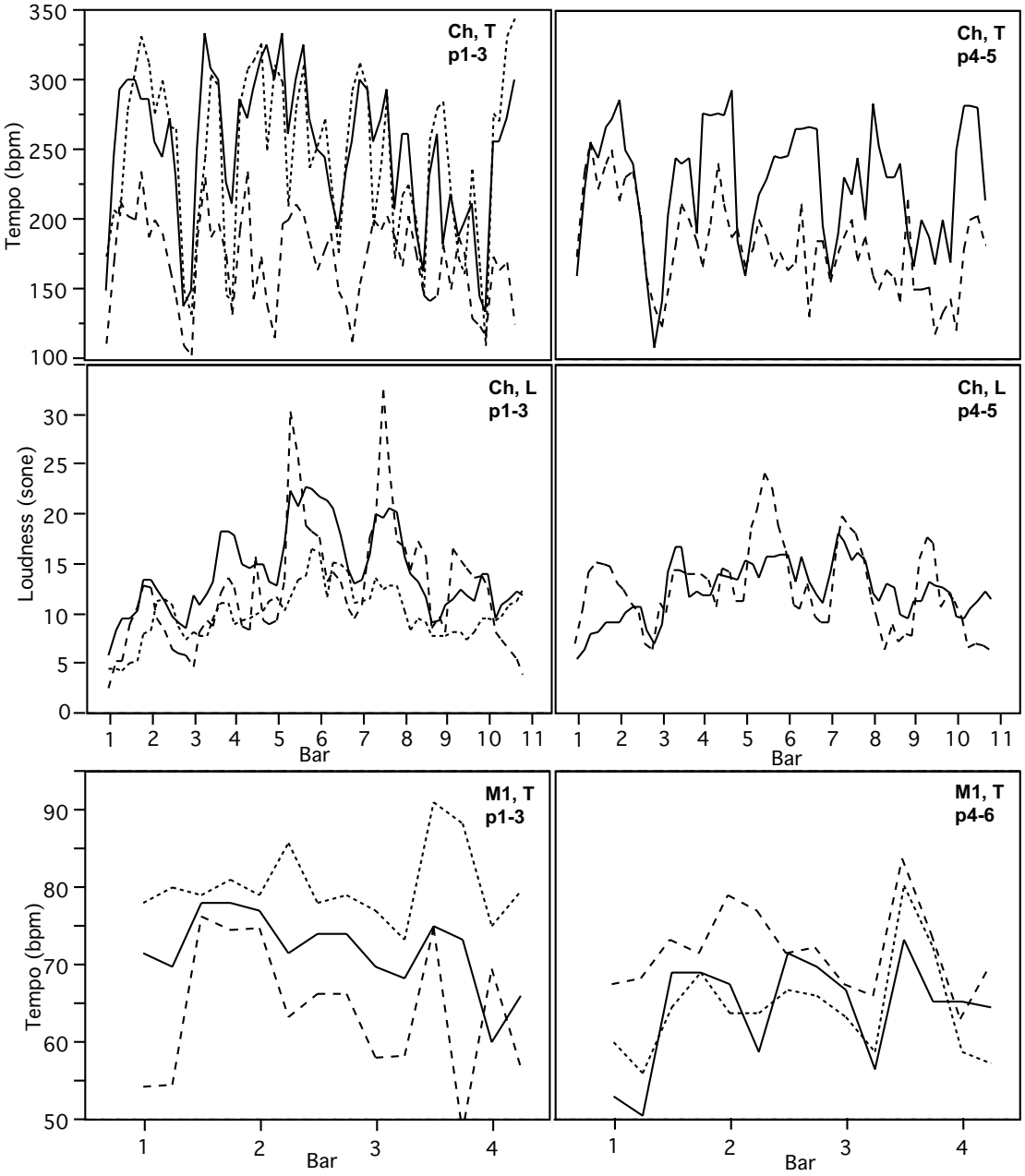
Thompson, W. F., Sundberg, J., Friberg, A., & Frydén, L. (1989). The use of rules for expression in the performance of melodies. *Psychology of Music, 17,* 63-82.

Timmers, R., Ashley, R., Desain, P., & Heijink, H (2000). The influence of musical context on tempo rubato. *Journal of New Music Research , 29* (2), 131-158.

Timmers, R. (2003). On the contextual appropriateness of expression. *Music Perception, 20* (3), 225-240.

