# Summary of the Music Performance Panel, MOSART Workshop 2001, Barcelona

Maarten Grachten

*Artificial Intelligence Research Institute, IIIA*
*Spanish Council for Scientific Research, CSIC*
*Campus UAB, 089193 Bellaterra, Catalonia, Spain*
maarten@iiia.csic.es

**Abstract**

This paper presents a summary of the Music Performance Panel, held at the MOSART Workshop 2001, in Barcelona. The approaches of the represented research groups are described briefly, and an overview is given of the topics that were addressed.

## 1 Introduction

During the MOSART Workshop 2001, on current research directions in computer music, a discussion panel addressed some issues on the topic of music performance. The panel consisted of the following members: Gerhard Widmer (ÖFAI, Vienna), Henkjan Honing (NICI, Nijmegen), Johan Sundberg (KTH, Stockholm) and Giovanni de Poli (DEI, Padua). The main topics that were discussed are:

**Research Strategies** What are the relative strengths and weaknesses of different research strategies (theory-driven vs. data-driven, oriented towards cognitive plausibility vs. computational simplicity, perception-oriented vs. production-oriented, etc)? And could there be more synergy between these strategies?

**Functions of performance** Expressive music performance seems to fulfill several functions (e.g. expressing emotional content, but also clarifying structural aspects of a piece); how do these functions fit together? How do current models of performance account for these functions?

**Evaluation** Given that expression is a subjective notion and that there is no such thing as the "correct" interpretation of a piece of music, can we nevertheless develop quantitative and scientifically rigorous procedures and standards for evaluating the quality/significance/validity of proposed models of expression? Would it be worthwhile to try to collect or construct 'benchmark problems' on which different models could be compared?

In the following sections, I will resume what was said with respect to the above questions during the panel. In section 2, an overview will also be given about the approaches that each of the research groups have been taking. This overview is partly based on the opening statements that each of the panel members made, and partly on the submitted position statement papers (included at the end of this paper).

## 2 Research strategies

As pointed out by Widmer et al. in [6], prior to the question of how research strategies relate to each other and how they can be compared, it should be clear what the *research aims* are. Three typical aims are respectively: 1) development of musical models that produce well-sounding musical results, 2) development musical models that produce results that maximally resemble (generalized patterns of) observed expert performances, and lastly, 3) development musical models that structurally resemble observed or hypothesized cognitive processes of musical performance. To capture the research of DEI in a category, I would like to propose a fourth category, in addition to these three categories. Namely: 4) development of performance models that provide optimal user control over the expressive renderings of performances. Depending on the aims of research, different methodological approaches may be appropriate.

**ÖFAI** ÖFAI's research is explicitly directed towards the second aim. More specifically, the aim is to discover regularities and patterns that can be found in the performance (particularly on piano) of musical pieces. This should result in models that as precisely and compactly as possible describe principles that emerge from performance data. Their strategy to achieve this aim can be described as analysis by machine induction. A large collection of musical data (expert performances of various complete Mozart piano sonatas) form the data for the inductive learning. These data are analyzed in terms of dynamics and tempo deviations. Structural representations (vz. transcriptions) of the performed music are also made. To this analysis of the data, the machine learning techniques are applied. The outcome of the learning process are the regularities that were found (co-occurrence of performance deviations and structural features of the music), typically in the form of prediction rules that predict expressive deviations based on the structural description of the music.

As the above suggests, rather than using the data to test preconceived hypotheses, the data are analyzed to generate hypotheses about music performances. In this respect, ÖFAI's research could be called bottom up, or data-driven. The knowledge that is gained in this way, is only about *what* is done during a performance, not *why* it is done; the intentions of the player are not accessible through this approach.

An advantage of ÖFAI's approach, being a data-driven approach, is that the results are independent of theories, hence independent of the need to verbalize and conceptualize in advance what will be analyzed. A data-driven approach may detect regularities in the data that were never noticed by human subjects.

**NICI** As opposed to the above approach, NICI tries to build *cognitive* models of music performance, rather than models that describe the performance itself. Thus, in terms of the three research aims mentioned above, the third aim is pursued by NICI. A major implication for methodology is that empirical data is obtained through controlled experiments (as is common in cognitive psychology), rather than by analyzing (large samples of) musical performances. In these experiments, the validity of constructs from musicological theories is tested. To do this, these theoretical constructs must first be formalized and implemented as algorithms. In this way, musicological theories are computationally modeled (cf. [3]). An important requirement for these models is that they are not primarily able to faultlessly reproduce human performances, but rather that parameters of the model are musically meaningful, i.e. correspond to musical concepts in the mind of the musician.

During the panel, Honing argued that controlled experiments are preferred over the analysis of large corpora of music performances. In experiments it would be possible to discard unintended deviations, and to give the performer instructions and measure their effects on performances. This kind of interaction is not possible in the case of corpora-analysis.

Another point that was made, is that the focus is on the re-construction of a 'performance-space', rather than on studying only expert-performances. By also observing non-expert performances and performances with particular intentions, a more complete view on the range of meaningful deviations can be accomplished.

Furthermore, Honing mentioned the relevance of perceptual studies of expressive deviations. Through controlled listening experiments for example, it can be established how great deviations in the performance of a rhythm can become before it is perceived as a different rhythm. In this way it is possible to elucidate the constraints on expressive transformations.

