02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 365

COMPUTATIONAL MODELING OF MUSIC COGNITION: A CASE STUDY ON MODEL SELECTION

HENKJAN HONING
Music Cognition Group, University of Amsterdam

WHILE THE MOST COMMON WAY of evaluating a computational model is to see whether it shows a good fit with the empirical data, recent literature on theory testing and model selection criticizes the assumption that this is actually strong evidence for the validity of a model. This article presents a case study from music cognition (modeling the *ritardandi* in music performance) and compares two families of computational models (kinematic and perceptual) using three different model selection criteria: goodness-of-fit, model simplicity, and the degree of surprise in the predictions. In the light of what counts as strong evidence for a model's validity—namely that it makes limited range, nonsmooth, and relatively surprising predictions—the perception-based model is preferred over the kinematic model.

Received October 11, 2004, accepted November 7, 2005

Meeting a friend in the corridor, Wittgenstein said: "Tell me, why do people always say it was natural *for men to assume that the sun went round the earth rather than that the earth was rotating?" His friend said: "Well, obviously, because it just* looks *as if the sun is going round the earth." To which the philosopher replied, "Well, what would it have looked like if it had looked as if the earth was rotating?"*

Tom Stoppard, *Jumpers*, 1972

HOW SHOULD WE SELECT among computational models of cognition? This question has recently attracted much discussion (Pitt, Myung, & Zhang, 2002; Roberts & Pashler, 2000, 2002; Rodgers & Rowe, 2002). While the most common way of evaluating a computational model is to see whether it shows a good fit with the empirical data, the discussion addresses problems that might arise with the assumption that this is actually strong evidence for the validity of a model. Some authors consider a fit between a theory and the empirical observations a necessary starting point, but clearly not the end point of model selection or verification (e.g., Desain, Honing, Van Thienen, & Windsor, 1998; Jacobs & Grainger, 1994; Rodgers & Rowe, 2002). Others suggest alternatives to a goodness-of-fit (GOF) measure, such as preferring the simplest model, in terms of both its functional form and the number of free parameters (e.g., Pitt & Myung, 2002; Pitt, Myung, & Zhang, 2002). Yet others have indicated a preference for theories that predict an empirical phenomenon that was least expected, as they consider a good fit to be of less relevance or even misleading (e.g., Roberts & Pashler, 2000).

It comes as no surprise that models stated in computational form are now subject to this discussion. One of the advantages that computational models of cognition have over alternative types of theories (e.g., verbal theories) is that the former are open to direct and immediate test (Longuet-Higgins, 1987) and allow, in principle, for easier evaluation, verification, or falsification. However, the aim of this article is not to add to this lively debate in a philosophical or methodological sense. Instead, the focus is on a specific problem from music cognition, that is, modeling *ritardandi* in music performance. It is a case study on how one can select between two computational models, informed by the methodological discussion mentioned above.

This article compares two families of computational models. The first takes a kinematic approach to the modeling of expressive timing in music performance. These models focus on commonality, that is, on the timing patterns that are commonly found in music performance and on how they conform to the laws of physical motion (see Honing, 2003; Shove & Repp, 1995). Here, this approach is contrasted with a perceptual approach. Rhythm perception models predict constraints on the use of timing and tempo in music performance. As such, this approach focuses on diversity: These models predict the degree of expressive freedom a performer has in the interpretation of a

Music Perception VOLUME 23, ISSUE 5, PP. 365-376, ISSN 0730-7829, ELECTRONIC ISSN 1533-8312 © 2006 BY THE REGENTS OF THE UNIVERSITY OF CALIFORNIA. ALL RIGHTS RESERVED. PLEASE DIRECT ALL REQUESTS FOR PERMISSION TO PHOTOCOPY OR REPRODUCE ARTICLE CONTENT THROUGH THE UNIVERSITY OF CALIFORNIA PRESS'S RIGHTS AND PERMISSIONS WEBSITE AT WWW.UCPRESS.EDU/JOURNALS/RIGHTS.HTM

02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 366

rhythmic fragment before it is "misinterpreted" by the listener as a different rhythm (see Clarke, 1999; Desain & Honing, 2003; Honing, 2005).

In this article, the two approaches are compared using three different model selection criteria, namely goodness-of-fit, parsimony, and the surprisingness of the predictions. However, before discussing further these issues in model selection, I shall elaborate the domain (music cognition), the problem (modeling expressive timing in music performance), and two approaches in modeling it (kinematic and perceptual).

Music Cognition and Computational Modeling

In recent decades, computational modeling has become a well-established research method in many fields (Fodor, 2000; Pylyshyn, 1984), including music cognition (Desain, Honing, Van Thienen, & Windsor, 1998). To characterize the current state of affairs in music research, one can distinguish between (at least) two approaches to computational modeling. One approach aims at modeling musical knowledge. These are models originating from music theory in which a thorough formalization contributes to an understanding of the theory itself, its predictions, and its scope (e.g., Lerdahl & Jackendoff, 1983; Narmour, 1992). The other approach aims at constructing theories of music cognition. Here, the objective is to understand music perception and music performance by formalizing the mental processes involved in listening to and in performing music (Clarke, 1999; Gabrielsson, 1999). Both approaches have different aims and can be seen as being complementary. Music cognition is the domain of the current article, which discusses model selection in the context of the computational modeling of the final *ritardandi* in music performance.

Modeling the Final Ritard

A considerable number of theories on the use of expressive timing in music performance make predictions on the final *ritardandi* (or "final ritard"), that is, the typical slowing down at the end of a music performance, especially in music from the Western Baroque and Romantic periods (Hudson, 1996). This characteristic slowing down can also be observed in, for instance, Javanese gamelan music and some pop and jazz genres. Together with the "fade-out," it is one of the most common ways of marking the ending a piece of music in Western culture. Several approaches have been suggested to explain the typical *ritardandi* found in music performance, most notably the relation between these timing patterns and physical motion (e.g., Shove & Repp, 1995; Truslit, 1938). However, also direct physiological (e.g., Todd, 1999) and perceptual explanations (in the Gibsonian sense; Clarke, 2001) have been proposed. Other theories focus more on the metaphorical relation between music and motion, investigating how far this is facilitated by cognition (Eitan & Granot, 2005; Gjerdingen, 1994). However, this is not the place to fully discuss the relation between music and motion. This article concentrates on two computational approaches to modeling final *ritardandi*: the kinematic model and the perception-based model.