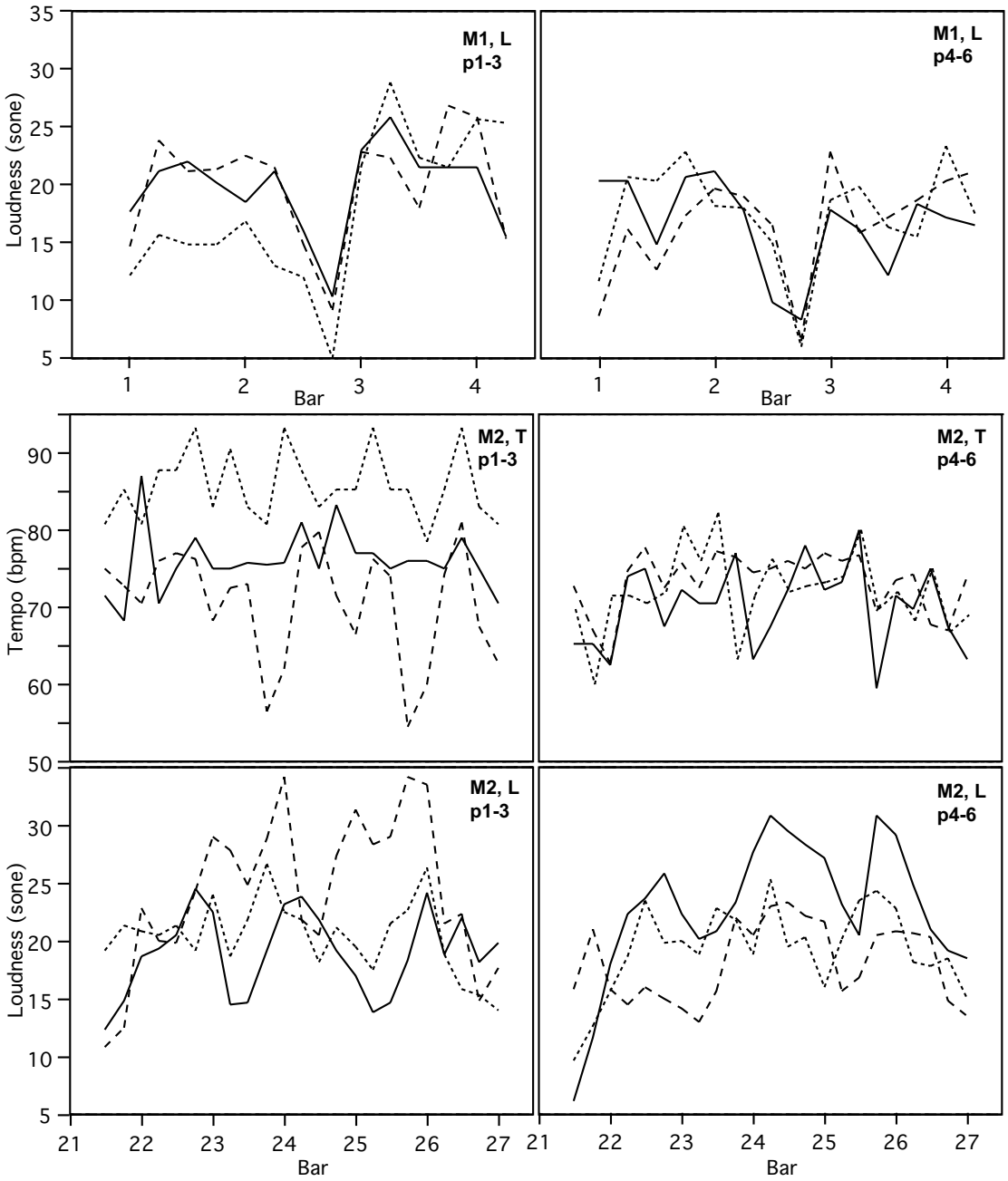
Widmer, G. (2003). Studying a creative act with computers: Music performance studies with automated discovery methods. *Musicae Scientiae* (in press).

Zanon, P., & Widmer, G. (2003). Learning to recognize famous pianists with machine learning techniques. In *Proceedings of SMAC03* (pp. 581-584), Stockholm, Sweden.



Appendix 1: Local tempo (in beats per minute) and local loudness (in sone) per beat for the three fragments. In the right corner, the fragment, the component (tempo or loudness) and the group of pianists of the graph are indicated. The solid line indicates the tempo or loudness of p1 (left panel) or p4 (right panel). The broken line indicates

these measures for p2 (left panel) or p5 (right panel). And the dotted line indicates these measures for p3 (left panel) or p6 (right panel). Bar numbers are indicated below the graphs.



Appendix 1 (continued)