Lastly, he argued that models of performance might very well benefit form a deeper understanding of the cognitive reality that accounts for the performances.

**KTH** The main research goal of KTH has been to gain a deeper understanding on music communication between musicians and listeners (possibly other musicians). Contrary to the idea that musical pieces can be performed in infinitely many equally well-sounding ways, the professional music performer Lars Frydén perceived that there are clear regularities in music performance. This gave rise to the idea of constructing a set of rules for musical performance (see [4] and [5]). To test the validity of the rules, an *analysis-by-synthesis* approach is adopted. This is to say that music performances are reconstructed from the score, by using the rules. The rules are applied one by one, each with an individual parameter that controls the magnitude of the effect of the rule, so the effect of each individual rule can be examined in detail.

In contrast to ÖFAI's approach, KTH's approach is primarily theory-driven. After one or more rules (which can be regarded as hypotheses) have been formulated, it's effect on performance is tested, by listening experiments where subjects rate the musical acceptability of performances that were generated by the rules. This way of evaluating is rather different from ÖFAI's evaluation method, where the performances generated by the system are not evaluated by listening, but by measuring their deviations from professional performances.

Sundberg mentioned some advantages of the analysis-by-synthesis method of testing the performance modeld. Firstly, the synthesis enables researchers to evaluate hypotheses under musically realistic conditions, namely by listening to performances that are generated under these hypotheses. Secondly, it is possible to test hypothesized rules separately and tune them, one at a time. Thirdly, the situation of rule-tuning is similar to a teacher student setting, so that the person that tunes the rules (usually a musical expert) can rely on his/her pedagogical skills. Lastly, this approach is independent of training data and as such, it is apt to produce non-obvious interpretations of a piece, that nevertheless comply to the musical performance principles.

Some limitations of this method that Sundberg mentioned, were that the rules are a reflection of the expertise of just one individual. Also, the system will produce identical performances with identical rule palettes.

**DEI** Music performance research at DEI is primarily concerned with *control* of sound, in order to give music composers useful and usable tools for generating music from sound/instrument models. As De Poli noted, the problem of music performance is in between music generation and sound production. An important question here is how the sounds can be controlled in an expressive manner, on a slow varying time scale (e.g. on the level of musical phrases). In general, there are two strategies: one is control by gesture, the other is the use of models for controlling the sound production. The latter approach has been adopted by DEI.

Thus, performance research at DEI is aimed at building models that map expressive intentions (through the use of dichotomous labels like ‘hard’, ‘soft’, ‘bright’, ‘dark’), to low-level acoustic features of the performance. These models can be used to render nominal performances in expressive ways, with (real-time) high-level control over the expressive parameters.

To establish the relation between expressive labels and acoustic performance features, performing/listening experiments have been done. These experiments showed that listeners ordered music performances that were played with different expressive intentions, along two abstract dimensions: ’kinematics’ and ’energy’ [1]. A mapping was then established between coordinates in the kinematics-energy space and deviations of expressive parameters like tempo, legato and intensity. This mapping serves as a model for expressiveness, translating the points in the abstract control space to expressive deviations.

De Poli noted that an important aspect of models of performance is that they convey a multi-level abstraction from the score, that is, the highest level expressive concepts should not be directly mapped to the lowest level (acoustic) parameters, but via several hierarchically ordered abstraction levels, corresponding to different time scales.

Another important question is the generalizability of the expressive models. The models were constructed based on Western classical music, of which the practice is relatively fixed. The practice of popular Western music, on the contrary, is less codified. This may imply that accurate expressive models are harder to build for this kind of music.

Future plans of DEI are to work in the reverse way; that is, instead of constructing performances based on expressive intentions, rather analyzing performances in order to derive the musicians expressive intentions from it.

# 3 Functions of performance

The question of how different approaches account for the functions of performance did not receive much direct attention during the panel. Nevertheless, some statements can be made about it, as to some extent, the stance toward the function of performance is inherent to a particular approach.

Firstly, in accord with Widmer’s remark that they study *what*, rather than *why*, it can be observed that ÖFAI’s research comprises only structural/syntactical analysis of performances. Performance elements are not related to anything external to the musical piece (like the performer’s intentions). For that reason, the functions of performance that can be investigated are bound to be about the performance itself; not e.g. communicative functions (like expressing emotions). Indeed Widmer noted that the function of performance they study, is performance as clarifying musical structure.

NICI’s research is not explicitly directed to accounting for the functions of performance. However, in studying expressive timing, the structural role of performance elements is identified as one of the factors that influence timing (see [7]).

KTH’s rule based approach also incorporates the function of performance as clarifying structure: this is evident by the categorizations within the rule base, where one category of rules is called ‘grouping rules’. These rules are intended to elucidate the boundaries between different structural units, like phrases.

On the other hand, the function of performance as communicating emotional content can also be modeled, as Sundberg noted, by the existence of magnitude parameters for the effect of each rule. Their hypothesis is that particular kinds of interpretations of a piece (‘sad’ or ‘happy’), correspond to particular settings of these parameters. This correspondence is nicely present in the metaphor of ‘rule palettes‘, that Sundberg used, suggesting the possibility of ‘painting’ with the magnitude parameters.

