9th International Conference on Music Perception and Cognition

Alma Mater Studiorum University of Bologna, August 22-26 2006

The role of surprise in theory testing: Some preliminary observations

Henkjan Honing
Music Cognition Group
ILLC / Universiteit van Amsterdam
www.hum.uva.nl/mmm
honing@uva.nl

ABSTRACT

While the most common way of evaluating a computational model is by showing a good fit with the empirical data, recently the literature on theory testing and model selection criticizes the assumption that this is actually strong evidence for a model. This paper explores the possibilities of developing a method selection technique that can serve as an alternative to a goodness-of-fit (GOF) measure. This alternative, a measure of surprise*, is based on the common idea that a model gets more support from the correct prediction of an unlikely event than the correct prediction of something that was expected anyway.*

Keywords

Cognitive science, computational modeling, model selection, rhythm perception.

INTRODUCTION

How should we select among computational models of cognition? This is a question that has attracted much discussion recently (Roberts & Pashler, 2000; 2002; Rodgers & Rowe, 2002; Pitt, Myung, & Zhang, 2002) and is at the heart of the scientific enterprise of cognition. A number of criteria have been proposed to assist in this endeavor (Jacobs & Grainger, 1994). They can be summarized as:

1. *Plausibility*; Are the assumptions of the model computationally and psychologically plausible?
2. *Explanatory adequacy*; Is the theoretical explanation reasonable and consistent with what is known?
3. *Interpretability*; Do the model and its parts —e.g., parameters— make sense? Are they understandable?
4. *Descriptive adequacy*; Does the model provide a good description of the observed data?

5. *Generalizability*; Does the model predict well the characteristics of data that will be observed in the future?
6. *Complexity*; Does the model capture the phenomenon in the least complex —i.e., simplest— possible manner?

The relative importance of these criteria may vary with the types of models being compared. For example, verbal (or informal) models are likely to be judged on the first three criteria. Computational models, on the other hand, may have already satisfied the first three criteria in an earlier stage of their development, making the last three criteria be the primary ones on which they are evaluated (Pitt, Myung & Zhang, 2002). In recent decades, computational modeling has become a well-established research method in many fields, including language (Pylyshyn, 1984; Fodor, 2000), vision (Longuet-Higgins, 1987), reasoning (Stenning & Van Lambalgen, 2005), and music cognition (Longuet-Higgins, 1987; Desain & Honing, 2004). It's nowadays hard to think of a cognitive theory that does not have a computational component. It even seems that computational modelling became a victim of its own success (see, for example, the sheer quantity of models of beat induction and tempo tracking in music; *cf.* Desain & Honing, 2004). Therefore, selecting among computational models has become a more important issue than before.

APPROACH

While the most common way of evaluating a computational model is to see whether it shows a good fit with the empirical data, recently the literature on theory testing and model selection criticizes 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, but clearly not the end point of model selection or verification (*e.g.,* Jacobs & Grainger, 1994; Desain, Honing, Thienen, & Windsor, 1998; 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). These authors identified serious shortcomings of GOF as a model selection method (summarized in Figure 1). 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).

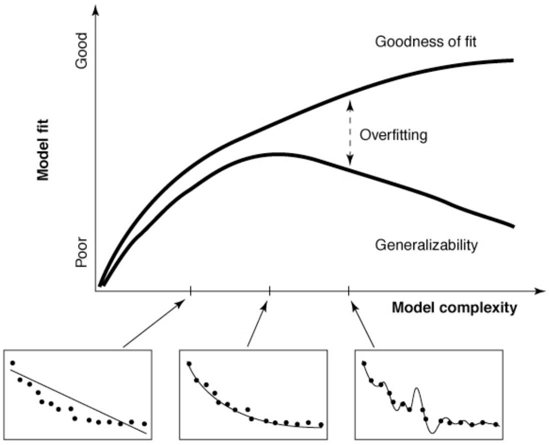

Figure 1. Goodness of fit and generalizability as a function of model complexity. The y-axis represents any fit index (a larger value indicating a better fit). The three smaller graphs along the x-axis show how fit improves as complexity increases. In the left graph, the model (represented by the line) is not complex enough to match the complexity of the data (dots). In the middle graph the two are well matched in complexity, which is why this occurs at the peak of the generalizability function. In the right graph, the model is more complex than the data, fitting random error. It has better goodness of fit, but is in fact overfitting the data. (Adapted from Pitt & Myung, 2002).

