frontiers in Digital Humanities

REVIEW
published: 24 October 2018
doi: 10.3389/fdigh.2018.00025

Computational Models of Expressive Music Performance: A Comprehensive and Critical Review

Carlos E. Cancino-Chacón[1,2]*, *Maarten Grachten*[2], *Werner Goebl*[3] *and Gerhard Widmer*[1,2]

[1] *Austrian Research Institute for Artificial Intelligence, Vienna, Austria,* [2] *Department of Computational Perception, Johannes Kepler University Linz, Linz, Austria,* [3] *Department of Music Acoustics – Wiener Klangstil, University of Music and Performing Arts Vienna, Vienna, Austria*

Expressive performance is an indispensable part of music making. When playing a piece, expert performers shape various parameters (tempo, timing, dynamics, intonation, articulation, etc.) in ways that are not prescribed by the notated score, in this way producing an *expressive* rendition that brings out dramatic, affective, and emotional qualities that may engage and affect the listeners. Given the central importance of this skill for many kinds of music, expressive performance has become an important research topic for disciplines like musicology, music psychology, etc. This paper focuses on a specific thread of research: work on *computational* music performance models. Computational models are attempts at codifying hypotheses about expressive performance in terms of mathematical formulas or computer programs, so that they can be evaluated in systematic and quantitative ways. Such models can serve at least two purposes: they permit us to systematically study certain hypotheses regarding performance; and they can be used as tools to generate automated or semi-automated performances, in artistic or educational contexts. The present article presents an up-to-date overview of the state of the art in this domain. We explore recent trends in the field, such as a strong focus on data-driven (machine learning) approaches; a growing interest in interactive expressive systems, such as conductor simulators and automatic accompaniment systems; and an increased interest in exploring cognitively plausible features and models. We provide an in-depth discussion of several important design choices in such computer models, and discuss a crucial (and still largely unsolved) problem that is hindering systematic progress: the question of how to *evaluate* such models in scientifically and musically meaningful ways. From all this, we finally derive some research directions that should be pursued with priority, in order to advance the field and our understanding of expressive music performance.

Keywords: music performance, music expression, computational modeling, machine learning, generative systems, evaluation

OPEN ACCESS

Edited by:
Mark Brian Sandler,
Queen Mary University of London,
United Kingdom

Reviewed by:
Peter Schubert,
McGill University, Canada
Roger B. Dannenberg,
Carnegie Mellon University,
United States

***Correspondence:**
Carlos E. Cancino-Chacón
carlos.cancino@ofai.at

Specialty section:
This article was submitted to Digital Musicology, a section of the journal Frontiers in Digital Humanities

Received: *28 February 2018*
Accepted: *04 October 2018*
Published: *24 October 2018*

Citation:
Cancino-Chacón CE, Grachten M, Goebl W and Widmer G (2018) Computational Models of Expressive Music Performance: A Comprehensive and Critical Review. Front. Digit. Humanit. 5:25. doi: 10.3389/fdigh.2018.00025

1. INTRODUCTION

The way a piece of music is performed is a very important factor influencing our enjoyment of music. In many kinds of music, particularly Western art music, a good performance is expected to be more than an exact acoustic rendering of the notes in the score. Performers have certain liberties in shaping various parameters (e.g., tempo, timing, dynamics, intonation, articulation, etc.) in ways that are not prescribed by the notated score, and are expected to use these to produce an *expressive* rendition of the piece in question. This applies not only to classical music, where interpretation and performance are perpetual topics of artistic and aesthetic discussion, but to virtually all kinds of music. *Expressive performance*, as we will call it in the following, is known to serve several purposes: foremost, to express and communicate the performer's understanding of structure and affective content ("meaning") inherent in a composition, and in this way to bring out dramatic, affective, and emotional qualities that, in the best case, may engage and affect the listeners emotionally. Expert musicians learn (mostly implicit) performance rules through many years of focused and intensive practice and intellectual engagement with music. Given the central importance of this subtle art, the principles behind, and processes involved in, expressive performance should be a central topic of research in music and music psychology.

The systematic study of expressive music performance is a relatively young field, starting in the first half of the twentieth Century with first quantitative investigations (Binet and Courtier, 1896; Seashore, 1938). The second half of the twentieth Century saw an increased interest in looking at performance from the perspective of music psychology and cognition (Clynes, 1969, 1986, 1987; Gabrielsson, 1974; Longuet-Higgins and Lee, 1982, 1984; Palmer, 1996). The field gained more attraction in the late 1980's, with advances in computers and electronic instruments, which facilitated more precise data capturing (Kirke and Miranda, 2013). Music performance science is a highly inter-disciplinary field, and a thorough review of the state of the art of the full field is outside the scope of this paper. We refer the interested reader to the very comprehensive review articles by Palmer (1997) and Gabrielsson (1999, 2003). For a review of performance research from a musicological point of view see Rink (1995, 2002, 2003). For philosophical perspectives on expressiveness in music, we refer the reader to Davies (1994, 2001).

The present article focuses on a narrower and more specific topic: *computational models of expressive performance*, that is, attempts at codifying hypotheses about expressive performance—as mappings from score to actual performance—in such a precise way that they can be implemented as computer programs and evaluated in systematic and quantitative ways. This has developed into a veritable research field of its own over the past two decades, and indeed the present work is not the first survey of its kind; previous reviews of computational performance modeling have been presented by De Poli (2004), Widmer and Goebl (2004), and Kirke and Miranda (2013).

The new review we offer here goes beyond these earlier works in several ways. In addition to providing a comprehensive update on newer developments, it is somewhat broader, covering also semi-automatic and accompaniment systems, and discusses the components of the models in more detail than previous reviews. In particular, it provides an extended critical discussion of issues involved in model choices—particularly the selection and encoding of input features (score representations) and output parameters (expressive performance dimensions)—and the evaluation of such models, and from this derives some research directions that should be pursued with priority, in order to advance the field and our understanding of expressive music performance. As in earlier papers, we focus on models for notated music, i.e., music for which a musical score (a symbolic description of the music) exists. This includes most Western art music. A review of models of expressive performance for non-western or improvised music traditions is outside the scope of this work.

The rest of this text is organized as follows: Section 2 introduces the concept of computational music performance models, including possible motivations, goals, and general model structure. Section 3 attempts to give a comprehensive overview of the current state of the art, focusing on several current trends in the field. Section 4 offers a critical discussion of crucial modeling aspects, and offers a critical view on the ways in which performance models are currently evaluated. Section 5 concludes the paper with a list of recommendations for future research.

2. COMPUTATIONAL MODELING OF EXPRESSIVE PERFORMANCE

2.1. Motivations for Computational Modeling

Formal and computational models of expressive performance are a topic of interest and research for a variety of scientific and artistic disciplines, including computer science, music psychology, and musicology, among others. Accordingly, there is an wide variety of motivations for this kind of modeling. Broadly speaking, we can categorize these motivations into two groups: on the one hand, computational models can be used as an analytical tool for understanding the way humans perform music; on the other hand, we can use these models to generate (synthesize) new performances of musical pieces in a wide variety of contexts.

As analysis tools, computational models permit us to study the way humans perform music by investigating the relationship between certain aspects of the music, like the phrase structure, and aspects of expressive performance, such as expressive timing and dynamics. Furthermore, they allow us to investigate the close relationship between the roles of the composer, the performer, and the listener (Kendall and Carterette, 1990; Gingras et al., 2016). Expressive performance and music perception form a feedback loop in which expressive performance actions (like a slowing down at the end of a phrase) are informed by perceptual constraints or expectations, and the perception of certain musical constructs (like grouping structure) is informed by the way the music is performed (Chew, 2016). In this way, computational

models could also be used to enhance our understanding of the way humans listen to music.

On the other hand, computational performance models can be interesting in their own right, as tools for generating automatic or semi-automatic performances. In this case, a generative system might attempt to produce a *convincing* or *human-like* performance of a piece of music given its score (Friberg et al., 2006; Grachten and Widmer, 2012; Okumura et al., 2014) or try to play alongside human musicians, not only tracking their expressive performance but also introducing its own expressive nuances (Xia et al., 2015; Cancino-Chacón et al., 2017a). Such systems might have many applications, including realistic playback in music typesetting tools (such as Finale or MuseScore) and automatic expressive accompaniment for rehearsing. Also, there is now a renewed interest in systems that automatically generate (i.e., compose) music. As pointed out by Herremans et al. (2017), automatic performance systems might be an important component in making automatic music generation usable by the general public.

From a philosophical perspective, the idea of musical expressiveness presents a number of issues (Davies, 2001). Among these is the fundamental question of whether an expressive performance *can be fully captured* using numerical descriptors. For example, Bergeron and Lopes (2009) discuss whether a complete sonic description of the music without any visual component can fully convey the expressivity of music. That hearing *and* seeing a musical performance provides for a richer experience[1] is an interesting and plausible hypothesis, but this question goes beyond the scope of the present article. In any case, it should be undisputed that there *is* more than enough expressivity to be perceived—and thus also modeled—from just a sonic representation; after all, listening to a recorded performance is still the predominant way of enjoying music, and it can be a rewarding experience.