Inherent to DEI’s approach to musical performance, is the function of communicating emotional content. As their aim is to render or transform music performances according to expressive labels like ‘dark’, ‘bright’, ‘light’ or ‘heavy’, it is clear that the focus is on performance as expressing intentions. As argued in [2], sensorial adjectives were preferred over emotional ones like ‘sad’/‘happy’, in order to limit the semantics under examination. The sensorial adjectives clearly have a more restricted meaning and related to performance more closely.

# 4 Evaluation

Finally, there was the question about the evaluation of performance models and the use and usefulness of benchmarks in the area of music performance research. There were quite diverse opinions about this among the panel members.

About the evaluation of their research, Widmer said that the focus was on two prime concepts: predictive accuracy and generality. Performance models should on the one hand predict the performance deviations of (expert) performers as accurately as possible, while on the other hand, the predictions should ideally be valid across different performers and musical styles. About the possibility of using benchmarks for evaluation of performance models, Widmer remarked that the use of benchmarks suggests that there is a set of pieces for which there is a ‘correct’ performance, which must be matched as close as possible by any good model. This is obviously not the case. Furthermore, using a standardized dataset as benchmark, introduces the risk of over-fitting the models to the benchmark data, as has apparently been the case in the area of machine-learning. This over-fitting should be avoided. Given these risks, Widmer supposes that, with much awareness, it could still be useful to propose a standardized test dataset. This dataset should at least be very diverse, different musical styles and performers should be represented.

At this point, I would like to note that the usefulness and justifiability of benchmarks for evaluation, is somewhat dependent on the goal of research. In ÖFAI’s case, where the goal is to match human expert performances as closely as possible, it makes makes more sense to use benchmarks, because the objective is purely quantitative: the deviation between test-data and predictions of the model should decrease to zero. If the goal is to produce musically acceptable results (as with the research of KTH), the use of benchmarks is less obvious, because the most important thing is that the performances resulting from the model should sound musically convincing in themselves, not that they are *similar* to musically convincing performances. When the aim is a truth-like cognitive model of music performance, a good use of benchmarks is neither very clear, because good cognitive models do not necessarily predict a particular set of actual data very accurately. Rather they predict the constraints that hold for music performance in general.

A related remark, made by Honing, is that in the case of human performances of music, not all the information is contained in the data, but that there is a lot of information which is only *suggested* by the data, but actually is in the minds of the listeners (e.g. tempo and evoked emotion are not measurable in the data themselves). This perceptual information is an important aspect of performances, which is not covered by a straight-forward use of benchmarks to evaluate models. Hence, benchmarks are only partly relevant as an evaluation tool.

There was a reply to this from the audience (by Jan Tro), that although the information may not all be conveyed by the data, at least the ‘triggers’ for this external information, are embodied in the data. Although this is obviously true, I would like to add that it does not take away the need for perceptual research to music as well, in order to establish to what kind of perceptual phenomena these triggers map.

# 5 Other remarks made during the panel session

Sundberg raised the point that how often a performance principle applies, may not touch the essence of such a principle (a rarely used performance rule may nevertheless be musically important). A more relevant question would be what the meaning of the deviation is, that is, what is expressed by it?

Widmer answered that the two kinds of research on music performance (looking for musical meaning of expressive deviations on the one hand and looking for communal expressive patterns in performances on the other), might very well co-exist at the same time. They should be regarded as complementary, where the regularities found by the latter approach could form a useful point of departure for the former. He furthermore noted that it should be made explicit that it is a *hypothesis* that the central function of expression is to communicate meaning to the listener.

A remark from Honing was that there is no such thing as an ‘average performance’. Averaging over several performances of the same piece (let alone different pieces), will not result in a ‘typical’ performance, and will probably not convey much useful information. Sundberg agreed that using averages in a quantitative way, will tend to diminish the magnitude of the measured effects of performance principles.

A critical remark from the audience (by Werner Goebl) was that it should be realized that performances generated under a controlled experiment cannot be taken to be exchangeable with other performances, like live performances or studio performances. Attempts to manipulate the performance by instructions may yield performances that are not representative, because it affects the performer in unnatural ways. Honing replied that through clever design of experiments it may be possible to manipulate the performers in unconscious ways. Sundberg noted that it could be interesting to study how performances in different settings differ.

A question from the audience was about the issue of instrument fingering. In what way does fingering affect the performance? Widmer suggested that fingering does not so much affect performance as fingering is chosen to achieve the expressive affect that is intended by the performer. Honing added that in addition to the effect of musical structure and emotion on expression, there is the effect of the instrument on expression. Typically musicians emphasize parts of a piece that are difficult to play on a particular instrument by a deliberate fingering. In general, Honing agrees with Widmer that fingering is chosen to maximize expressive control.

A final appeasing point made from the audience (by Roger Dannenberg) with respect to the problem of evaluation and collaboration of different performance models was that criticism and skepticism about the right way to proceed and combine research may be an obstacle for progression. It may be fruitful to share results and data, even with the limitations that hold.

# References

[1] S. Canazza, G. de Poli, A. Rodà, A. Vidolin, and P. Zanon. Kinematics-energy space for expressive interaction in music performance. In *Proceedings of the Workshop on current research directions in computer music*, pages 35–40, Barcelona, november 2001.

[2] S. Canazza, G. Poli, and A. Vidolin. Perceptual analysis of the musical expressive intention in a clarinet performance. In M.Leman, editor, *Music, Gestalt and Computing*, pages 441–450. Springer Verlag, Berlin, 1997.