Kinematic Model

An important contribution to the modeling of expressive timing is made by a family of computational theories, namely kinematic models (Honing, 2003). They make an explicit relation between the laws of physical motion in the real world and expressive timing in music performance (Epstein, 1994; Feldman, Epstein, & Richards 1992; Friberg & Sundberg, 1999; Kronman & Sundberg, 1987; Longuet-Higgins & Lisle, 1989; Sundberg & Verrillo, 1980; Todd, 1985; Todd, 1992; Todd, 1995). Most of this research suggests that musicians, in using tempo and timing as an expressive device, allude to physical motion, even so far as modeling it as a force that causes the tempo to "push forward" or "hold back" (Gabrielsson, 1999). In the case of the final ritard, the analogy is made with deceleration in human motion, as in walking or running. Such a deceleration pattern can be described by a well-known model (from elementary mechanics) of velocity (v) as a function of time (t):

$$v(t) = u + at \tag{1}$$

where u is the initial tempo and a the acceleration factor (deceleration when a is constrained to be less than 1). This function can be generalized and expressed in terms of score position x, resulting in a model in which normalized tempo v (normalized with respect to the initial or pre-*ritardando* tempo) is defined as a function of normalized score position x (normalized with respect to the length of the ritard), with q for curvature and w determining slope (Friberg & Sundberg, 1999; see Figure 1):

$$v(x) = [1 + (w^q - 1)x]^{1/q} \tag{2}$$

In this generalized model, w is defined over the interval $< 0, 1]$, and q over the interval $< 0, \infty]$. Furthermore, the curvature of q is dependent on w. When w

02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 367

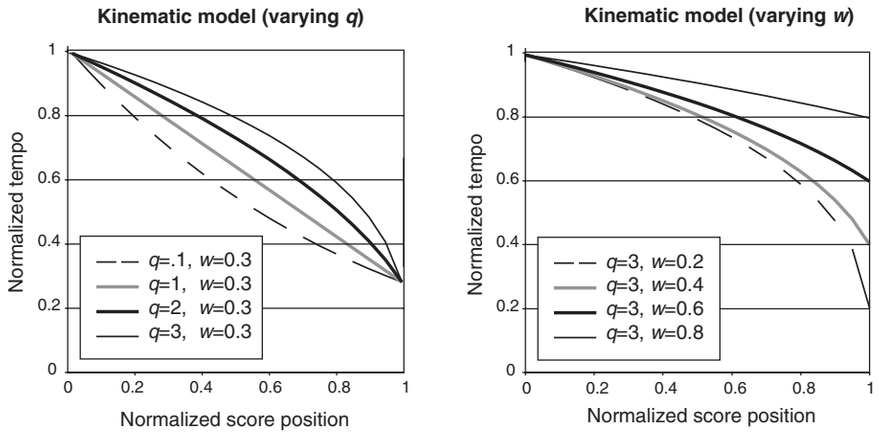

FIG. 1. Predictions of the final ritard made by the kinematic model. The *x* axis indicates the normalized score position, the *y* axis the normalized tempo. The left panel shows some of the predictions that can be made by varying *q* for curvature (with $q = 2$ being a model of constant braking force, and $q = 3$ a model of constant breaking power); the right panel shows some of the predictions that can be obtained by varying *w* for slope.

approaches 1, curvature is reduced (see Figure 1). Two instances of this model are most commonly found in the literature: a model of constant braking force ($q = 2$, Epstein, 1994; Kronman & Sundberg, 1987; Longuet-Higgins & Lisle, 1989; Todd, 1992), and a model of constant braking power ($q = 3$, Friberg & Sundberg, 1999). Both were found to be good predictors of performance and perception data (Friberg & Sundberg, 1999). Furthermore, the latter model ($q = 3$) was shown to be similar to the way dancers stop running. Alternatives, such as a model of duration (IOI instead of tempo) of score position (Repp, 1992; Todd, 1985), were shown to fit the empirical data less well.

The rationale for the kinematic model is that it models types of movement with which the listener is quite familiar, and consequently facilitates the prediction of the actual end, the final stop of the performance.

Perception-Based Model

An alternative to the kinematic approach is based on computational models of rhythm perception: It is referred to as a "perception-based model" (Honing, 2005). It consists of two components. The first component, a model of perceived regularity (or "tempo tracker"), tracks the perceived tempo of the performance using an adaptive oscillator (Large & Kolen, 1994; McAuley, 1995; Toiviainen, 1998). The output of a tempo tracker can be described by:

$$o(t) = 1 + tanh\left[\gamma(\cos 2\pi\phi(t) - 1)\right] \quad (3)$$

$$\phi(t) = \frac{t - t_x}{p}, t_x - \frac{p}{2} \leq t < t_x + \frac{p}{2} \quad (4)$$

where p is the period, t_x the time at which an event is expected, and γ the "temporal receptive field," that is, the area within which the oscillator can be changed (a higher value being a smaller temporal receptive field). At the point in time that an event occurs (referred to as t_*), the period and phase are adapted according to:

$$\Delta p = \eta_p \frac{p}{2\pi} sech^2\gamma(\cos 2\pi\phi(t^*) - 1)\sin 2\pi\phi(t^*) \quad (5)$$

$$\Delta t_x = \eta_\phi \frac{p}{2\pi} sech^2\gamma(\cos 2\pi\phi(t^*) - 1)\sin 2\pi\phi(t^*) \quad (6)$$

where η_p and η_ϕ are the coupling-strength parameters for period and phase tracking. If an event occurs within the temporal receptive field, but before t_x (i.e., it occurs earlier than expected), the period is shortened. If an event occurs outside the temporal receptive field, the period remains unchanged.