2.2. Components of the Performance Process

In her seminal review, Palmer (1997) groups the reported work in three sections that can be taken to reflect the human cognitive processes involved in performing a piece of notated music:

1. **Interpretation.** According to Kendall and Carterette (1990), music performance is a communication process in which information (emotional and semantic content of the piece) flows from the composer to the performer to the listener. We note here that these should be regarded as roles rather than agents, since, for example, the composer and performer may be embodied by the same person. An important task for the performer is to determine how to convey the message from the composer to the listener. Palmer refers to *interpretation* as the act of arriving at a conceptual understanding of structure and emotional content or character of a given piece, in view of a planned performance. Examples of relevant structural aspects are the grouping and segmentation of sequences into smaller subsequences to form hierarchical levels—such as those proposed by Lerdahl and Jackendoff (1983) in their Generative Theory of Tonal Music (GTTM).
2. **Planning.** Through planning the performer decides how to relate the syntax of musical structure to expression through style-specific actions and constraints. Such actions include, e.g., the use of arch-like patterns in dynamics and tempo to elucidate the phrasing structure (Todd, 1992; Friberg and Sundberg, 1999).
3. **Movement.** Finally, a performer needs to transform a performance plan into a concrete execution of the piece by means of physical movement. These movements can be seen as embodied human–music interactions which have an impact on the way humans perform and perceive music (Leman et al., 2017a).

In section 4.2, we will present a discussion on how different aspects and methods of computational modeling of performance fit into these categories, as well as the implications of the choice for the modeling.

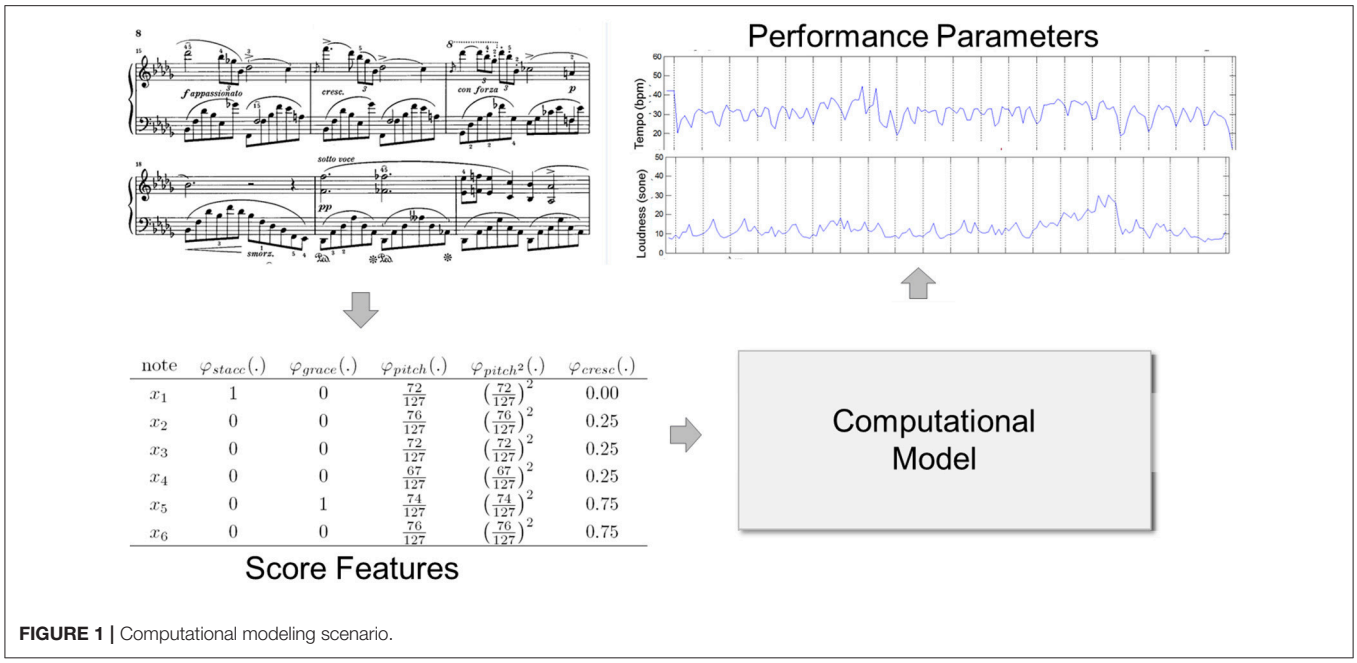
2.3. Components of Computational Models

Ideally, a full computational model of expressive performance should cover all three of the above aspects. However, the models described in the literature so far focus almost exclusively on the *planning* process, conceptualizing it as a *mapping* from a given score to specific patterns in various performance parameters (e.g., timing or dynamics) and, eventually, to an acoustic realization of the piece (De Poli, 2004). Thus, in the remainder of this review we will adopt this (admittedly too) limited view and discuss existing performance models in this context.

Kirke and Miranda (2013) proposed a generic framework for describing research in expressive performance. In the present article, we adopt a simplified version of this framework involving three main components by which computational performance models (in the limited sense as explained above) can be characterized. The resulting simple modeling scenario is shown in **Figure 1**, along with a fragment of a musical score.

By *score features*—which are the inputs to the computational model—we denote descriptors used to represent a piece of notated music. Some of these features may be given directly by the score (such as notated pitches and durations), while others may be computed from the score in more or less elaborate ways, by some well-defined procedure (such as the cognitive features discussed in section 3.3). Features can range from low-level descriptors such as (MIDI) pitches (Friberg et al., 2006; Grindlay and Helmbold, 2006; Cancino Chacón and Grachten, 2015) and hand-crafted features, like encodings of metrical strength (Grindlay and Helmbold, 2006; Giraldo S. and Ramírez, 2016); to cognitively inspired features, like Narmour's Implication-Realization (IR) descriptors (Flossmann et al., 2013; Giraldo S.I. and Ramirez, 2016), or even features learned directly from the score using unsupervised machine learning (Grachten and Krebs, 2014; van Herwaarden et al., 2014). The process of extracting score features corresponds to the *Music/Analysis* module in Kirke and Miranda (2013)'s framework and can be seen as at least partly related to Palmer's *Interpretation* aspect (see above).

[1] Or, as Bergeron and Lopes (2009) (quoting Robert Schumann) put it: "if Liszt played behind a screen, a great deal of poetry would be lost."



note	$\varphi_{stacc}(.)$	$\varphi_{grace}(.)$	$\varphi_{pitch}(.)$	$\varphi_{pitch^2}(.)$	$\varphi_{cresc}(.)$
x_1	1	0	$\frac{72}{127}$	$\left(\frac{72}{127}\right)^2$	0.00
x_2	0	0	$\frac{76}{127}$	$\left(\frac{76}{127}\right)^2$	0.25
x_3	0	0	$\frac{72}{127}$	$\left(\frac{72}{127}\right)^2$	0.25
x_4	0	0	$\frac{67}{127}$	$\left(\frac{67}{127}\right)^2$	0.25
x_5	0	1	$\frac{74}{127}$	$\left(\frac{74}{127}\right)^2$	0.75
x_6	0	0	$\frac{76}{127}$	$\left(\frac{76}{127}\right)^2$	0.75

FIGURE 1 | Computational modeling scenario.

An *expressive parameter*—the output of a model—is a numerical encoding of an aspect of expressive performance. Since most systems deal with piano music, the most common descriptors relate to loudness, expressive tempo and timing, and articulation (Widmer and Goebl, 2004; Kirke and Miranda, 2013), but of course they can also include other parameters like timbral features (Raphael, 2009; Ohishi et al., 2014) and intonation (Clynes, 2005), or higher-level patterns such as "pulse microstructure" (Clynes, 1987). The expressive parameters correspond to the outputs of the *Performance Knowledge* module in Kirke and Miranda (2013)'s framework. Section 4.1 presents a critical review of the choices involved in selecting and encoding these parameters.

A *computational model* then, in our context, is any computable function that maps score features to expressive parameters or, to be more precise, can make a *prediction* of the values of expressive parameters, given a score (represented via score features) as input. In music performance modeling, this is typically done by means of mathematical functions (probabilistic models, artificial neural networks, etc.) (Teramura et al., 2008; Kim et al., 2010; Grachten and Widmer, 2012) or by means of rules (Friberg et al., 2006; Canazza et al., 2015). Some of these models can be *trained* using a dataset of expressive performances. The model/function corresponds to the *Performance Knowledge* and the *Performance Context* in Kirke and Miranda (2013)'s framework; the training of the model corresponds to the *Adaptation Process*, and the datasets are the *Performance Examples*.

3. A SYNOPSIS OF CURRENT STATE AND RECENT TRENDS

In this section we discuss some of the recent trends in computational performance modeling. This brief overview is meant as an update to earlier review papers by De Poli (2004), Widmer and Goebl (2004), and Kirke and Miranda (2013). In the following, we will refer to a model as *static* if its predictions only depend on a single event in time (e.g., linear regression, feed forward neural networks), and *dynamic* if its predictions can account for time-dependent changes (e.g., hidden Markov models or recurrent neural networks).

3.1. Data-Driven Methods for Analysis and Generation of Expressive Performances

A first noteworthy trend in recent research is an increasing focus on *data-driven* approaches to performance modeling, relying on *machine learning* to infer score-performance mappings (and even the input score features themselves) from large collections of real data (scores and performances). This is in contrast to *rule-based* approaches where performance rules are manually designed, based on musical hypotheses.