Honing (2006) presents a case study that was inspired by this debate. It focuses on the possibility of selecting among computational models of expressive timing in music, in particular those trying to explain the shape of final *ritardandi* (or *R* for short*)*: the typical slowing down at the end of a music performance (see Figure 2). It compared two families of computational models using three different model selection criteria: 1) goodness-of-fit (GOF), 2) model's simplicity, and 3) the amount of surprise in the predictions. These three criteria form the basis for a more general study on which this paper makes some preliminary observations.

Next to well-known *GOF measures* (*e.g.*, percent variance accounted for or RMS error), two alternative measures are considered. The first is based on a *measure of simplicity* using two well-known candidates: minimum description length (MDL), which provides an intuitive and theoretically well-grounded understanding of why one model should be chosen (Grünwald, Myung & Pitt, 2005), and Bayesian model selection (BMS). The latter method assesses a model's generalizability by combining a GOF-measure with a measure of complexity (Kass & Raftery, 1995; *cf.* Sadakata, Desain & Honing, 2006). The two approaches reflect the classical duality of 'simplicity' versus 'likelihood'.

The second, and relatively novel selection method that will be considered is a *measure of surprise*. In this approach the key idea is to try and capture the predictive power of a model in making precise and potentially unexpected predictions; The amount of 'surprise' in the predictions being an important contribution to the impact of a theory.

An example, using the case study mentioned above, might make this more clear. While a computational model might be designed and fine-tuned to explain one particular phenomenon, it could, in principle, say something about the consequences for a related phenomenon as well. For instance, a model that was designed to capture perceived regularity (*cf.* Honing, 2005a) can be used to make predictions on *R*: how much slowing down (or speeding up) still allows for an appropriate categorization of the performed rhythm according to the model. This was not what the components of the model were designed to capture. However, they could be interpreted that way relatively easily. Calculating the predictions of this model on the possible shapes of final *ritardandi* turned out to be relatively surprising.

In short, I would like to argue that the level of surprise of a model's predictions is more relevant and interesting than a model that simply makes a good fit with the data it was designed to fit.

WHAT MAKES A MODEL SURPRISING?

To give some structure to the notion of surprise, in Figure 2 a distinction is made between possible, plausible, and predicted observations, again using the example of models of *R*. The total area of the square indicates the possible

[*] This research is in progress (May 2006). See Honing (2006; in preparation) for a more elaborate discussion.

tempo values (e.g., a horizontal line would indicate a constant tempo, a vertical line an instant tempo change).

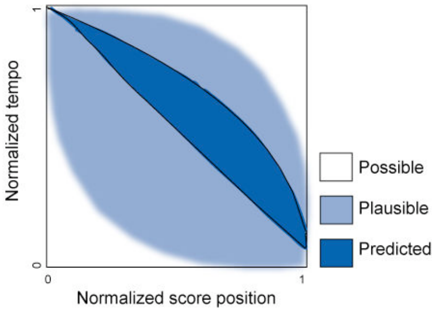

Figure 2. Schematic diagram of possible, plausible, and predicted values of a model of *R*. The x-axis indicates position in the musical score, the y-axis indicates tempo (relative speed of the performance normalized w.r.t. the overall pre-*R* tempo).

However, the plausible values —the values one can expect to occur in the case of a slowing-down in tempo— are roughly within the lighter area. It includes all curves that predict slowing-down. The darker area indicates the predicted values of a model of *R*. The predictions of the model shown in Figure 2 should therefore be judged as being non-surprising, based on the intuitive idea that a model gets more support from the correct prediction of an unlikely event than the correct prediction of something that was expected anyway.

For a model to be surprising, first, all predicted outcomes should be a small fraction of the plausible outcomes. Only when few observations and precise predictions across all parameter values are made, is this substantial evidence for a model. 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. A model that exhibits a larger forbidden zone is less flexible, and thus potentially easier to falsify — a characteristic that is considered a strong aspect of a model. Thus, as an example, in Figure 3 we should prefer B and D over A and C.