The second component, a model of rhythmic categorization (or "quantizer"), takes the residue—the timing pattern after a tempo interpretation—and predicts the perceived duration category (e.g., an eighth note as opposed to a sixteenth note in common music notation). Three quantizers are considered. Each takes as input an inter-onset interval (IOI) pattern and returns a categorized version of it. The Dannenberg and Mont-Reynaud (1987) model does this using techniques from control theory, the Longuet-Higgins (1987) model uses AI-based techniques, and the Desain and Honing

02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 368

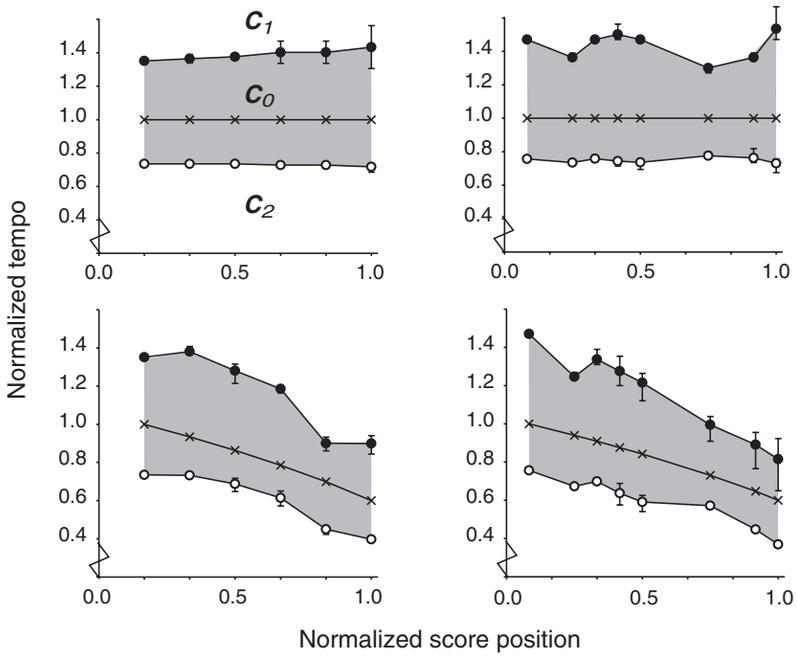

FIG. 2. The influence of rhythmic structure (isochronous vs. nonisochronous; left vs. right column) and curvature (top vs. bottom row) on the predicted degree of expressive freedom (gray area) in a final ritard, as predicted by the perception-based model. Crosses indicate the input data; circles mark the average upper and lower perceptual category boundaries. The bars indicate the minimum and maximum values predicted by different quantizers[2] (adapted from Honing, 2005).

(1989) model takes a connectionist approach. Each of these models has two parameters. They are all used with their default values. Hence, in this article I shall consider the categorization component as parameter-free. For details on the formalization of these models, the reader is referred to Desain and Honing (1992).

The perception-based model was evaluated on artificial data (Honing, 2005), an example of which is given in Figure 2. It shows the results of simulations of the model on data that were varied for rhythm (isochronous and nonisochronous durations; see the columns in Figure 2) and curvature (without and with ritard; see the rows in Figure 2). The gray areas indicate the degree of expressive timing (tempo change or variance) that a performed note can exhibit before being categorized as a different duration category. It was shown that the rhythmic structure constrains the expressive freedom: More complex rhythmic patterns restrict the degree of slowing down; the narrower the gray area, the less freedom a performer has in varying the duration of that particular note (as predicted by the model).

In more formal terms, when a certain inter-onset interval (IOI) is categorized as C (e.g., C_0 in Figure 2a) by the categorization component, the upper border (filled circles) indicates the tempo boundary at which an input duration will be categorized as $C\times3/4$ (C_1 in Figure 2a), and the lower border (open circles) the boundary at which it will be categorized as $C\times4/3$ (C_2 in Figure 2a).[1] The error bars indicate the maximum and minimum prediction of the rhythmic category boundary over the three quantization models (Dannenberg & Mont-Reynaud, 1987; Desain & Honing, 1989; Longuet-Higgins, 1987).[2] Note that just one possible category boundary is shown here.

The rationale for the perception-based model is that, in general, a performer would like the listener to recognize the original, notated rhythm. The model makes precise predictions about when a rhythm performed with some tempo and timing variations will still be recognizable as such by the listener (e.g., those

[1]A log scale will make the areas shown in Figure 2 symmetrical. A linear scale is used for easier comparison with the existing literature.

[2]As it turned out, it is not too important which models are used for the categorization component: Different combinations of quantizers and a tempo tracker gave roughly similar behavior (Honing, 2005). See error bars in Figure 2.

02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 369

performances that stay within the gray areas shown in Figure 2). It also predicts when the perceived rhythmical structure will change or break down because there was too much tempo change (e.g., a performance that crosses a category boundary).[3] So, in short, the perception-based model does not predict the specific shape of a *ritardando*, but the perceptual boundaries between which *ritardandi* are expected to occur.

General Differences between the Two Approaches

Before starting a more detailed comparison, I shall first point out some general differences between the two models.

As mentioned, the main difference between the two approaches is the focus on either commonality (What is the shape of a "prototypical" ritard?) or diversity (What is the variability observed?). However, both models can make predictions regarding both questions. The perception-based model, while characterized as a model of diversity, can make predictions on the shape of the final ritard by defining it as the shape that can best be tempo-tracked using the perceived regularity component of the perception-based model. For the kinematic model, while characterized as a model of commonality, one can define the range of possible ritards as those that can be successfully fitted. As such, this model can make predictions on the variability of ritards.

Another important difference between the two models is the type of input they use. The kinematic model assumes the score position to be known (x in Eq. 2), whereas the perception-based model takes real performance data as input (t in Eq. 3). So while the kinematic model has to be informed of score information, the perception-based model derives this categorical information from the input data by using the rhythmic categorization component. This makes the perception-based model more complex, but arguably also more perceptually realistic. This has implications for the comparison of both models on empirical data. It informs the kinematic model of score information (i.e., duration categories) such that it can distinguish between tempo and timing variations, while the perception-based model has to try and separate these (using, respectively, period and phase adjustments). However, since there seems to be no elegant way to resolve this issue, I shall ignore this difference here. Hence the complexity of the rhythmic categorization component in the analyses below (see under "Measure of simplicity") is not considered.