An important example of the rule-based variety is the KTH model (Sundberg et al., 1983; Friberg et al., 2006), developed at the Royal Institute of Technology (KTH) in Stockholm. These rules were developed and evaluated through an iterative *analysis-by-synthesis* approach involving judgments by experts and listening tests. A performance is shaped by a (linear) combination of the effects of the rules, which the user can weigh individually. The KTH model has been implemented as a software package called *Director Musices* (DM) (Friberg et al., 2000; Masko et al., 2014). Recent versions of the KTH model include cognitively motivated rules regarding musical accents (Bisesi et al., 2011). Friberg and Bisesi (2014) study the use of the system for modeling stylistic variations for Baroque, Romantic and Contemporary art music. The KTH model won the first prize at the RenCon, a competition for computational models of performance, in 2004 (see also section 4.3 below).

While early data-driven approaches (Widmer, 1995, 1996, 2000, 2003; Widmer and Tobudic, 2002) aimed at learning explicit performance rules at various structural levels (from individual notes to higher phrasing levels), using methods like instance-based learning and inductive logic programming, recent advances in machine learning—in particular relating to probabilistic graphical models and (deep) neural networks—have led to a surge of such methods in computational performance modeling, which will be reviewed in the following.

3.1.1. Data-Driven Methods for Performance Analysis

As tools for analysis, computational methods can be used for several purposes, including studying the relationship between structural aspects of the score and specific aspects of a performance, or for comparing expressive renderings by different performers.

3.1.1.1. Explaining/modeling aspects of performance

An important question in analyzing expressive performance is determining the likely "causes" of observed performance patterns, i.e., structural or other aspects of a piece that would "explain" why a certain passage was (or needs to be) played in a certain way. By analyzing large amounts of data, data-driven methods can find systematic relations between measured performance aspects (e.g., changes in tempo and dynamics) and various structural aspects of a musical score (e.g., pitch content, metrical, and phrasing structure), notated performance indications (e.g., dynamics markings such as ***piano*** and ***forte*** or articulation marks such as ***legato*** slurs and ***staccato***), or even aspects related to our perception of music, like melodic expectation (as far as we are able to infer or compute these in a reliable way from a score).

Examples of such approaches include the work by Kosta et al. (2014, 2015, 2016), who focus on the relationship between dynamics markings and expressive dynamics, and the *Basis Function Model* (Grachten and Widmer, 2012)—a framework that encodes score properties via so-called basis functions—which attempts to quantify the contribution of a variety of score descriptors (such as pitch, metrical position, and dynamics markings) to expressive dynamics (Cancino-Chacón C.E. et al., 2017) and timing (Grachten and Cancino-Chacón, 2017). Fu et al. (2015) study timing deviations in arpeggiated chords with statistical methods. Gingras et al. (2016) and Cancino-Chacón et al. (2017b) focus on linking information-theoretic features quantifying the expectation of musical events in listeners, to expressive timing. Caramiaux et al. (2017) study performers' skill levels through variability in timing and features describing finger motion. Marchini et al. (2014) study the use of score features describing horizontal (i.e., melodic) and vertical (i.e., harmonic) contexts for modeling dynamics, articulation, and timbral characteristics of expressive ensemble performances, focusing on string quartets. Using machine learning and feature selection techniques, Giraldo S.I. and Ramirez (2016) and Bantula et al. (2016) evaluate a number of score descriptors in modeling expressive performance actions for jazz guitar and jazz ensembles, respectively.

A second form of analysis focuses on specific patterns and characteristics in curves of expressive parameters. This includes work on methods for visualizing expressive parameters and their characteristics (Langner and Goebl, 2003; Grachten et al., 2009; Chew and Callender, 2013), on inferring performance strategies like phrasing from expressive timing (Chuan and Chew, 2007) or dynamics (Cheng and Chew, 2008), and clustering of patterns of (phrase-level) tempo variations (Li et al., 2014, 2015, 2016, 2017). The results obtained with such methods support the existence of common performance strategies (Cheng and Chew, 2008; Li et al., 2014; Kosta et al., 2016). Quantitative studies on the contribution of various score features to expressive parameters reveal well-known relationships, like the importance of pitch (height) for predicting expressive dynamics, and the relationship between metrical features and timing deviations. At the same time, some results indicate that aspects of performance (like expressive tempo and dynamics) might be related in more than one way to structural aspects of the music, e.g., phrasing has been shown to be related to dynamics (Cheng and Chew, 2008) or timing (Chuan and Chew, 2007). An interesting finding is the importance of features and models that allow for describing the musical contexts, both horizontal (temporal) (Gingras et al., 2016; Kosta et al., 2016; Grachten and Cancino-Chacón, 2017) and vertical (i.e., harmonic) (Marchini et al., 2014).

3.1.1.2. Comparing expressive performances

A different approach to analyzing expressive performances is to compare different renditions of the same piece by different performers, which allows for studying of commonalities and differences in performance strategies. Some of the work in this direction follows an unsupervised approach, which does without any score information, instead focusing on comparing aligned curves of expressive parameters that encode the performances. Sapp (2007, 2008) presents a graphical approach and explores different metrics for comparing collections of performances of the same piece. Liem and Hanjalic (2011) and Liem et al. (2011) propose a method for comparing expressive timing by studying alignment patterns between expressive performances of the same piece using standard deviations and entropy, respectively. Liem and Hanjalic (2015) use Principal Components Analysis (PCA) to localize areas of cross-performance variation, and to determine similarities between performances in orchestral recordings. Peperkamp et al. (2017) present a formalization of relative tempo variations that considers performances as compositions of functions which map performance times to relevant feature spaces. Rather than focusing on a single aspect like dynamics or timing, Liebman et al. (2012) present a phylogenetic approach that compares and relates performances of two solo violin pieces by different performers, using performance descriptors like bowing, tempo changes, and phrase duration. Grachten et al. (2017) use Basis Function models to assess the contribution of score features pertaining to individual orchestral instruments to the overall loudness curves, using differential sensitivity analysis, which allows for graphically comparing pairs of recordings of the same piece by different conductors and orchestras. Methods for comparing performances can be used for identifying musicians by their individual performance styles. This has

been demonstrated for violinists (Molina-Solana et al., 2008, 2010a), saxophone players (Ramírez et al., 2007), and pianists (Stamatatos and Widmer, 2005; Saunders et al., 2008; Grachten and Widmer, 2009; Molina-Solana et al., 2010b).

Computational methods for performance comparison have produced some interesting results. They support the idea of common performance strategies across performers, as well as consistent individual differences between performers. Furthermore, they seem to support musicologically plausible hypotheses such as the change in playing style over the years and differences between mainstream and historically informed performance styles, while only providing weak evidence for the existence of "performance schools" (Liebman et al., 2012). The formalization of tempo proposed by Peperkamp et al. (2017) provides an interesting mathematical constraint on tempo curves as convex linear combinations of tempo variation functions.

In spite of all this, there has been only little progress in really understanding the way humans perform music expressively. An important issue is that effectively all studies are limited to small datasets (at least compared to other machine learning domains) that only contain a small selection of pieces and/or performers. This raises the question how well (or if) the insights gained from these studies generalize to other performers or kinds of music. Also, most models rely on features that capture only small local contexts, so that the resulting models cannot properly account for long temporal dependencies that might be important for understanding global aspects of performance expression. We still largely fail to understand how to model long-term, non-contiguous relationships in complex music. The hope is that recent advances in (deep) machine learning may open new possibilities here (Widmer, 2017).

3.1.2. Data-Driven Methods for Performance Generation

In this section we examine recent work on autonomous generative systems. While computational methods for analysis tend to focus on explaining a single aspect of expression, generative models most commonly have to consider more expressive parameters, with expressive tempo/timing and dynamics being the most popular. As previously discussed, a major trend in this area is the use of complex probabilistic approaches and the use of neural-network-based methods.

3.1.2.1. Probabilistic Approaches

In a nutshell, probabilistic approaches describe expressive performances by modeling the probability distribution of the expressive parameters given the input score features. **Table 1** presents some of the recent probabilistic performance systems in terms of their computational models, expressive parameters, and score features. Please note that the column relating to score features in **Table 1** is not exhaustive, given the potentially large number of features used in each model.

While each model conceptualizes a music score and its corresponding expressive performance differently, there are some interesting commonalities. Several researchers use variants of Hidden Markov Models (HMMs) to describe the temporal evolution of a performance, such as Hierarchical HMMs (Grindlay and Helmbold, 2006), Dynamic Bayesian Networks (DBNs) (Widmer et al., 2009; Flossmann et al., 2011, 2013), Conditional Random Fields (CRFs) (Kim et al., 2010, 2011, 2013), or Switching Kalman Filters (Gu and Raphael, 2012). Furthermore, most models assume that the underlying probability distribution of the expressive parameters is Gaussian (Grindlay and Helmbold, 2006; Teramura et al., 2008; Gu and Raphael, 2012; Flossmann et al., 2013; Okumura et al., 2014). A different approach is taken by Kim et al. (2013) and Moulieras and Pachet (2016), who use maximum entropy models to approximate the underlying probability distributions. While most models focus on Western classical music, Moulieras and Pachet (2016) focus on expressive performance of jazz piano.