[3] P. Desain, H. Honing, and R. Timmers. Music performance panel: NICI/MMM position statement. MOSART Workshop on current research directions in computer music, november 2001.

[4] A. F. Johan Sundberg and R. Bresin. Music performance panel: Position statement kth group. MOSART Workshop on current research directions in computer music, november 2001.

[5] J. Sundberg, A. Friberg, and L. Frydén. Common secrets of musicians and listeners: An analysis-by-synthesis study of musical performance, 1991.

[6] G. Widmer, S. Dixon, W. Goebl, E. Stamatatos, and A. Tobudic. Empirical music performance research: ÖFAI's position. MOSART Workshop on current research directions in computer music, november 2001.

[7] W. Windsor, P. Desain, H. Honing, R. Aarts, H. Heijink, and R. Timmers. On time: the influence of tempo, structure and style on the timing of grace notes in skilled musical performance. In *Rhythm perception and production*, pages 217–223. Swets & Zeitlinger, Lisse, NL, 2000.

# Empirical Music Performance Research: ÖFAI's Position

Gerhard Widmer, Simon Dixon, Werner Goebl, Efstathios Stamatatos, Asmir Tobudic
Austrian Research Institute for Artificial Intelligence (ÖFAI)
Schottengasse 3, A-1010 Vienna, Austria
{gerhard|simon|wernerg|stathis|asmir}@ai.univie.ac.at

## Abstract

This short paper presents our view on some general questions regarding empirical research on expressive music performance. The main direction of performance research going on at the Austrian Research Institute for Artificial Intelligence (ÖFAI) is briefly reviewed and positioned relative to three general issues, namely, different research strategies, different dimensions of performance, and the question of empirical evaluation of performance models.

## 1 Introduction

The *Music Performance Panel* held at the MOSART 2001 workshop is dedicated to three principal questions that try to put current research on expressive music performance into perspective: what are different *research strategies*, and what are their respective roles? what are different *functions* or *dimensions* of performance, and how are these accounted for by different research approaches? and how should formal, computational models of performance be *evaluated*?

We believe that when, and indeed before, trying to answer these questions it is crucial to define for oneself what the *goal* and *purpose* of one's research is: (a) do we aim at computational models of performance that produce well-sounding musical results and thus are useful to the music software industry? or (b) do we aim at models that as much as possible fit the patterns and regularities observed in expert performance and can make predictions regarding aspects of expert performances? or (c) do we want models that, through their very structure and conceptual design, reflect an observed or hypothesized cognitive reality?

These are quite distinct goals. For instance, in the first case (a), we will probably not care about whether the model itself is cognitively adequate, or we will care about that only to the extent that a model expressed in more "intuitive" terms is also easier to use and control (cf. Desain et al.'s point on FM synthesis vs. physical modeling [4]). Also, the different goals will necessitate different strategies for evaluating the usefulness (a) or precision and generality (b) or plausibility (c) of proposed models.

Different research groups (some of which are represented in the MOSART consortium) capitalize on different goals, and thus both their approaches, theoretical and technical, and the way they present and evaluate their results, are different. In the rest of this paper, we will focus on our own research as it relates to musical performance, and will try to position it relative to the above issues.

## 2 Inducing Models from Large Collections of Expert Performances

### 2.1 Research Goals

ÖFAI's immediate research goals focus on the second of the above three alternatives: we want to find *descriptive* and *predictive (partial) models* of certain aspects of expressive performance. These models should "explain" (i.e., fit) as much as possible of the observed phenomena, and they should be predictive in the sense that they generalize to other performers and possibly other types of music. The starting point for these investigations are large collections of "real-world" performances (in particular, performances by concert pianists made not specifically for research purposes).

To this end, we develop and use *Artificial Intelligence* and, more specifically, *Inductive Machine Learning* techniques to find computational models of typical performance strategies [10]. We take a strictly data-driven approach: expert performances are collected, quantitative details concerning expressive performance (timing, dynamics, articulation) are measured, and the resulting data are analyzed with the help of machine learning algorithms that try to find common patterns and regularities in these data. In this way, the computer is used as a tool or assistant in the process of inductive model building.

Cognitive adequacy of the resulting models is not an immediate goal; that would probably require a different kind of approach (and it would require expertise in cognitive psychology that we do not have). The main point of our research is to discover potentially new, general patterns that have hitherto been neglected in performance research. These may then be studied in more focused and controlled experiments.

## 2.2 Research Strategy

We see our approach as complementary to the research strategies followed by other performance researchers, be they based on systematic controlled experimentation (e.g., [13]), on ‘analysis-by-synthesis’ [8] or on purely statistical methods (e.g., [6]).

What distinguishes our work from most of the other work in empirical performance research is the use of computational learning and knowledge discovery methods and, connected with that, the strictly data-oriented approach. We use algorithms that can search for and discover complex dependencies and regularities in extremely large data sets, and can describe their discoveries to the user in intelligible terms [12].

A distinct advantage of such an approach is that the computer is free of any musical preconceptions and expectations and thus may more easily come up with novel and possibly surprising hypotheses [11]. These hypotheses may not necessarily always relate to a conceptual framework that musicians or musicologists find musically intuitive or cognitively plausible. In other words, they may not be directly interpretable as a model that reflects the musical reality of a performer. But the discovered patterns may point to interesting phenomena that have not been looked at so far and that can then be studied in more focussed and controlled experiments. In our view, that is the main role of this machine induction approach.