And, lastly, although the output of the two models is also different, both can easily be interpreted as a tempo prediction. For the kinematic model, this is simply the result of Eq. 2, while for the perception-based model this is the reciprocal of the period, the output of the oscillator (Eq. 3). The other differences between the models will be discussed in more detail below.

Theory Testing and Model Selection

I shall now compare the two approaches (kinematic and perception-based) using the perspectives on model selection mentioned in the introduction, namely (a) how well the model fits the empirical data (measure of good fit), which is the most parsimonious model (measure of simplicity), and (b) how surprising the predictions are (what could one expect). For brevity, I shall henceforth refer to the kinematic model as the "K model" and to the perception-based model as the "P model."

Measure of Good Fit

Goodness of fit (GOF) is the most common way of testing the validity of a theory. It indicates the precision with which a model fits a particular sample of observed data. The predictions of the model are compared with the observed data and the discrepancy between the two is measured using an error measure.

I shall use GOF to compare the two models on measurements of final *ritardandi* taken from Friberg and Sundberg (1999). This set—the "F&S99" set—consists of twelve harpsichord performances of compositions by J. S. Bach. Sundberg and Verrillo (1980) used slightly larger set of twenty[4] performances that includes the F&S99 set; these are referred to as the "S&V80" set, and will be used for comparison.

To obtain the best fit, for the K model the q and w parameters were varied (see Eq. 2), as was the v_{offset} parameter, which adds a constant to the model. This third parameter was added to eliminate a propagating effect of sometimes misfitting the first normalized

[3]Of course, some performers challenge this. For instance, the late Glenn Gould took tempi or used timing that would alter the actual notated rhythm on occasion. However, it could be argued that the mere existence of such perceptual boundary, and just crossing it, makes an interpretation ambiguous and interesting (cf. Desain & Honing, 2003).

[4]Sundberg and Verillo (1980) originally reported on 24 measurements. Of this set, only 20 measurements were made available to the author.

02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 370

TABLE 1. Results for the K (Kinematic) and P (Perception-Based) Model as Fitted on the F&S99 Dataset.

			K model				P model			
Music examples	Id	Notes	q	w	v_{offset}	r^2	γ	η_Φ	η_p	r^2
W. clav. I Prel. 1	WIP	10	2.4	.32	.010	.98	0.1	0.7	1.0	.85
W. clav. II Prel. 1	WP1	8	2.1	.50	.000	.98	0.0	1.0	1.0	.99
W. clav. II Prel. 2	WP2	7	2.5	.51	.030	.97	0.0	0.6	1.0	.99
W. clav. II Fug. 3	WF3	6	1.1	.48	−.020	.97	0.0	0.7	1.0	.95
W. clav. II Fug. 5	WF5A	7	4.3	.51	.010	.99	0.0	0.2	1.0	.93
W. clav. II Fug. 5	WF5B	8	2.6	.38	−.020	.98	0.2	0.0	0.8	.89
Eng. Suite 1 All.	E1A	6	2.0	.45	−.030	.98	0.0	0.4	1.0	.88
Eng. Suite 2 All.	E2A	11	3.8	.37	.020	.98	0.0	1.0	1.0	.85
Fr. Suite 4 Cour.	F4C	6	4.1	.50	.020	.99	0.0	0.5	1.0	.89
Fr. Suite 6 All.	F6A	7	2.3	.44	.020	.98	0.0	0.6	1.0	.96
Fr. Suite 6 Cour.	F6C	7	5.0	.46	−.010	.99	0.0	0.0	0.8	.76
It. Conc. Mvt. 3	IC3	7	1.2	.34	−.010	.97	0.1	0.9	1.0	.85
		Mean	2.7	.44	.002	.98	0.03	0.55	0.97	.90
		SD	1.2	.07	.019	.01	0.07	0.35	0.08	.07

TABLE 2. Results for Datasets F&S99 and S&V80 with (+) and without (−) the Last IOI.

		K model		P model		
Set	Last	Mean r^2	*SD*	Mean r^2	*SD*	
F&S99	+	.98	.01	.90	.07	*
F&S99	−	.89	.08	.97	.04	*
S&V80	+	.95	.05	.91	.07	n.s.
S&V80	−	.86	.09	.97	.05	**

* $p < .01$. ** $p < .001$.

value (Friberg & Sundberg, 1999:1478). For the P model, the γ , η_p, and η_ϕ parameters were varied (see Eqs. 3, 5, and 6).[5] Table 1 shows the results of fitting the two models to the F&S99 dataset minimizing the root mean squared error (RMSE) between each model and the observed data. Note that the results for the K model replicate those presented in Table IV of Friberg and Sundberg (1999, p. 1478).

From these results it can be concluded that both models actually fit the F&S99 dataset quite well. The K model does slightly better ($r^2 = .98$) than the P model ($r^2 = .90$). A t-test comparing the mean correlations indicates that this difference is significant ($p < 0.1$).

These results could, however, be affected by the specific selection criteria used in determining which and what part of the measured performances to use as a final ritard. For instance, only *ritardandi* with "smooth shapes" were selected by Friberg and Sundberg (1999, p. 1478), resulting in a set of twelve performances, instead of the twenty performances of the original S&V80 dataset. Furthermore, at least for one performance the last IOI was removed (WP1). In this performance, the penultimate IOI was longer than the last IOI: Apparently, the performance speeded up at the end of the ritard.[6] Table 2 shows the influence of the specific data set used (F&S99 or S&V80) and the effect of a systematic inclusion or exclusion of the last IOI (note that the first row in Table 2 is the same as the means and standard deviations reported in Table 1).