In terms of expressive parameters, most models describe expressive dynamics using the note-wise MIDI velocity. This is mostly done by either making predictions from a static model (Teramura et al., 2008), focusing only on monophonic melodies (Grindlay and Helmbold, 2006; Gu and Raphael, 2012; Moulieras and Pachet, 2016), or assuming a decomposition of the piece into monophonic streams (Kim et al., 2013; Okumura et al., 2014). On the other hand, there seems to be a variety of descriptors for expressive tempo and timing, with some models focusing on the *inter-beat interval* (IBI; a local estimation of the time between consecutive beats) or inter-onset interval (IOI; the time interval between consecutive onsets), some on the local *beats per minute* (bpm; the inverse of the IBI). Other models target local changes in their expressive parameters, by means of modeling their first differences[2]. Most models use a combination of low-level features—pitch, onset, and duration of notes, as well as encodings of dynamics and articulation markings—and high-level features describing musically meaningful structures, such as metrical strength. Most systems only model expressive parameters independently, and the few exceptions focus on specific combinations of parameters, such as the ESP system (Grindlay and Helmbold, 2006) that jointly models tempo and tempo changes, but describes dynamics independently, and the model by Moulieras and Pachet (2016), which jointly models timing and dynamics.

3.1.2.2. Artificial neural network-based approaches

Broadly speaking, artificial neural networks (ANNs) can be understood as a family of mathematical functions describing hierarchical non-linear transformations of their inputs. The success of ANNs and deep learning in other areas, including computer vision and natural language processing (Goodfellow et al., 2016) and music information retrieval (Humphrey et al., 2012; Schlüter, 2017), has motivated their use for modeling expressive performance in recent years. A description of systems using ANNs for performance generation is given in **Table 2**. As in the case of probabilistic models, the list of score features for each model is not exhaustive.

Some ANN-based approaches use feed forward neural networks (FFNNs) to predict expressive parameters as a function

[2]For a parameter p_i, the first (finite) difference refers to $\Delta p_i = p_i - p_{i-1}$.

TABLE 1 | Probabilistic models for performance generation.

System	Computational model	Expressive parameters	Score features
ESP Grindlay and Helmbold, 2006	Hierarchical HMMs	• Tempo: log IBI ratio and its first difference. • Dynamics: MIDI velocity	• Low-level: melodic interval. • High-level: metrical hierarchies and phrase structure (annotated)
NAIST Teramura et al., 2008	Gaussian Processes	• Dynamics: MIDI velocity, • Timing: onset deviations (bpm) • Articulation: offset deviations (bpm)	• Low-level: pitch, duration, dynamics markings. • High-level: time signature, melody, relative pitch.
YQX Flossmann et al., 2011	DBNs (+ learned articulation rules from Widmer, 2003)	• Tempo: low frequency components of the log-IOI. • Timing: high-frequency components of log-IOI. • Articulation: duration ratio • Dynamics: log-MIDI velocity.	• Low-level: Pitch, contour. • High-level: Narmour's IR features, harmonic consonance
Gu & Raphael Gu and Raphael, 2012	Switching Kalman Filter	• Tempo: IOI • Dynamics: MIDI velocity	• Low-level: position in the score
Polyhymnia Kim et al., 2013	3 CRFs modeling the highest and lowest voices (m) and a harmony model (h) for the inner voices (+ rules for dynamics markings and ornaments from statistical analysis)	• Tempo: log IBI ratio (m) • Timing: onset deviations • Articulation: note-wise duration ratio (m, h). • Dynamics: log MIDI velocity ratio (m, h).	• Low-level: pitch (m), duration (m,h), interval to outer voices (h) • High-level: metrical strength (m)
Laminae Okumura et al., 2014	Performance cases modeled by Gaussian distributions and Tree-based clustering + first order Markov model.	Voice-wise first differences of • Tempo: ave. bpm per beat • Timing: onset deviations (bpm) • Articulation: duration ratio • Dynamics: MIDI velocity	• Low-level: pitch class, octave, dynamics markings • High-level: phrasing, voice (human annotated)
SONY Moulieras and Pachet, 2016	Maximum Entropy model	• Timing: onset deviation. • Dynamics: MIDI velocity	• Low-level: onset position in the bar

of the score features (Bresin, 1998; Cancino Chacón and Grachten, 2015; Giraldo S. and Ramírez, 2016). These systems tend to compensate for the static nature of FFNNs by including score features that describe some of the musical context of a performed note (e.g., features describing the adjacent rhythmic/melodic context). Other approaches use recurrent neural networks (RNNs), a class of dynamic ANNs, to model temporal dependencies between score features and expressive parameters (Cancino Chacón and Grachten, 2016; Cancino-Chacón C.E. et al., 2017). While early versions of the *Basis Mixer*, an implementation of the Basis Function model, used a simple linear model (Grachten and Widmer, 2012; Krebs and Grachten, 2012), current incarnations (Cancino Chacón and Grachten, 2016) use both FFNNs and RNNs as non-linear function classes, either in the form of deterministic ANNs, or as Gaussian mixture density networks—probabilistic ANNs in which the outputs of the network parameterize the joint probability distribution of a Gaussian Mixture Model.

Neural network models closely follow probabilistic approaches in terms of their expressive parameters. Instead of expecting a human-annotated or heuristically computed decomposition of a polyphonic score into monophonic streams, Cancino Chacón and Grachten (2016) decompose a performance into a series of sequential and non-sequential expressive parameters, which permits to model both temporal trends in dynamics and tempo, and local effects (note-level) in timing, articulation, and dynamics deviations. Giraldo S. and Ramírez (2016) present an approach for modeling jazz guitar performances which allows for describing not only dynamics and timing, but also ornamentation.

In terms of input features, most ANN models again tend to rely on a combination of low-level hand-crafted features describing local aspects describing individual notes, and some higher-level features relating to structural properties of the music. On the other hand, some researchers have tried to use ANNs to automatically *learn* features from low-level representations of the score. Grachten and Krebs (2014) and van Herwaarden et al. (2014) use Restricted Boltzmann Machines (RBMs), a probabilistic class of neural networks, to learn features from note-centered piano rolls in an unsupervised fashion.

3.2. Expressive Interactive Systems: Models for Technology-Mediated Performance

A second major trend that can be observed in recent years is a growing interest in developing human–computer interaction systems that generate expressive music performances. Rowe (1992) proposed a terminology for categorizing interactive music systems in three dimensions: *score-driven* vs. *performance-driven*, referring to whether the system follows a musical score or responds to a human performance; *instrument paradigm* vs. *player paradigm*, if the system is meant for solo or ensemble performances; and *transformative* vs. *generative*

TABLE 2 | Neural network based models.

System	Computational model	Expressive parameters	Score features
Bresin Bresin, 1998	FFNNs	• Tempo: IOI • Articulation: performed duration • Dynamics: change in loudness	• Low-level: pitch, duration, melodic interval • High-level: encodings of conditions for KTH rules, like leap articulation, melodic charge, articulation repetition
Unsupervised RBM Grachten and Krebs, 2014; van Herwaarden et al., 2014	RBM (features) + Linear Models	• Dynamics: MIDI velocity	• Low-level: note-centered piano-roll, MIDI-velocity history
Giraldo & Ramírez Giraldo S. and Ramírez, 2016	• Ornamentation (classification): FFNNs, decision trees, SVMs, k-NN • Timing, Articulation and Dynamics (regression): FFNNs, regression trees, SVMs, k-NN	• Timing: onset deviation • Articulation: duration ratio • Dynamics: energy ratio • Ornamentation	• Low-level: pitch, duration, position in bar • High-level: Narmour's IR features, key, metrical position, phrase position
Basis Mixer Grachten and Widmer, 2012; Cancino Chacón and Grachten, 2016	• Onset-wise model: RNNs • Note-wise: models FFNNs Models can be either deterministic NNs or probabilistic GMDNs	• Tempo: log-IBI (onsetwise) • Timing: Onset deviations (notewise) • Articulation log-duration (notewise) • Dynamics: MIDI velocity trend (onsetwise) and deviations (onsetwise)	Encoding of score aspects through basis functions. • Low-level: pitch, duration, dynamics and articulation markings, position in bar • High-level: tonal tension, harmonic analysis

vs. *sequenced*, describing how the system renders the music. The focus of the present survey is on expressive score-driven systems; performance-driven approaches such as interactive improvisation systems are beyond the scope of this paper. A more thorough review of interactive systems is provided by Chew and McPherson (2017).

3.2.1. Conductor Systems

Conductor systems allow the user to shape a solo performance in real-time, and in Rowe's taxonomy would classify as score-driven, instrument paradigm, transformative systems. Such models divide the rendering of an expressive performance into three parallel subtasks: capturing the input from the user, mapping such input to expressive parameters, and providing feedback to the user in real time. **Table 3** shows several feedback and conductor models. For a more thorough review of feedback models we refer the reader to Fabiani et al. (2013).

Common ways for a user to control certain aspects of a performance are either via high-level semantic descriptors that describe the intended expressive character—often selected from some 2D space related to Russell (1980)'s valence—arousal plane (Friberg, 2006; Canazza et al., 2015); or via physical gestures, measured either through motion capture (Fabiani, 2011) or by using physical interfaces (Chew et al., 2005; Dixon et al., 2005; Baba et al., 2010). Some systems even attempt to provide a realistic simulation of conducting an orchestra (Baba et al., 2010; Fabiani, 2011).