## 2.3 Aspects of Performance Studied

Starting from given collections of expert performances also has consequences on the types of things we can and cannot study. To put it simply, what can be hypothesized from given performances is *what* the performer did and what s/he is likely to do in other pieces, but not (or not directly, at least) *why* s/he did it (the performer’s musical or communicational intentions) or what effect the observed performance strategies have on the *listener* (the perception of performed music). The latter questions would require controlled experiments with performers and/or listeners, where performers are asked to play pieces under different conditions (as, e.g., in [9]) or with different kinds of ‘target emotion’ [3]. If we only take given performances, we cannot, for instance, make any quantifiable statements about emotional aspects, either in terms of production or perception. What we can hope to discover from large collections of precisely measured expert performances is general expressive patterns that seem to be common across a wide range of pieces and different performers and thus seem to indicate general performance strategies [11]. The same kind of material can also be used to study systematic *differences* between performers, again with inductive methods [7].

To widen the range of questions we can answer, and to clarify some very basic, but elusive notions (e.g., what really is “tempo”?), we have recently also started to perform controlled experiments with human subjects, both listeners and performers (e.g., [2, 5]). Here we can study certain specialized questions (e.g., the phenomenon of *melody lead*) in more detail, but with a narrower data basis (because producing controlled experimental data with human subjects is expensive).

## 2.4 Evaluation Issues

In empirical research, testing inductively obtained hypotheses on independent data is essential. In order to make it possible to compare competing models and algorithms, they have to be tested on a common set of data of an appropriate level of complexity.

In the area of machine learning, for instance, this has led to the establishment of a database of common benchmark data sets on which new algorithms must be tested so that their results can be compared to the results of other methods. The database is maintained by a group at the University of California at Irvine [1] and is continually updated with new data sets contributed by members of the scientific community.

In the area of music performance research, establishing such a database of common test data would be an interesting (and laborious) task. Whether it would be worthwhile would depend on a consensus, within the research community, on a set of basic evaluation criteria. To prevent a possible misunderstanding, let us make clear that we do *not* mean that such a set of test performances would in any sense represent *the* “correct” interpretations, in the sense of an absolute benchmark. On the contrary, it should contain performances of the same pieces by different performers and possibly under different conditions.

What we consider crucial is that the *generality* of the models should be established experimentally, and that requires testing them on large sets of diverse musical situations. Using only a few hand-selected pieces for model building or testing always comes with the danger of *overfitting* (either by fitting the model too tightly to the data, or by (consciously or unconsciously) selecting the test data in such a way that they confirm the model).

Of course, working with large sets of training and test pieces makes it difficult to attend to all the details and possible artifacts that may be hidden in the data, and to have a fine control on all experimental conditions. On the other hand, the kinds of patterns we find with our data-driven approach have a certain empirical weight and generality simply by virtue of the fact that they are based on (and their predictive potential has been tested on) a large set of diverse musical pieces. We do believe that the size, complexity, and musical diversity of experimental test data can give a new kind of quality and validity to experimental results. To put it (overly) simply, in our current machine learning experiments, we sacrifice observation precision for stronger or broader empirical support.

# 3 Conclusions

It seems clear that no research approach alone will lead to complete models of expressive performance that do justice to the complexity of the phenomenon and that are adequate from every possible point of view. More cooperation between different approaches will be needed (for instance, discovering novel types of patterns with our approach and then investigating these further in more detailed and controlled experiments). That requires first and foremost the definition of a common set of problems and evaluation criteria. This panel has at least made explicit some of the differences between current approaches, but has also revealed a lot of common ground that we can build on in future work.

## Acknowledgements

ÖFAI's research on expressive music performance is supported by a very generous START Research Prize (project no. Y99-INF) by the Austrian Federal Government, administered by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung (FWF)*, and by the EU project HPRN-CT-2000-00115 MOSART. The Austrian Research Institute for Artificial Intelligence acknowledges basic financial support by the Austrian Federal Ministry for Education, Science, and Culture.

## References

[1] Blake, C. and Merz, C. (1998). UCI Repository of Machine Learning Databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`. Department of Information and Computer Science, University of California at Irvine, Irvine, CA.

[2] Cambouropoulos, E., Dixon, S., Goebl, W., and Widmer, G. (2001). Human Preferences for Tempo Smoothness. In *Proceedings of the VII International Symposium on Systematic and Comparative Musicology, III International Conference on Cognitive Musicology*, Jyväskylä, Finland.

[3] Canazza, S., De Poli, G., and Vidolin, A. (1997). Perceptual Analysis of the Musical Expressive Intention in a Clarinet Performance. In M. Leman (ed.), *Music, Gestalt, and Computing*. Berlin: Springer Verlag.

[4] Desain, P., Honing, H., and Timmers, R. (2001). Music Performance Panel: Position Statement. *MOSART Workshop on Current Research Directions in Computer Music*, Nov. 2001, Barcelona.

[5] Goebl, W. (2001). Melody Lead in Piano Performance: Expressive Device or Artifact? *Journal of the Acoustical Society of America* 110(1), 563-572.