Table 2 shows that the specific dataset chosen has an effect on the result: For the F&S99 set the difference in fit is significant, while for the S&V80 set the difference is nonsignificant (compare the first with the third row in Table 2). Furthermore, systematically removing the last IOI from both datasets has a large effect on the results. When it is excluded, the P model is significantly better, and when it is included the K model makes better overall fits (compare the first and the second row, and the third and the fourth row in Table 2).

Based on these results, we cannot conclude that one model fits the empirical data better than the other. In

[5]For the other parameters of the P model, the default settings were used (i.e., the period of the oscillator is initially set to the first observed interval in the input, and the initial phase is set to zero). Furthermore, since the global tempo (absolute IOIs) is not considered relevant to the K model, for the P model the fit was selected for the tempo that gave the best results.

[6]This special behavior of the last IOI in the context of the final ritard was also observed by Repp (1992). Sometimes performers give the last IOI the same length as the penultimate IOI, but they mostly give it a much longer or even a shorter length (e.g., WP1).

02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 371

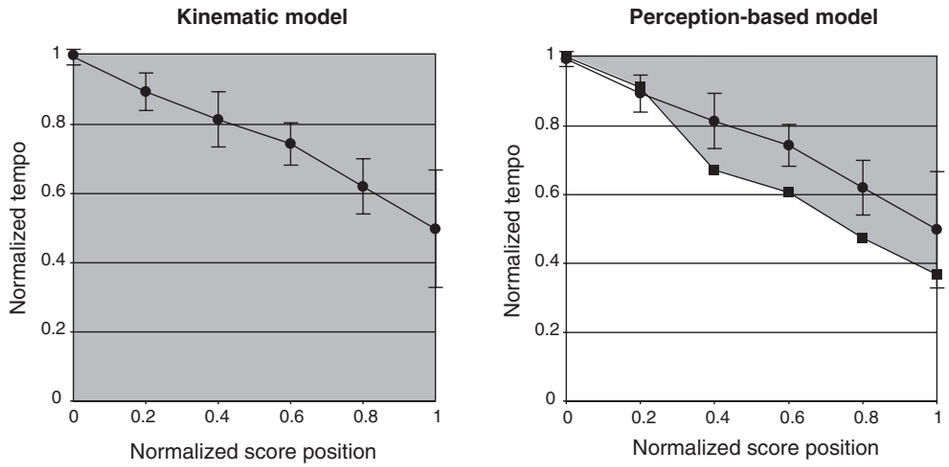

FIG. 3. The response area (gray) for the K model (left) and the P model (right). Circles indicate the measurements of the F&S99 dataset (closed circles indicate mean tempo; error bars indicate ± one standard deviation).

addition, we have to realize that the K model has an advantage in that it has access to score information (see under "General Differences between the Two Approaches"). However, even if we were to restrict our evidence to the results shown in Table 1, we could not select one model over another. The reason for this is that such measures as RMSE or percentage of variance accounted for (PVAF) only assess fit. These measures are not able to distinguish between variations in the data caused by noise and those that the model was designed to capture. Therefore, several alternative model selection criteria were proposed (for an overview, see Pitt & Myung, 2002), some of which will be discussed below.

Measure of Simplicity

How can we select between models that fit a particular sample of observed data equally well? One way of doing so is to relate the complexity of a model to the degree of success in making a good fit. Complexity is a property of a model that enables it to fit diverse patterns of data; it is the flexibility of a model (Pitt & Myung, 2002). Dimensions of complexity that can be evaluated are, for example, the functional form—the way in which parameters and data are combined in a model's equation (e.g., $y = ax$ and $y = a + x$ have the same number of parameters but different functional forms)—and the number of free parameters of a model (e.g., in $y = ax$, a is the free parameter) that can be adjusted to improve a model's fit to the data. As such, it provides a measure of the flexibility of a model (the GOF measure used in the previous section does not consider any dimension of complexity).

With regard to the number of free parameters, it is clear that we cannot make a selection between the two models under discussion, simply because both use the same number of parameters. But even if one model were to have fewer parameters than the other, this would be too crude a comparison since the functional form of both models is not considered. While a model might have fewer free parameters, it could well have a functional form that allows much greater flexibility in making fits; in fact, it might even suffer from over-fitting the data. Therefore, we need a way of showing the flexibility—that is, the range of predictions—a model can make.

I shall investigate this by plotting the response area of a model. This graphical representation shows the range of predictions a model can make.[7] The larger the response area, the more flexible the model.

Figure 3 shows the response area for both models as well as the *ritardandi* in the F&S99 dataset. Figure 3a shows the response area containing all the predictions that can be made by the K model, in fact the full square (all possible ritards). For the P model (see Figure 3b), the response area contains all the ritards that can be successfully tempo-tracked by the model ($r^2 > .99$), a

[7]For alternatives in visualizing the complexity of a model, see Pitt, Myung, and Zhang (2002).

02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 372

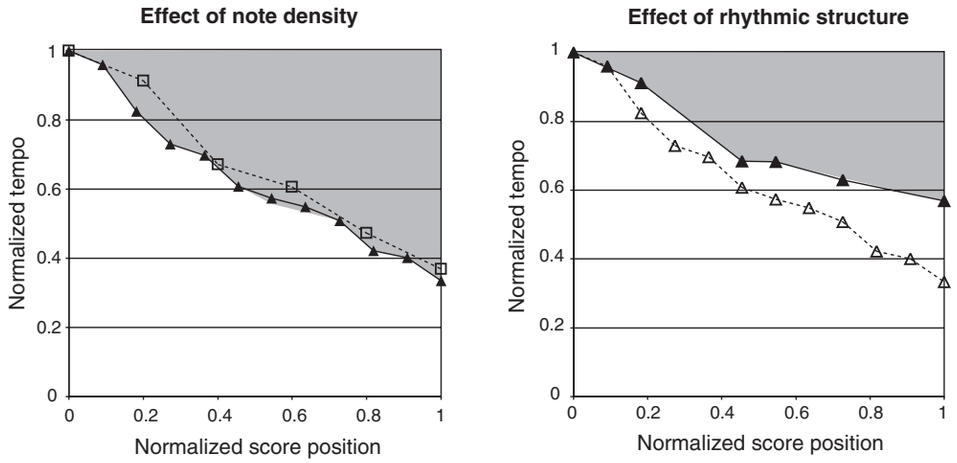

FIG. 4. The response area (gray) for the P model modulated by note density (left) and rhythmic structure (right). The left panel show the influence of note density for a rhythm of six IOIs (open squares) and twelve IOIs (closed triangles). The right panel shows the influence of isochronous data (open triangles) versus nonisochronous data (closed triangles).

much smaller portion of the full square. So, in conclusion, the P model exhibits much less flexibility than the K model, despite the larger complexity in terms of the functional form of the former model (cf. equations).