Regarding the mapping of expressive intentions to performance parameters, some systems give the performer direct control of expressive parameters (e.g., tempo and MIDI velocity) via their input (Chew et al., 2005; Dixon et al., 2005). This allows for analyzing the way humans perform music (Chew et al., 2005, 2006). On the other hand, most systems use rule-based models, like the KTH model, to map the user input to expressive parameters (Friberg, 2006; Baba et al., 2010; Fabiani, 2011; Canazza et al., 2015).

3.2.2. Accompaniment Systems

Accompaniment systems are score-driven, player paradigm systems, according to Rowe's taxonomy. In order to successfully perform together with a human, accompaniment systems must solve three tasks: detecting the solo part, matching the detected input to the score, and generating an expressive accompaniment part (Dannenberg, 1984). The first tasks refers to the ability of the system to capture a human performance in real time (either from a microphone or a MIDI instrument) and identify the performed notes, while the second refers to matching these performed notes to notes in the score (also in the presence of errors). The third task involves generating an expressive accompaniment that adapts to the performance of the soloist. The first two tasks are commonly referred to as real-time *score following*. In this review we focus mostly on accompaniment systems for notated Western classical music. For perspectives on accompaniment systems for popular music, we refer the reader to Dannenberg et al. (2014).

Perhaps the most well-developed accompaniment systems are *Antescofo* (Cont, 2008; Cont et al., 2012) and *Music Plus One* (Raphael, 2001a,b, 2010). Antescofo is not only a polyphonic accompaniment system, but a synchronous programming language (i.e., a computer language optimized for real-time reactive systems) for electro-acoustical musical composition. Both systems solve the score following problem using dynamic probabilistic graphical models such as variants of HMMs and DBNs. *Eurydice* (Nakamura et al., 2013, 2014a, 2015a,b) is a robust accompaniment system for polyphonic music that allows for skips, repetitions and ornaments using hidden semi-Markov models.

In spite of the great progress in automatic accompaniment systems, Xia (2016) points out that most of the work

TABLE 3 | Feedback and conductor systems.

System	Computational model	User input	Feedback	Expressiveness controlled by the user
pDM Friberg, 2006	Rule based: • KTH model: user controlled weighting of the rules	Semantic descriptors (Russell's space)	Audio (MIDI)	• Tempo • Dynamics • Timing • Articulation
Home Conducting Friberg, 2005		Gestures	Audio (MIDI) and Visual (emotion)	
PerMORfer Fabiani, 2011		Gestures/ semantic descriptors (Russell's space)	Audio (modified recordings)	
Air worm Dixon et al., 2005	Direct control	MIDI theremin	Audio (MIDI) / Visual (performance worm)	• Tempo • Dynamics
ESP-Chew Chew et al., 2005, 2006	Direct control	Car controller interface (steering wheel, gas and brake pedals)	Audio / Visual (road simulation)	• Tempo (car control) • Dynamics (car acceleration)
Virtual Philharmony Baba et al., 2010	Rule based: • Linear models defining tempo adjusting heuristics	Physical Interface	Audio	• Tempo
Caro 2.0 Canazza et al., 2015	Rule based: • Naturalizer: rules controlling base performance • Expressivizer: user controlled expressive deviations using linear models	Semantic Descriptors (Russell's space)	Audio	• Tempo • Dynamics • Articulation

on accompaniment systems has focused on solving the score following problem, while overlooking the generation of expressivity in the accompaniment part, or mostly focusing on expressive timing. However, in recent years there has been a growing interest in expressive accompaniment systems. Specifically, Xia and Dannenberg (2015) and Xia et al. (2015) show how to use linear dynamical systems trained via spectral learning, to predict expressive dynamics and timing of the next score events. The *ACCompanion* (Cancino-Chacón et al., 2017a) is a system that combines an HMM-based monophonic score follower with a variant of the Basis Mixer to predict expressive timing, dynamics, and articulation for the accompaniment.

Another interesting recent development in accompaniment systems is embodied human–computer interactions through *humanoid robots* (Hoffman and Weinberg, 2011; Lim et al., 2012; Solis and Takanishi, 2013; Xia, 2016). These robots could be used for studying the way humans interact with each other.

3.3. Use of Cognitively Plausible Features and Models

A third clearly recognizable trend in performance modeling has to do with using features and models inspired by music psychology and cognition. While in early work (e.g., Widmer, 2003) the focus was on features rooted in music theory, such as scale degrees, melodic intervals, and metrical positions, recent years have seen an increased interest in developing descriptors that capture some aspects of the way humans—both listeners and performers—hear music. Wiggins et al. (2010) suggest that music theory is a kind of folk psychology, and thus, might benefit from being more explicitly informed by music cognition. The music cognition literature supports the hypothesis that much of the way we perform music is informed by the way we perceive music (Farbood, 2012; Goodchild et al., 2016).

3.3.1. Cognitively Inspired Features

From a computational modeling perspective, perhaps the most straightforward approach toward cognitively plausible models is to use features related to aspects of cognition. An important aspect of music cognition is the *expectation* of musical events. One of the most commonly used frameworks of music expectation in computational models of expression is Narmour's *Implication–Realization (IR) model* (Narmour, 1990). The IR model is a music-centered cognitive framework based on Gestalt theory that has emerged from Schenkerian analysis. It defines a number of patterns of listeners' ongoing expectations regarding the continuation of a melody, and how these expectations can be realized to different degrees by the actual continuation. Methods that include features based on IR include YQX (Flossmann et al., 2013), Giraldo S.I. and Ramirez (2016)'s approach to studying expression in Jazz guitar, and Marchini et al. (2014)'s approach for string quartets. More recently, there has been an interest to use information theoretic features computed using the *IDyOM model* (Pearce, 2005), a probabilistic model of statistical learning whose expectations have been shown to match human listeners'. Gingras et al. (2016) use entropy and information content as features to study expressive timing and perceived tension. This work supports Kendall and Carterette (1990)'s hypothesis regarding the communication between the composer, the performer, and the listener by linking expectation features, defined by the composer, to expressive timing, controlled by the performer, which is linked to perceived tension by the listener. Cancino-Chacón et al. (2017b) explore the use of these

information-theoretic features for actually predicting expressive tempo and dynamics of polyphonic piano music.

Other related cognitive aspects that influence the way humans perform music are the perception of tonality and tonal tension (Farbood, 2012; Chew, 2016). Several systems incorporate features relating to the tonal hierarchies defined by Krumhansl and Kessler's profiles (Krumhansl, 1990), including YQX (Flossmann et al., 2013; Giraldo S.I. and Ramirez, 2016) and the Basis Function models (Cancino Chacón and Grachten, 2016; Cancino-Chacón and Grachten, 2018; Cancino-Chacón C.E. et al., 2017), which also include tonal tension features by Herremans and Chew (2016) to predict expressive tempo, timing, dynamics, and articulation.

3.3.2. Cognitively Inspired Models

On the other hand, some researchers incorporate aspects of cognition as part of the design of the computational model itself. Recent versions of the KTH model includes some rules that refer to musical *accents* (Bisesi et al., 2011), local events that attract a listener's attention through changes in timing, dynamics, articulation, or pedaling; and musical tension rules (Friberg et al., 2006). The approach presented by Gu and Raphael (2012) decomposes expressive timing into discrete "behaviors": constant time, slowing down, speeding up and accent, which, as the authors argue, are more similar to the way human performers conceptualize expressive performance actions. *Polyhymnia* (Kim et al., 2013) uses 3 Conditional Random Fields (CRFs) to independently model the highest, lowest and internal voices. This decomposition allows the model to define the expressive parameters for the internal voices in terms of the outermost voices, following the hypothesis that listeners perceive the expressivity of the uppermost and lowermost voices more clearly than that of the inner voices (Huron and Fantini, 1989).

3.4. New Datasets

Data-driven modeling requires data—in the present case, corpora of music performances from which aspects of expressive performance can be readily extracted. This is a non-trivial problem, particularly for notated music, since performances not only have to be recorded (as audio or MIDI files), but they also have to be aligned to the corresponding score, so that we obtain a mapping between elements in the performance (temporal position in the case of audio recordings, or MIDI pitch, onset, and offset times) and elements in the score (score position, or an explicit mapping between a performed MIDI note and a note in the score). This is required in order to be able to calculate, e.g., expressive timing as the deviation of played on- and offsets from the corresponding time points implied by the score.