[6] Repp, B. (1992). Diversity and Commonality in Music Performance: An Analysis of Timing Microstructure in Schumann's 'Träumerei'. *Journal of the Acoustical Society of America* 92(5), 2546–2568.

[7] Stamatatos, E. (2001). A Computational Model for Discriminating Music Performers. In *Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*, Nov. 2001, Barcelona.

[8] Sundberg, J., Friberg, A., and Frydén, L. (1991). Common Secrets of Musicians and Listeners: An Analysis-by-Synthesis Study of Musical Performance. In P. Howell, R. West & I. Cross (eds.), *Representing Musical Structure*. London: Academic Press.

[9] Timmers, R., Ashley, R, Desain, P, and Heijink, H. (2000). The Influence of Musical Context on Tempo Rubato. *Journal of New Music Research* 131–158.

[10] Widmer, G. (2001). Using AI and Machine Learning to Study Expressive Music Performance: Project Survey and First Report. *AI Communications* 14 (in press).

[11] Widmer, G. (2001). Inductive Learning of General and Robust Local Expression Principles. In *Proceedings of the International Computer Music Conference (ICMC'2001)*, La Habana, Cuba.

[12] Widmer, G. (2001). *A Meta-learning Method for Discovering Extremely Simple Partial Rule Models*. Submitted. Available as Technical Report OEFAI-TR-2001-30, Austrian Research Institute for Artificial Intelligence, Vienna.

[13] Windsor, L., Desain, P., Honing, H., Aarts, R., Heijink, H., and Timmers, R. (2000). On Time: The Influence of Tempo, Structure and Style on the Timing of Grace Notes in Skilled Musical Performance. In Desain, P. and Windsor, W. L. (Eds.), *Rhythm Perception and Production*. Lisse: Swets & Zeitlinger.

# Music Performance Panel: Position Statement

*KTH* Group

Johan Sundberg, Anders Friberg, Roberto Bresin

**Research Strategy**

**The goal of our research on music performance is to gain a deeper understanding of music communication. Our research was initiated in the 1970s. Those days the general belief was that any piece of music could be performed in a number of widely differing and yet musically acceptable ways. Therefore, it was argued, there is no chance that a decent performance can be generated by rules. On the contrary, music performances are unique, and this is what makes them musically interesting and attractive. In this situation, a statistical analysis of, e.g., tone durations in a set of performance of a given piece did not seem promising.**

At the same time the idea that performances are completely independent of rules did not agree with the vast musical experience of the late musician Lars Frydén. During his violin playing in string quartet and in orchestras, he had found a number of regularities that he wanted to test.

In this situation it seemed advantageous to choose an analysis-by-synthesis strategy, i.e., to use the score transformed into a music file as the input for a rule system that converts it into a sounding performance. This allowed the testing of the performance rules that Frydén wanted to test. Each rule could be tested on a number of music examples, thus allowing him to listen to what extent the performance was improved by the rule and to find out if the effect was of an appropriate magnitude.

This analysis-by-synthesis strategy has certain unique advantages. Its main strength is the synthesis that allows the researcher to test hypotheses by listening to performances under musically reasonably realistic conditions. The method further allows a good control of the performance in the sense that one rule can be tuned and tested at a time. Also, it provides examples that are judged in a setting somewhat similar to the student-teacher setting, so that the musician can rely on his pedagogical expertise. The method also has the advantage of allowing a systematic build up of rules, in each state giving priority to salient effects. Hence, rules tend to be developed in the order of importance. The method also has some advantage over methods where statistical data on performance drive the research process. The analysis-by-synthesis strategy is independent of such data. It may generate quite unusual interpretations of a piece that still are musically acceptable and/or interesting.

The strategy also has limitations. The basic idea is that performance is determined by regularities. This implies that the machine will generate exactly the same performance effect each time its context conditions are met. In reality, however, musicians may play a given piece quite differently. In particular, the same sequence of tones may be played differently the second time it appears in a piece. As the magnitude of the effect of each rule is controlled by a quantity parameter, the rule system can indeed generate differing performances of a piece, e.g., eliminating or exaggerating various rules.

Within our group we have an ongoing discussion regarding the perceptual relevance of random variation, but as yet, we have failed to reach a common view. Measurable random variation obviously occur, the crucial question being to what extent it is subliminal and whether or not it contributes to the esthetical quality of a performance.

The assumption that performances are controlled by regularities may prove to be unrealistic in the future. On the other hand, it seems wise to test the simplest models first and to abandon them only when their limitations have been clearly exposed.

Another limitation of our rule system is that basically it is a formalised description of the musical competence of one professional musician only. We have found this a minor concern, as our musician is generally acknowledged as an outstanding expert. Therefore, his competence must be, by and large, representative.

Research using the analysis-by-synthesis strategy is driven by data rather than by theory. Indeed, our results have sometimes driven theory. An example is the concept of melodic charge. Here, the playing of music examples demonstrated the need for variations reflecting the relation between the tone and the underlying harmony. We tried a number of different existing alternatives as control parameters for the dynamic variations, all with inappropriate result. Eventually we arrived at the relationship along an asymmetric version of the circle of fifths, that we called melodic charge. Thus, the playing of melodic lines void of characteristics that reflected this melodic charge seemed to lack an important aspect of an ideal rendering. The fact that the introduction of the melodic charge into the performance grammar improved the musical acceptability of the performance seems to imply that this novel concept is relevant to music perception.