Another aspect that can be analyzed using the response area visualization is the influence of the structural characteristics of the input data. In Honing (2005) it was shown that the K model is insensitive to note density and rhythmic structure: The predicted shape of the ritard is not affected by these factors. However, these factors were shown to have an effect on the predictions made by the P model. Using the response area visualization, this can also be shown in the current context.

In Figure 4a the response area for the P model is shown for six notes (marked with open squares, repeated from Figure 3b) and for an input twice as dense (i.e., 12 notes; marked with closed triangles). It shows that note density has an effect on the size of the response area: The more notes (per time unit), the larger the response area. This is in line with the idea that one would expect a ritard of many notes to allow for a deep *rubato*, while one of only a few notes is likely to be less deep (i.e., less slowing down), simply because there is less material with which to communicate the change of tempo to the listener (cf. Honing, 2005).

Furthermore, as is shown in Figure 4b, the rhythmic structure also has an effect on the response area of the P model. When the input pattern (marked with open triangles, repeated from Figure 4a) is grouped into notes of different duration (shown for rhythm 1-2-3-1-2-3 in Figure 4b; marked with closed triangles), the response area shrinks. This is in line with research in rhythmic categorization that showed that the expressive freedom in timing—the amount of timing that is allowed for the rhythm still to be interpreted as the same rhythmic category (i.e., the notated score)—is affected by the rhythmic structure. Simple rhythms can be expected to allow for more timing variation than more complex ones (cf. Honing, 2005).

In short, the P model shows less flexibility for rhythmically varied input than for isochronous input, and more flexibility with higher note density. Interestingly, this has no influence on the predictions made by the K model: For the latter model, these factors are irrelevant.

In conclusion, in addition to making roughly similar fits to the empirical data, the P model shows less flexibility than the K model does. As such, it becomes the preferred model. Still, we can wonder how surprising all this is in the context of the phenomenon modeled.

Element of Surprise

How surprising is the prediction of a slowing-down pattern in music performance when selected from a musical genre known for its use? What could we actually expect?

To give some structure to the notion of surprise, in Figure 5 a distinction is made between possible, plausible, and predicted observations of a final ritard. The total area of the square indicates the possible tempo values (e.g., a horizontal line would indicate a constant tempo, a vertical line an instant tempo change). However, the plausible values—the values one can expect to happen in the case of a slowing-down in tempo—are

02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 373

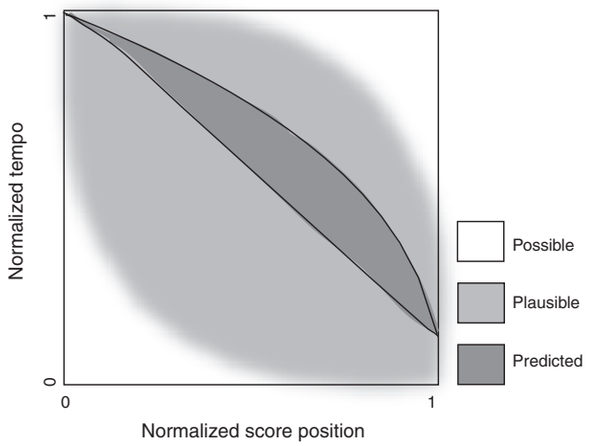

FIG. 5. Schematic diagram of possible, plausible, and predicted final ritards (see text for details).

roughly within the gray area (note, however, that this is a very loosely defined area, since there is no model available of what is and what is not perceived as a ritard).[8] Finally, the dark gray area indicates the predicted values of several settings of the K model (*w* set to .3 and *q* varied between 1 and 3).

For a model to be surprising, all predicted outcomes should be a small fraction of the possible—or, even better, the plausible—outcomes (see Figure 6). Only when few observations and precise predictions across all parameter values are made, is this substantial evidence for a model (Roberts & Pashler, 2002). A good fit in itself does not say much; what is more important is what the model rules out. This is characterized by the "forbidden zone" (Roberts & Sternberg, 1993), namely the outcomes that a model *cannot* predict. In this case study, these are the ritards that fall outside of the response area (cf. the white areas in Figures 3 and 4). So for the P model, since the response area is a small portion of the plausible outcomes, it has a relative large forbidden zone. And, since the model is less flexible, it is potentially easier to falsify. This is considered a strong aspect of a model (independent of whether empirical data might actually support this). By contrast, the K model can predict ritards in the whole space of possible ritards, hence it has no forbidden zone, and as such it is unclear what the K model predicts to be an unexpected or impossible shape of a ritard. As an example, in Figure 6 we should prefer B and D over A and C.

Furthermore, a model that predicts simple or smooth shapes is less surprising than one that predicts nonsmooth or complex shapes, because smooth and simple functions (as often used in psychology research) are likely on the basis of experience and are easily explained. "It is hard to think of a theory that would not produce a smooth function" (Roberts & Pashler, 2000). So in Figure 6, we should prefer C and D over A and B.

In conclusion, and in addition to preferring models that make a good fit and are the least complex, we should favor models that are relatively surprising—surprising in the sense that they make a limited range of predictions and predict nonsmooth shapes. All this puts the P model in a better position than the K model (as summarized in Figure 6).

Summary and Conclusion

In summary, it was shown that both the K (kinematic) and the P (perception-based) model could roughly fit the data and do so equally well, dependent of the specific dataset and selection criteria used. It was also shown that even if one model were to fit the data significantly better than the other, it would be impossible to select between these models on the basis of a goodness of fit (GOF) measure alone.