Table 4 presents some of the datasets used for modeling expressive performances in current research. Note that this list is not exhaustive; it is intended to give representative examples of the kinds of existing datasets. Performance datasets can be characterized along various dimensions, which are also shown in **Table 4**:

1. **Instrumentation and Solo/Ensemble Setting.** Performance datasets can include a variety of instruments, ranging from solo to ensemble performances. By far the most studied instrument in computational modeling is the piano, partially due to the existence of computer-controlled instruments such as the Bösendorfer SE/CEUS or the Yamaha Disklavier. However, recently there is also an increased interest in modeling ensembles (Marchini et al., 2014; Liem and Hanjalic, 2015; Grachten et al., 2017). For datasets relating to ensemble performances, an important distinction is between those which only reflect collective characteristics of the performance (as might be the case with datasets containing audio recordings where, e.g., timing and loudness of individual instruments are hard or even impossible to disentangle), and datasets where note-precise data is captured for each performer in the ensemble (as is the case with the Xia dataset described in Xia and Dannenberg, 2015).
2. **Performer(s).** Research on music performance has studied a wide range of musical skill levels, from novices and amateurs to advanced music students (i.e., enrolled in advanced undergraduate and post-graduate music programs), professionals and world-renowned performers. (Whether the performances by "world-renowned" performers are in any way better than those of "professional" performers, or who qualifies as a famous artist, are, of course, subjective matters.). Again, "performer" might not be singular, as some datasets relate to ensemble performances (cf. the Xia, Marchini, and RCO/Symphonic datasets in **Table 4**).
3. **Genre and Epoch** refer to the kind of music contained in the database and the period in which the music was composed. Most of the work on expressive performance modeling has focused on 19th century Western classical music. In **Table 4**, "Classical" denotes Western classical music and "Popular" denotes music genres such as jazz, folk, rock, and pop.
4. **Multiple Performances**. Different musicians perform the same pieces in different ways, and it is highly unlikely that the same performer would generate exactly the same performance more than once. Datasets that include multiple performances of the same piece by different performers allow modeling commonalities and systematic differences among performers, while multiple performances of a piece by the same performer could bring insights into the different aspects that contribute to specific realizations of expressive performance actions.
5. **Source** refers to whether the performances are taken from audio recordings or played on a computer-controlled instrument. Another related issue is whether the performances are recorded in front of a live audience, in a recording studio, or in a research lab. Such differences may have an influence on expressive parameters (Moelants et al., 2012).
6. **Alignment** refers to whether there is a mapping between elements in the performance and the score. (Producing such mappings is generally a very tedious task.) Alignments can be *note-wise*, i.e., individual performed notes are matched to their corresponding symbolic representations in the score; or *onset-wise*, where there is just a mapping between temporal position in the performance and score position.

In spite of the apparent richness and variety of data, it is important to raise awareness to some issues, like the fact that

TABLE 4 | Datasets of expressive performances.

Dataset	Performer	Pieces	Genre & epoch	Multiple performances		Source	Score alignment
				Different performer	Same performer		
PIANO							
Repp Repp, 1996	Advanced students	4	Classical: 1830–1920	Yes	Yes	Computer-controlled piano	Note-wise
Vienna 4x22 Goebl, 1999	Advanced students & professionals	4	Classical: 1780–1840	Yes	No	Computer-controlled piano	Note-wise
Batik/Mozart Widmer and Tobudic, 2002	Professional	30+	Classical: 1750–1800	No	No	Computer-controlled piano	Note-wise
Magaloff/Chopin Flossmann et al., 2010	World-renowned	150+	Classical: 1800–1850	No	No	Computer-controlled piano	Note-wise
Zeilinger/Beethoven Cancino-Chacón C.E. et al., 2017	Professional	15+	Classical: 1790–1830	No	No	Computer-controlled piano	Note-wise
Mazurka Sapp, 2007	World-renowned	45+	Classical: 1800–1850	Yes	Some	Audio recordings	Onset-wise
CrestMuse PEDB Hashida et al., 2008, 2017	World-renowned & professionals	40+	Classical: 1700–1900	No	Yes	Audio recordings and computer-controlled piano	Note-wise
LeadSheet Moulieras and Pachet, 2016	Professional	170+	Popular: 1950–2000	No	Yes	Computer-controlled piano	Note-wise
e-Piano Competition Simon et al., 2017	Professional	900+	Classical: 1700–2000	Yes	No	Computer-controlled piano	None
Xia Xia and Dannenberg, 2015	Advanced Students (duets)	3	Popular: 1800–1990	Yes	Yes	Computer-controlled piano	Note-wise
OTHER							
RCO/Symphonic Grachten et al., 2017	World-renowned (orchestra)	20+	Classical: 1800–1900	Yes	No	Audio recordings	Onset-wise
Marchini Marchini et al., 2014	Professional (string quartet)	1	Classical: 1790–1800	No	Yes	Audio recordings	Onset-wise

it is unlikely that the same performance would happen in two different kinds of rooms with different audiences (Di Carlo and Rodà, 2014). Furthermore, in the case of computer-controlled instruments, the mapping from MIDI velocities to loudness and timbre is dependent on the instrument.

But perhaps one of the most pressing issues is the availability of the datasets. Part of the impressive progress in other Artificial Intelligence domains is due to the availability of large standard datasets, which allow for comparing different approaches. In our case, however, only a few of the performance datasets are publicly available, often due to rights issues. (Of the datasets reported in **Table 4**, only CrestMuse PEDB, Xia, Vienna 4 x 22, Mazurka and the e-Piano competition datasets are publicly available). A noteworthy effort toward the compilation of large and varied performance datasets is being made by the CrestMuse group in Japan (Hashida et al., 2017), who not only provide a second edition of the PEDB database, but also have provided some tools for aligning MIDI performances to scores (Nakamura et al., 2017). A more in-depth review of methods for extracting information from performances can be found in Goebl et al. (2008) and Goebl and Widmer (2009).

In particular for score-based music it is of capital importance to have central datasets that combine score information, structural annotation, performance data and performance annotation. Crowd-sourcing platforms for creating and maintaining such databases are an avenue that should definitely be pursued.

3.5. Computational Models as Tools for Music Education

A final recent trend we would like to mention here is the increased interest in exploring computational models of expressive performance for educational purposes. Juslin (2003) already pointed out that insights learned by developing such models can help understand and appreciate the way musicians perform music expressively. Furthermore, initiatives like the RenCon competition have stressed from the beginning the importance of using computational models for educational purposes, stating in a tongue-in-cheek manner that RenCon's long-term goal is to have a human performer educated using a computational system win the first prize at the International Chopin Piano Competition by 2100 (Hiraga et al., 2002).

A possible use of computational models as tools for education is to analyze performance strategies from visualizations of expressive parameters (Langner and Goebl, 2003; Grachten et al., 2009; Chew, 2012, 2016) or comparing characteristics of a performance (Sapp, 2007, 2008; Liem and Hanjalic, 2015; Grachten et al., 2017). By highlighting similarities and variations in expressive patterns and qualities in performances and relating these to aspects of the written score, this kind of analyses might be interesting not only to music students, buy also to general audiences, stimulating listeners' engagement with, and understanding of, music. All of this could be built into *active music listening interfaces* (Goto, 2007), such as the integrated prototype of the PHENICX project[3] (Liem et al., 2015).

Computer accompaniment systems can help musicians to practice. First concrete examples are commercial applications such as Smartmusic[4], which commercializes Roger Dannenberg's research, Cadenza[5], based on work by Chris Raphael and Antescofo (Cont, 2008), which has been developed into commercial applications for providing adaptable backing tracks for musicians and music students[6]. Conductor and feedback systems can be also be used for educational purposes, either as a simulation of orchestra conducting for conducting students (Peng and Gerhard, 2009; Baba et al., 2010), or as interactive experiences for helping to introduce general audiences to classical music (Sarasúa et al., 2016).

Another dimension is the technical and mechanical aspects of instrument playing and practicing. Here, for example, algorithms that can determine the difficulty of a piece (Sébastien et al., 2012; Nakamura et al., 2014b) or propose appropriate fingering strategies (Al Kasimi et al., 2007; Nakamura et al., 2014b; Balliauw et al., 2015) would be useful. Furthermore, computational models might help determine a performer's skill level (Grindlay and Helmbold, 2006; Caramiaux et al., 2017). Musical e-learning platforms such as Yousician[7] and Music Prodigy[8] (and many more, as this is a rapidly growing business segment for start-ups) might benefit from models of performance to provide a more engaging experience, as well as to develop better musicianship.

4. A CRITICAL DISCUSSION OF PARAMETER SELECTION AND MODEL EVALUATION

The following section presents a discussion of how certain choices in the score features, expressive parameters and models affect what a computational performance model can describe. We focus on three main aspects, namely, the effects of the choice of expressive targets (section 4.1), the level at which a system models expressive performance, based on Palmer's categories described in section 2.2 above (section 4.2), and on the way models are evaluated (section 4.3).

[3]http://phenicx.com
[4]https://www.smartmusic.com
[5]http://www.sonacadenza.com
[6]https://www.antescofo.com
[7]https://yousician.com
[8]https://www.musicprodigy.com

4.1. Encoding Expressive Dimensions and Parameters

As explained in section 2.3 above, expressive parameters are numerical descriptors that capture certain aspects of a performance. As already discussed by De Poli (2004) (and this remains true today), there seems to be no consensus on the best way of describing a music performance. Instead, each formulation uses variants of these parameters, which has some consequences on the kinds of performances or performance aspects that can be modeled.

The most commonly modeled performance aspects (for the piano) are expressive tempo/timing, dynamics and articulation. To keep the discussion manageable, we will also restrict ourselves to these parameters here, leaving out other dimensions such as timbral parameters, vibrato, or intonation. A piano performance can be represented in the most simplistic way by the three MIDI parameters note onset, offset, and key velocity. Other instruments might involve other parameters such as bow velocity for string instruments (Marchini et al., 2014). Furthermore, in some instruments, like winds and strings, there might be a discussion whether to model perceptual or physical onsets (Vos and Rasch, 1981), or indeed whether the notion of a well-defined, exact onset time is meaningful.