Perturbation of tone duration is an important channel for musical expression. A remaining question is when such perturbation should be controlled by proportion or in terms of absolute duration. In some rules, such as the inegalle, we use proportions, while other rules work with absolute duration. Both alternatives seem relevant to music listening. For example, a tone appears to loose its autonomy and sound like a grace note as soon as its duration is shorter than about 100 ms.

An important task in our research is to sort out the roles of rules that operate at same level. For example, the phrasing rule should not operate on tone sequences treated by the final ritard rule. There are also other as yet not quite resolved interference between certain rules, such as the punctuation and leap articulation rules.

**Functions of performance** Our formulation of performance rules have yielded a generative grammar of music performance that has invited us to speculation regarding to function of performance, or, more precisely, regarding the function of the expressive deviations. Thus, we have seen that the rules can be divided into three major categories depending on their apparent function in music communication. One category seems to serve the purpose of differentiating tones belonging to different tone categories, i.e., to enhance the differences between pitch and interval classes and between note values. Another category seems to mark which tones belong together and where the structural boundaries are. In this way, the performer facilitates the listener's processing of the signal flow. Interestingly, the same two principles, differentiation and grouping, can be observed also in spoken communication. As yet, we have not tested these cognitive aspects of music performance, though. The third group concerns technical aspects of ensemble playing related to synchronisation of voices and tuning.

Music performances are also coloured emotionally. We have found that emotional colouring can be achieved by varying the rules' quantity parameters. Thus, by enhancing some rules and suppressing others, emotionally differing performances of the same piece can be generated. We have already constructed a set of palettes that add different emotional colours to performances (angry, sad, happy, scared, tender, and solemn) and we plan to build special rule palettes that will generate agitated and peaceful performances.

**Evaluation** Synthesised performances appear to represent a powerful tool for evaluating the perceptual relevance of research findings. It seems advantageous, however, to use expert listeners. We have had good experiences of listening tests where musicians were asked to adjust the quantity parameter to an optimum for different music examples. In these experiments, rules have been tested one by one. As zero is thereby an available choice the results show if the rule tested provides a desirable effect.

In case performance research relies on statistical data from real performances, the evaluation may be more problematic. It appears that synthesis will greatly facilitate the verification of such results.

**Future work/Remaining problems** The score we now use as input for the performance grammar is rudimentary in the sense that it contains information on nothing but pitch and duration. Thus phrase markers and chord symbols are introduced manually. Also, the realisation of conventional items like trill, point, and grace notes requires hand editing of the score. Our plan is to complement the input score with signs for such events. We also plan to implement Craig Sapp's algorithm for automated chord analysis.

Another planned improvement is to test the usefulness of a realtime control of rule quantity. This will be realised within the MEGA project; hopefully, this may solve the problem that the grammar performs the same music material exactly the same way if it reappears in a piece.

We have lately been cooperating with Max Mathews and Gerald Bennett implementing the performance grammar in the Radio Baton system. This has been an informative experience, elucidating the boundaries between the musician's and the conductor's responsibilities in shaping a performance.

Basically the research method seems unproblematic. We do not regard the analysis-by-synthesis strategy as the only possible method. An exchange of data assembled by various methods will improve quality of research and promote progress. The Vienna material represents an extremely valuable resource for the further development of the performance grammar. A crucial condition, however, would be the use of synthesis, apparently representing an indispensable opportunity to test the perceptual relevance of findings. The MOSART project comprises exchange of research results and computer synthesis of music performance as two of its core aims and thus offers a perfect opportunity to proceed along these lines.

# Music Performance Panel: NICI / MMM Position Statement

Peter Desain, Henkjan Honing and Renee Timmers
*Music, Mind, Machine* Group
NICI, University of Nijmegen
mmm@nici.kun.nl, www.nici.kun.nl/mmm

In this paper we will put forward our view on the computational modeling of music cognition with respect to the issues addressed in the *Music Performance Panel* held during the MOSART 2001 workshop. We will focus on issues that can be considered crucial in the development of our understanding of human performance and perception in its application to computer music systems. Furthermore, they were chosen such as to complement the issues brought forward by the other contributing institutes (i.e. OFAI/Vienna, KTH/Stockholm, and DEI/Padua). In summary these are:

- A computational model in agreement with music performance data is *starting point* of research, rather than an *end product* (cognitive modeling is preferred over a descriptive model)
- Importance of empirical data obtained in controlled experiments (rather than using individual examples of music performances)
- Preference for the concept of *performance space* (over the use of large corpora of music performances)
- Study performance through perception, focusing on the constraints of expression rather than studying the ideal or “correct” performance (as such avoiding the issue of performance style, and enabling the study of important aspects that are not directly measurable in the performance data itself, e.g., those of a perceptual and/or cognitive nature)

## Research aims

The panel addresses a number of dichotomies in the study of music performance, such as theory-driven vs. data-driven, oriented towards cognitive plausibility vs. computational simplicity, perception-oriented vs. production-oriented. The discussion aims to reveal research aims and methods, which are quite varied among research groups.