While the K model captures some common timing patterns in music performance (a "prototypical" ritard), this by itself is no strong evidence for the validity of such a model. Just because a certain model fits the empirical data well does not necessarily imply that the regularity one seeks to capture in the data is well approximated by the model (cf. epigraph).

Furthermore, while the shape of a final ritard might often resemble a cubic root function (a K model with $q = 3$), the P model predicts the shape of the final ritard to be modulated by several structural and temporal aspects of the music. Because these factors have an effect on the predictions made by the P model, but not on those made by the K model, the former is a stronger model (independent of whether empirical data would actually support this).

And, lastly, it could be argued that a model that focuses on the perceptual boundaries—that is, on the maximum diversity or expressive freedom predicted (what the theory predicts will *not* happen)—is far more selective than one that looks for common patterns in the final ritard. The P model makes explicit what cannot be explained (i.e., it has a clear "forbidden zone"; cf. the white area in Figure 4b), while the K model does

[8]While there are studies of just noticeable differences (JND) of single timing deviations in an isochronous sequence (e.g., Michon, 1964), one cannot make a tempo curve prediction from these.

02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 374

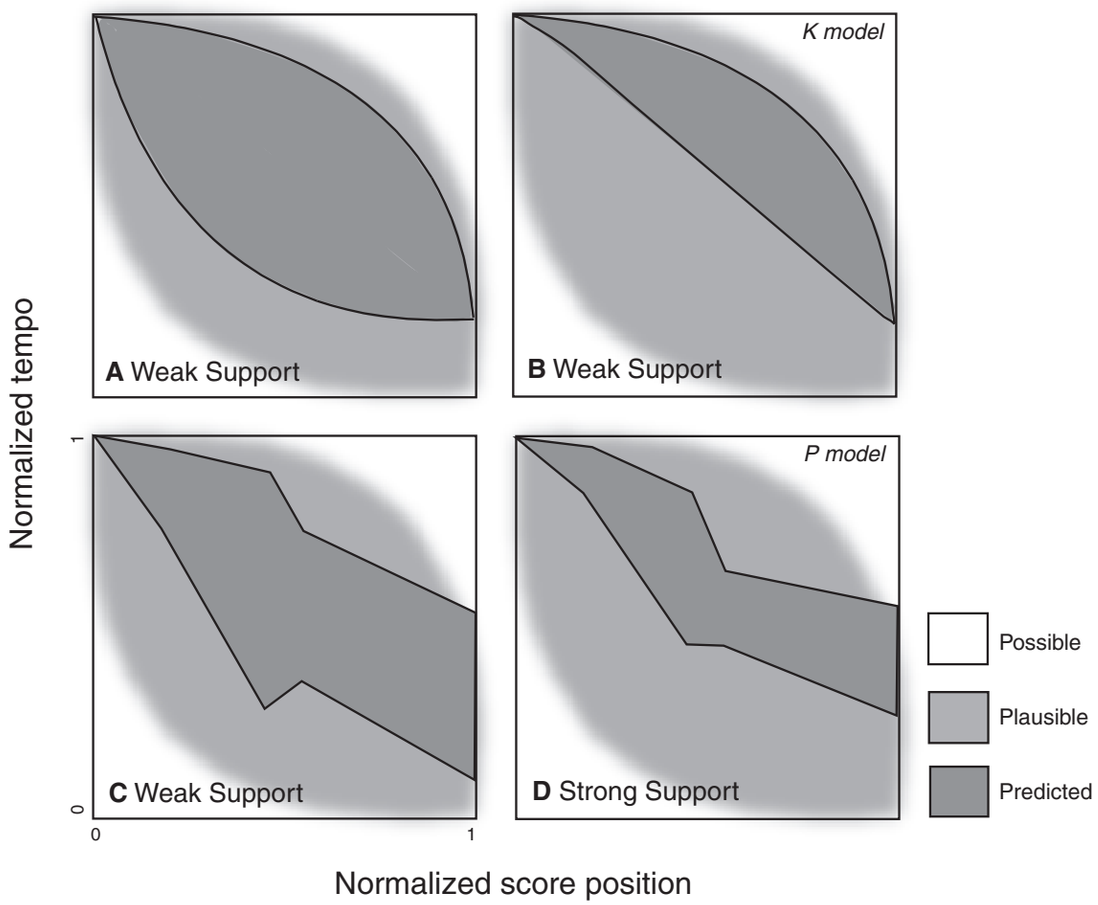

FIG. 6. Schematic diagram of strong and weak support for a model of the final ritard (see text for details).

not have this quality (cf. Figure 4a). The latter can fit a considerably larger number of shapes of ritards, and is hence more complex. A challenge for future research is to see whether the predictions made by the P model generally hold, that is, whether empirical data that is systematically varied for note density and rhythmic structure will stay outside the predicted forbidden zone. For now, and in the light of what counts as strong evidence for a model—namely making precise (constraint), nonsmooth, and relatively surprising predictions—the P model can be preferred over the K model. The general aim of this article was to show how current issues in model selection can be of use to the computational modeling of music cognition.

Author Note

Portions of this work were presented at the XXVII Annual Conference of the Cognitive Science Society (CogSci2005), University of Turin, Stresa, Italy (July 21–23, 2005) and the 8th International Conference on Music Perception & Cognition (ICMPC8), Northwestern University, Evanston (August 3–7, 2004). Part of this article was written at Northwestern University (Evanston/Chicago, IL) in the spring of 2004, while on a visiting professorship at the kind invitation of Richard Ashley (Music Cognition Program). The article has benefited from the insightful remarks made by Peter Pfordresher, Edward Large and two anonymous referees. I like to thank Anders Friberg and Johan Sundberg for generously sharing their measurements of final ritards.

Address Correspondence to: Henkjan Honing, Music Cognition Group, ILLC / Department of Musicology, University of Amsterdam, Nieuwe Doelenstraat 16, 1012 CP Amsterdam, The Netherlands. E-MAIL *honing@uva.nl*

02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 375

References

CLARKE, E. F. (1999). Rhythm and timing in music. In D. Deutsch (Ed.), *Psychology of music* (2nd ed., pp. 473–500). New York: Academic Press.