4.1.1. Tempo and Timing

Expressive tempo and timing ultimately relate to the "temporal position" of musical events. Broadly speaking, *tempo* refers to the approximate rate at which musical events happen. This may refer to the *global tempo* of a performance (which is often roughly prescribed in the score by the metronome number), or to *local tempo*, which is the rate of events within a smaller time window and can be regarded as local deviations from the global tempo. *Expressive timing*, finally, refers to deviations of the individual events from the local tempo. Setting these three notions apart is of crucial importance in quantitative modeling of performance, computational, or otherwise.

There is support from the music psychology literature that timing patterns are tempo-dependent (Desain and Honing, 1994; Repp et al., 2002; Honing, 2005; Coorevits et al., 2015). Although there is no clear-cut definition of where local tempo variations end and expressive timing starts, the distinction between local tempo and timing was shown to be perceptually relevant in a study by Dixon et al. (2006) where listeners rated beat trains played along with expressive performances, and were shown to prefer slightly smoothed beat trains over beat trains that were exactly aligned to the note onsets. This reinforces the idea that note level irregularities should be not be regarded as as micro-fluctuations of local tempo, but rather as *deviations* from local tempo. A similar result was presented by Gu and Raphael (2012). Honing (2005, 2006) provides valuable insight into the limits of expressive timing by observing that very strong deviations from a steady beat may interfere with the rhythm that is perceived by the listener. Assuming that a goal of the performer is to make the listener accurately recognize the rhythmic categories of the score being played, this constrains the freedom of expressive timing. Honing (2006) then uses a model of human rhythm perception

to infer limits on expressive timing for phrase endings based on their rhythmic patterns.

Several computational models explicitly separate tempo and timing. Recent versions of the KTH model (Friberg et al., 2006 see **Table 1**) have rules dealing with tempo (e.g., phrasing rules) and timing (e.g., melodic sync, micro-level timing). In *Laminae* (Okumura et al., 2014), tempo is represented by the average BPM per beat, while timing is defined as the onset deviations relative to the beat. *Polyhymnia* (Kim et al., 2011) decomposes tempo into two expressive parameters, calculating tempo curves for the highest and lowest melodic lines. *YQX* (Flossmann et al., 2013) represents tempo as the lower frequency components of the log-IOI ratio series, and timing as the residual high frequency components. In a similar fashion, the most recent version of the *Basis Mixer* (Cancino Chacón and Grachten, 2016) computes expressive tempo from the smoothed log-IOI series, where the estimated IOIs come from a smoothed (spline) interpolation of the performed onsets, and timing as the deviations from these estimated IOIs. There are some practical issues with the use of smooth tempo targets, such as the problem of phrase boundaries, where tempo changes are not necessarily smooth. A solution involving adaptive smoothing (Dixon et al., 2006)—splines with manual knot placement at phrase boundaries—would require human annotation of the phrase structure. Dannenberg and Mohan (2011) describe an interesting dynamic programming optimization algorithm to find the best spline fit allowing a finite number of knots without manual annotations. Other approaches involve local linear approximations of the tempo (Xia, 2016) or multiple hierarchical decompositions (Widmer and Tobudic, 2002).

Another issue related to the modeling of tempo and timing is *scaling* of the expressive parameters, which determines whether we model relative tempo changes, or the actual tempo itself. Chew and Callender (2013) argue in favor of using log-tempo for analysis of performance strategies. Flossmann et al. (2013), Kim et al. (2013), and Grachten and Cancino-Chacón (2017) use logarithmic tempo parameters, while most works focus on linear parameters (Grindlay and Helmbold, 2006; Teramura et al., 2008; Gu and Raphael, 2012; Okumura et al., 2014; Gingras et al., 2016; Cancino-Chacón et al., 2017b; Peperkamp et al., 2017).

Some choose to focus on modeling the dynamic *change* in the parameters instead of the parameters themselves, by calculating *differences*. Gingras et al. (2016) model both IOIs and their first differences— also for a technical reason, since the IOI series is not stationary, and thus not suitable for linear time-series analysis. Okumura et al. (2014) focuson the changes in expressive tempo, by explicitly modeling the conditional probability distribution of the current expressive tempo given its previous difference, using Gaussian distributions. Grindlay and Helmbold (2006) jointly model expressive tempo and its first differences, which leads to more coherent predictions.

4.1.2. Articulation

Articulation, in the case of the piano, refers to the ratio between the performed duration of a note and its notated value and therefore also describes the amount of overlap between consecutive notes. Common articulation strategies include ***staccato*** (shortening compared to notated duration) and ***legato*** (smooth connection to following note). While most generative models deal with expressive tempo/timing, not all of them model articulation. As with tempo, there are several variants of quantitatively describing articulation, including the use of linear (Flossmann et al., 2013) or logarithmic scaling of the parameters (Kim et al., 2011; Cancino Chacón and Grachten, 2016).

To the best of our knowledge, no data-driven generative system has attempted to model *pedaling*, a subtle art that has complex consequences for note durations, but also for the overall sound of a passage. The effect of pedaling on articulation may still be modeled implicitly, by distinguishing between the events of a piano key release and the actual ending of the associated sound (when the sustain pedal is released), as is done in the Basis Function models, for example.

4.1.3. Expressive Dynamics

To simply relate performed dynamics to loudness would miss a number of important aspects of expressive performance. As discussed by Elowsson and Friberg (2017), there is a difference between mere loudness and perceived dynamics. For example, it has been noted that the timbral characteristics of instruments (and therefore, their spectra) change with the performed intensity. Liebman et al. (2012) choose not to focus on loudness since analysis of loudness might not be entirely reliable.

Most approaches for the piano use MIDI velocity as a proxy for loudness. However, it must be noted that the mapping of MIDI velocities to performed dynamics and perceived or measured loudness in piano is not standardized in any way—it may be non-linear, and change from instrument to instrument. Some systems simply use MIDI velocity as an expressive target for each note, while others—particularly those for polyphonic music—decompose the MIDI velocity into several parameters. Early versions of the Basis Function model (Grachten and Widmer, 2012; Cancino Chacón and Grachten, 2015), as well as the unsupervised approach by van Herwaarden et al. (2014) and the NAIST model (Teramura et al., 2008), are non-sequential models and thus predict MIDI velocity for each score note. Sequential models such as ESP (Grindlay and Helmbold, 2006), Laminae (Okumura et al., 2011), and Polyhymnia (Kim et al., 2011) decompose a piece of music into several melodic lines, either automatically (Polyhymnia) or manually (ESP, Laminae), and predict the MIDI velocity for each voice independently. The latest version of the Basis Function models decomposes a performance into a dynamic trend, either the average or the maximal MIDI velocity at each score position (Cancino Chacón and Grachten, 2016; Cancino-Chacón et al., 2017b), and a local parameter describing the deviations from the trend for each score note. The rationale for this decomposition is that it allows for modeling the temporal evolution of expressive dynamics, something that cannot easily be done in polyphonic music when dynamics is represented as an attribute of individual notes.

In the case of *audio*, the problem of choosing a metric for expressive dynamics is more complicated due to the large number of measures of loudness. A common trend is to use loudness measures that take into account human perception, such as the

EBU-R-128 measure defined for regulation of loudness in the broadcasting industry (Grachten et al., 2017), and smoothed loudness curves in sones (Kosta et al., 2016).

4.1.4. Joint Modeling of Parameters

Musicians' expressive manipulations of tempo, timing, dynamics, and articulation have been studied from a cognitive perspective, both individually and in combination, to determine how they shape listeners' perceptions of performed music. A number of studies have sought to identify interactions between pairs of expressive parameters like timing and dynamics (Tekman, 2002; Boltz, 2011), and timing and tempo (Desain and Honing, 1994; Repp et al., 2002; Coorevits et al., 2015, 2017). While the music psychology literature provides some indication of how listeners expect pairs of expressive parameters to relate in certain (simplistic) contexts, it remains unclear whether these relationships are upheld during normal music performance, when the underlying piece is complex and many expressive parameters must be manipulated in parallel.

The influential model of expressive tempo and dynamics by Todd (1992) states that both aspects are linearly coupled by default (unless the musical context demands a decoupling), and suggests that this coupling may be especially tight for romantic piano music. The model predicts arc-like dynamics and tempo shapes to express phrase structure. Grindlay and Helmbold (2006)'s HHMM-based ESP system allows for the joint modeling of expressive parameters, however the focus in their work is strongly on local tempo. No quantitative results are given for the modeling of tempo in combination with dynamics and articulation. The KTH model (Friberg et al., 2006) includes rules that prescribe the joint variation of multiple parameters, such as a *phrasing* rule that accounts for arc-like shapes in dynamics and tempo, similar to those in Todd (1992). Several other authors combine separate models for each expressive parameter, and do not consider interactions (Teramura et al., 2008; Widmer et al., 2009), or consider only a single expressive parameter (Kosta et al., 2016; Peperkamp et al., 2017). Recent versions of the Basis Function models (Cancino Chacón and Grachten, 2016) allow for joint estimation of parameters using Gaussian mixture density networks (GMNs); parameters defined for individual notes and parameters defined only per score time point are modeled in separate sets. Xia and Dannenberg (2015) and Xia et al. (2015) jointly model expressive dynamics and tempo using linear dynamical systems, with the underlying assumption that the joint distribution of the parameters is Gaussian. The approach presented by Moulieras and Pachet (2016) models dynamics and timing jointly with a joint probability distribution approximated using a maximum entropy approach. Since this approach is not Gaussian, the form of the distribution depends on the training data.