Second, a model that predicts simple or smooth shapes is less surprising than one that predicts non-smooth 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 (Roberts & Pashler, 2000). Hence, in Figure 3 we should prefer C and D over A and B.

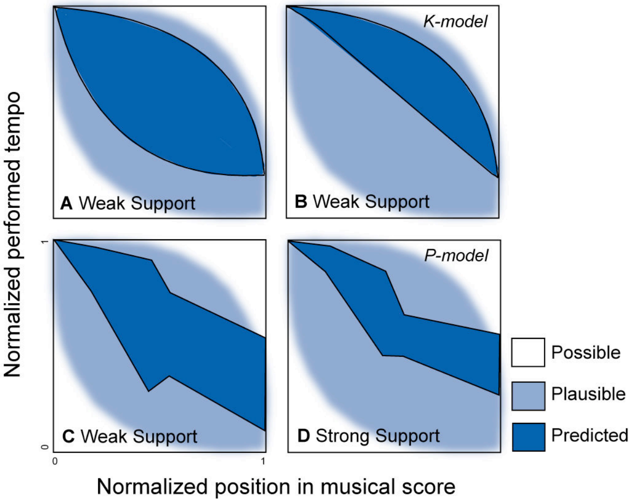

Figure 3. Schematic diagram of strong and weak support for a model of *R*. The model that makes limited range, non-smooth and surprising predictions is favored. The x-axis indicates position in the musical score, the y-axis indicates tempo (Adapted from Honing, 2006).

DISCUSSION

While for most scientists the limitations of GOF might be clear (*cf.* Pitt & Myung, 2002), the recent discussion in the cognitive science literature (summarized here in the Approach section) shows that this selection method is still (or again) in the center of scientific debate. However, Honing (2006) can be seen as a proof of concept with regard to the applicability of the element of surprise to theory testing and model selection in music cognition research. It demonstrates that a measure of surprise —an index of the limited range, non-smooth and surprising predictions of a model— could potentially serve as an alternative to more common model selection techniques, including GOF and measures of simplicity. Of course, the challenge is to elaborate, formalize and evaluate such a measure of surprise. This is the topic of current research.

ACKNOWLEDGMENTS

Thanks to Leigh M. Smith and Olivia Ladinig for their comments on an earlier version of this paper.

REFERENCES

Desain, P., & Honing, H. (2003). The formation of rhythmic categories and metric priming. *Perception*, *32*, 341-365.

Desain, P., & Honing, H. (2004). *Final Report NWO-PIONIER Project 'Music, Mind, Machine'*. ILLC, X-2004-02. [http://dare.uva.nl/en/record/117783]

Desain, P., Honing, H., Thienen, H. van, & Windsor, W. L. (1998) Computational Modeling of Music Cognition: Problem or Solution? *Music Perception, 16*(1), 151-166.

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

Grünwald, P., Myung, I.J., & Pitt, M. (eds)(2005). *Advances in Minimum Description Length: Theory and Applications*. Cambridge: MIT Press.

Honing, H. (2005a). Is there a perception-based alternative to kinematic models of tempo rubato? *Music Perception*, *23*(1), 79-85.

Honing, H. (2005b). Music cognition: Theory testing and model selection. *Proceedings of the XXVII Annual Conference of the Cognitive Science Society* (CogSci2005), 38, Stresa: University of Turin.

Honing, H. (2006). Computational modeling of music cognition: a case study on model selection. *Music Perception*, *23*(5). [http://dare.uva.nl/en/record/146627]

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.

Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Longuet-Higgins, H. C. (1987). *Mental processes. Studies in cognitive science.* Cambridge, Mass.: MIT 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, Mass.: MIT Press.

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.

Sadakata, M., Desain, P., & Honing, H. (2006). The Bayesian way to relate rhythm perception and production. *Music Perception, 23*(3), 267-286.

Stenning, K., & van Lambalgen, M. (2005) Semantic interpretation as computation in nonmonotonic logic: The real meaning of the suppression task. *Cognitive Science: A Multidisciplinary Journal*, *29*(6), 919-960