In our group, we study music perception and performance using an interdisciplinary approach that builds on musicology, psychology and computer science (hence the name *Music, Mind, Machine*). The aim is to better understand music cognition as a whole. The method is to start with hypotheses from music theory, to formalize them in the form of an algorithm, to validate the predictions with experiments, and, often, to adapt the model (and theory) accordingly. In other words, in the method of *computational modeling*, theories are first formalized in such a way that they can be implemented as computer programs. As a result of this process, more insight is gained into the nature of the theory, and theoretical predictions are, in principle, much easier to develop and assess. With regard to computational modeling of musical knowledge, the theoretical constructs and operations used by musicologists are subjected to such formalization. Conversely, with computational modeling of music cognition, the aim is to describe the mental processes that take place when perceiving or producing music, which does not necessarily lead to the same kind of models. As such, for us, a computational model that mimics human behavior is not enough. It in fact is more a starting point of analysis and research, than an end product (see [1] for an elaborate description).

## Evaluation and validation of music performance models

One of the key issues in developing algorithms and computational models is their validation on empirical data. In the case of the MOSART project, music that is artificially generated should respect human perception and performance such as to assure seamless interaction and intelligible control by its users. For evaluating and validating models of expression, it is problematic to search for a “correct”, general or benchmark interpretation of music [2], to which the models can be compared. Though this approach is quite common in AI modeling, it is very unattractive for music cognition research. Not only is the notion of an ideal performance questionable, comparing the input-output relation between the model and the musical performance is also too limited an evaluation. A data-driven perspective might eventually result in an accurate description [2,3], it will, however, not be a model, in the cognitive sense. It needs to describe more than just an input-output transformation. In fact, a good model is a model for which changes in parameter settings that relate to manipulated aspects of the performance (e.g. by instruction to the performer) remains to show agreement between model and performance. As such step by step further validating the model.

As an illustration of the difference between a model and a good description from another domain, one can take difference between FM-synthesis and physical modeling. It is possible to generate very convincing sounds with FM synthesis (after careful selection of the parameters). However, the whole space of sounds is unintuitive and difficult to control. In contrast, physical models have more similarity with the human world and succeed in replicating the behavior of existing objects (e.g., made of tubes and strings) that are known to the user and are therefore easier to control, despite their more restricted expressive power.

In general, a computational model that captures important aspects of human perception and action will be more successful in computer music systems. Models that simply aim at an input-output agreement do not necessarily give us a better understanding of the underlying perceptual or cognitive processes, which is essential for the development of convincing and intuitive models for human interaction with machines (see [4] for a discussion on the psychological validation of models of music cognition). A solely data-driven approach ignores the fact that important aspects of music performance are not directly measurable or present in the data itself. For instance, tempo (or expressive rubato for that matter) is a percept, and cannot be directly measured. The same applies for syncopation and other temporal aspects of music that exist due to (violations of) listener’s expectations.

With regard to the methodology of evaluating models of expression, we assign great importance to the systematic collection of empirical data, experimentally manipulating the relevant parameters. For instance, in our research on expressive vibrato [5, 6], we explicitly control for global tempo to reveal how it is adapted to the duration of notes. And we record repeated performance to get a better grip on consistency (e.g. to be able to separate between intended and non-intended expressive information). Similarly, in our studies on piano performances (e.g., [7]), only careful experimental manipulation of a few parameters (like global tempo, or the addition or removal of one note) will give a precise insight in the underlying mechanisms that we need to reveal in order to make better computer music editing software or music generation systems. Blindly examining very large samples of music performance is clearly not an alternative to this.

And, finally, in our work in rhythm perception, we put quite some effort in developing methods that allow us to investigate the concept of *performance space*, abstracting from individual examples. The idea here is to consider all possible interpretations, including musical and unmusical ones, in a variety of styles. While currently we only applied this approach to relatively short fragments of music [8], we find this method a more systematic and insightful alternative for randomly grown corpora of music performances. In addition, studying the perception of rhythm is also a way to identify the constraints on expressive timing in music performance (instead of focusing on an ideal or unique performance) as such avoiding the notion of a "correct" performance, which is an important advantage that allows for models to be elaborated independent of performance style.

## References

[1] Documents on `http://www.nici.kun.nl/mmm` under heading "Research".

[2] Sundberg, J., Friberg, A., and Frydén, L. (1991) Common Secrets of Musicians and Listeners: An Analysis-by-Synthesis Study of Musical Performance. In P. Howell, R. West & I. Cross (eds.). *Representing Musical Structure*. London: Academic Press.

[3] Widmer, G. (2001) Using AI and Machine Learning to Study Expressive Music Performance: Project Survey and First Report. *AI Communications*, 14.

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

[5] Desain, P. & Honing, H. (1996) Modeling Continuous Aspects of Music Performance: Vibrato and Portamento [ICMPC Keynote address], *Proceedings of the International Music Perception and Cognition Conference*. CD-ROM, Montreal: McGill University.

[6] Rossignol, S., Desain, P. & Honing, H. (2001). State-of-the-art in fundamental frequency tracking. *Proceedings of the Workshop on Current Research Directions in Computer Music.* Barcelona: UPF.

[7] Timmers, R., Ashley, R., Desain, P., Honing, H., and Windsor, L. (in press) Timing of ornaments in the theme of Beethoven's Paisiello Variations: Empirical Data and a Model. *Music Perception*.

[8] Desain, P. & Honing, H. (submitted). *The Perception of Time: The Formation of Rhythmic Categories and Metric Priming*. See `http://www.nici.kun.nl/mmm/time.html`