CLARKE, E. F. (2001). Meaning and the specification of motion in music. *Musicae Scientiae, 5*, 213–234.

DANNENBERG, R. B., & MONT-REYNAUD, B. (1987). Following a jazz improvisation in real time. In *Proceedings of the 1987 International Computer Music Conference* (pp. 241–248). San Francisco: International Computer Music Association.

DESAIN, P., & HONING, H. (1989). The quantization of musical time: A connectionist approach. *Computer Music Journal, 13*, 56–66.

DESAIN, P., & HONING, H. (1992). The quantization problem: Traditional and connectionist approaches. In M. Balaban, K. Ebcioglu, & O. Laske (Eds.), *Understanding music with AI: Perspectives on music cognition* (pp. 448–463). Cambridge, MA: MIT Press.

DESAIN, P., & HONING, H. (2003). The formation of rhythmic categories and metric priming. *Perception, 32*, 341–365.

DESAIN, P., HONING, H., THIENEN, H. VAN, & WINDSOR, W. L. (1998). Computational modeling of music cognition: Problem or solution? *Music Perception, 16*, 151–166.

EITAN, Z., & GRANOT, R. Y. (2005). How music moves: Musical parameters and listeners' images of motion. *Music Perception, 23*, 221–247.

EPSTEIN, D. (1994). *Shaping time*. New York: Schirmer.

FELDMAN, J., EPSTEIN, D., & RICHARDS, W. (1992). Force dynamics of tempo change in music. *Music Perception, 10*, 185–204.

FODOR, J. (2000). *The mind doesn't work that way. The scope and limits of computational psychology*. Cambridge, MA: MIT Press.

FRIBERG, A., & SUNDBERG, J. (1999). Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. *Journal of the Acoustical Society of America, 105*, 1469–1484.

GABRIELSSON, A. (1999). Music performance. In D. Deutsch (Ed.), *Psychology of music* (2nd ed., pp. 506–602). New York: Academic Press.

GJERDINGEN, R. (1994). Apparent motion in music? *Music Perception, 11*, 335–370.

HONING, H. (2003). The final ritard: On music, motion, and kinematic models. *Computer Music Journal, 27*, 66–72.

HONING, H. (2005). Is there a perception-based alternative to kinematic models of tempo rubato? *Music Perception, 23*, 79–85.

HUDSON, R. (1996). *Stolen time: History of tempo rubato*. London: Clarendon Press.

JACOBS, A. M., & GRAINGER, J. (1994). Models of visual word recognition—Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 1311–1334.

KRONMAN, U., & J. SUNDBERG (1987). Is the musical ritard an allusion to physical motion? In A. Gabrielsson (Ed.), *Action and perception in rhythm and music: No. 55* (pp. 57–68). Stockholm: Royal Swedish Academy of Music.

LARGE, E. W., & KOLEN, J. F. (1994). Resonance and the perception of musical meter. *Connection Science, 6*, 177–208.

LERDAHL, F., & JACKENDOFF, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

LONGUET-HIGGINS, H. C. (1987). *Mental processes: Studies in cognitive science*. Cambridge, MA: MIT Press.

LONGUET-HIGGINS, H. C., & LISLE, E. R. (1989). Modelling music cognition. *Contemporary Music Review, 3*, 15–27.

MCAULEY, J. D. (1995). *Perception of time as phase: Towards an adaptive oscillator model of rhythmic pattern processing.* Unpublished doctoral dissertation, Indiana University, Bloomington.

MICHON, J. A. (1964). Studies on subjective duration 1. Differential sensitivity on the perception of repeated temporal intervals. *Acta Psychologica, 22*, 441–450.

NARMOUR, E. (1992). *The analysis and cognition of basic melodic complexity: The implication-realization model*. Chicago: Chicago University Press.

PITT, M. A., & MYUNG, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Science, 6*, 421–425.

PITT, M. A., MYUNG, I. J., & ZHANG, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472–491.

PYLYSHYN, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: MIT Press.

REPP, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei." *Journal of the Acoustical Society of America, 92*, 2546–2567.

ROBERTS, S., & PASHLER, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

ROBERTS, S., & PASHLER, H. (2002). Reply to Rodgers and Rowe (2002). *Psychological Review, 109*, 605–607.

ROBERTS, S., & STERNBERG, S. (1993). The meaning of additive reaction-time effects: Test of three alternatives. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience—A silver jubilee* (pp. 611–653). Cambridge, MA: MIT Press.

RODGERS, J. L., & ROWE, D. C. (2002). Theory development should begin (but not end) with good empirical fits: A comment on Roberts and Pashler (2000). *Psychological Review, 109*, 599–604.

02.MUSIC.23_365-376.qxd 30/05/2006 11:11 Page 376

SHOVE, P., & REPP, B. H. (1995). Musical motion and performance: Theoretical and empirical perspectives. In J. Rink (Ed.), *The practice of performance* (pp. 55–83). Cambridge, UK: Cambridge University Press.

SUNDBERG, J., & VERILLO, V. (1980). On the anatomy of the ritard: A study of timing in music. *Journal of the Acoustical Society of America, 68*, 772–779.

TODD, N. P. M. (1985). A model of expressive timing in tonal music. *Music Perception, 3*, 33–58.

TODD, N. P. M. (1992). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America, 91*, 3540–3550.

TODD, N. P. M. (1995). The kinematics of musical expression. *Journal of the Acoustical Society of America, 97*, 1940–1949.

TODD, N. P. M. (1999). Motion in music: A neurobiological perspective. *Music Perception, 17*, 115–126.

TOIVIAINEN, P. (1998). An interactive MIDI accompanist. *Computer Music Journal, 22*, 63–75.

TRUSLIT, A. (1938). *Gestaltung und Bewegung in der Musik* [Shaping and motion in music]. Berlin-Lichtenfelde: Chr, Fiedrich Vieweg.