To the best of our knowledge, there has not been an extensive computational study analyzing whether the joint estimation of parameters improves the generative quality of predictive models. Furthermore, in some cases performers will manipulate two parameters in different ways during the course of a single piece to achieve different expressive goals (e.g., slowing down while simultaneously getting softer, then elsewhere slowing down while getting louder). Whether the consistent use of particular parameter relationships relates to the aesthetic quality of a performance, increases its predictability, or makes the communication of expression more successful likewise requires further study.

4.2. Relation to Palmer's Categories

4.2.1. Interpretation

As stated in section 2.2, expressive performance of notated music can be seen as a communication process in which information flows from the composer to the listener through the performer (Kendall and Carterette, 1990). In this case, the role of the performer involves semantically and affectively *interpreting* the score. Gingras et al. (2016) provide evidence supporting this relationship by linking information-theoretic features (related to the role the composer) to expressive timing (performer), which is a good predictor of perceived tension (listener).

An important aspect of the interpretation of a score is to highlight structural content. A common approach taken by many systems is to rely on input features describing *group boundaries* and *phrase structure*. Friberg et al. (2006) and Grindlay and Helmbold (2006) use features related to phrase structure, which is assumed to be manually annotated in the score. Giraldo S. and Ramírez (2016) and Giraldo S.I. and Ramirez (2016) use LBDM, an automatic segmentation algorithm based on Gestalt theory (Cambouropoulos, 1997).

Another important aspect in polyphonic Western music is the hierarchical relations and interactions between different *voices*, which in most cases involves distinguishing the main (or most salient) melody. Several models require the melody to be annotated (Grindlay and Helmbold, 2006; Okumura et al., 2014; Cancino-Chacón et al., 2017b). Other models simply assume that the main melody is composed of the highest notes (Teramura et al., 2008; Flossmann et al., 2013).

Another marker of music structure are the patterns of *tension* and *relaxation* in music, linked to several aspects of *expectedness*. Farbood (2012) showed a relationship between expressive timing and perceived tension. Grachten and Widmer (2012) use Narmour (1990) Implication–Realization model to link expressive dynamics to melodic expectation, but observe no substantial improvement over simpler models that use only pitch and dynamics annotations as predictors. Chew (2016) introduces the idea of tipping points, i.e., extreme cases of pulse elasticity, and their relation to tonality, in particular harmonic tension. The KTH model includes features describing harmonic tension (Friberg et al., 2006). Gingras et al. (2016) show relationship of expressive timing and perceived tension. Recent versions of the Basis Function models (Cancino Chacón and Grachten, 2016) include harmonic tension features computed using the methods proposed by Herremans and Chew (2016).

Beyond the identification of structural aspects, another important aspect of interpretation is to highlight particular *emotional content* of the music. Juslin (2003) points out that "*[a] function of performance expression might be to render the performance with a particular emotional expression*." Research in music and emotion is a very active field (see Juslin and Sloboda, 2011 for an overview), which includes studying

the relationship between intended emotion and performance gestures and strategies (Juslin, 2001; Gabrielsson and Lindström, 2010; Bresin and Friberg, 2011). Eerola et al. (2013) study the contribution of expressive dimensions such as tempo, dynamics, articulation, register, and timbre to determining emotional expression. Their results suggest that expressive dimensions interact linearly, and their contributions seem to be additive. While some generative models allow the user to control the intended emotion or expressive character (Bresin and Friberg, 2000; Friberg, 2006; Canazza et al., 2015), to the best of our knowledge no autonomous generative model attempts to recognize emotive content of a piece directly from analysis of the score, and render it appropriately.

4.2.2. Planning

While interpretation of a musical score aims at uncovering its semantic and affective content, performance planning refers to how this content, along with more or less specific artistic or expressive intentions of the performer, is turned into specific expressive performance decisions. In this view, most computational models of expressive performance act at this level, since they focus on explicitly (i.e., quantitatively) relating structural aspects of the score to parameters encoding an expressive performance.

An important characteristic of Western classical music is the hierarchical nature of its structure. Repp (1998) points out that "*[t]he performer's (often subconscious) intention seems to 'act out' the music's hierarchical grouping structure and thereby communicate to the listeners*." It is therefore, important to determine how the different hierarchical levels interact with each other and contribute to the overall expression. The relation between the hierarchical structure and expression has been explored in the cognitive literature (Clarke, 1993; Repp, 1998; Toiviainen et al., 2010). Widmer and Tobudic (2002) explore the relationship between hierarchical levels of the phrase structure and expressive tempo, using a multilevel decomposition of the tempo curves corresponding to each level of the phrase structure, and an inductive rule learning method to model the note-wise performance residuals. Tobudic and Widmer (2006) expand on this work using an instance-based learning method in which the hierarchical phrase structure is represented using first-order logic.

An important design issue relating to the structure–expression relationships is how the choice of score (feature) representation affects the possible performance gestures that can be modeled (i.e., planned). An example of this would be whether the possible patterns of dynamics and timing deviations that a system can describe are "implicitly" assumed from the encoding of features—as might be the case with systems using features describing metrical strength and metrical hierarchy (Grindlay and Helmbold, 2006; Teramura et al., 2008; Kim et al., 2011; Marchini et al., 2014; Giraldo S. and Ramírez, 2016)—or can be inferred directly from human performances using more agnostic features denoting metrical position (Xia et al., 2015; Cancino-Chacón C.E. et al., 2017).

4.2.3. Movement

Humans need to transform the result of the interpretation and planning stages into an actual acoustic rendering of a piece by means of movements of their bodies (i.e., actually playing the instrument). In this regard, we can consider movement and embodiment as necessary conditions for (human) expressive performance. Similar to the concept of embodied cognition (Leman et al., 2017a), neuroscientific accounts refer to the "action–perception loop," a well-trained neural connection between the aim of an action, here the musical sound, and its execution, the necessary body movements at the musical instrument (Novembre and Keller, 2014). Musicians, having practiced over decades, will "hear" or imagine a certain sound and execute the appropriate body movements automatically. Likewise, co-musicians or the audience will perceive a performance through hearing and seeing the performer (Platz and Kopiez, 2012); and even from only hearing the sound, experienced listeners will be able to deduce bodily states and movement characteristics of the performer. Leman et al. (2017b) discuss the role of the hand as a co-articulated organ of the brain's action–perception machinery in expressive performance, music listening and learning.

Body motion is an essential means of non-verbal communication not only to the audience, but also among musicians. Goebl and Palmer (2009) showed in ensemble performances of simple melodies that visual information became more important to stay in synchrony (i.e., musicians' head movements were more synchronized) as auditory cues were reduced. Body movements serve specific roles at certain places in a piece (e.g., at the beginning, after fermatas). Bishop and Goebl (2017, 2018) study specific head motion kinematics in ensemble performance used to cue-in a piece without upbeat. They found characteristic patters including acceleration peaks to carry relevant cueing information.

In spite of the progress in music psychology and embodied cognition, few computational approaches take into account aspects of motion while modeling expressive performance. However, the availability of motion capture technology as well as new trends in psychological research might open the field of modeling expressive movement. The KTH model includes performance noise as a white noise component relating to motor delay and uses 1/f noise to simulate noise coming from an internal time-keeper clock (Friberg et al., 2006). Dalla Bella and Palmer (2011) show that finger velocity and acceleration can be used as features to identify individual pianists. Marchini et al. (2013, 2014) study expressive performance in string quartets using a combination of music-only related expressive parameters, as well as bow velocity, a dimension of movement directly related to performed dynamics. Caramiaux et al. (2017) assess whether individuality can be trained, that is whether the differences in performance style are related to development in skill and can thus be learned. Their results suggest that motion features are better than musical timing features for discriminating performance styles. Furthermore, the results suggest that motion features are better for classification.

4.3. Evaluating Computational Performance Models

How the quality or adequacy of computational performance models can be evaluated in a systematic and reliable fashion is a difficult question. First of all, the evaluation will depend on the purpose of the model. A model designed to serve an explanatory purpose should be evaluated according to different criteria than a model for performance generation. In the former case, the simplicity of the model structure may be of prime importance, as well as how easily the model output can be linked to aspects of the input. In the latter, we may be more interested in how convincing the generated performance sounds than how easy it is to understand the decisions of the model.

Furthermore, when we evaluate a model by the quality of its output, an important issue is the ultimately subjective nature of judging the musical quality of an expressive performance. And while we might even be able to formulate principles to which a good performance should adhere, it is entirely conceivable that a performance conforming to all these principles fails to please us, or conversely, that a performance defying these principles is nevertheless captivating.

Bresin and Friberg (2013) formulate several more formal aspects of computational performance models that can be evaluated, including their ability to reproduce/reconstruct (specific) human performances and their capacity to adapt to different expressive intentions/contexts.

4.3.1. Attempts at Quantitative, "Objective" Evaluation

Most of the work described above relies on quantitative evaluation in terms of predictive capabilities and/or goodness of fit, relative to a given set of human performances. These measures tend to focus on the prediction or reconstruction error—e.g., in the form of the correlation or the mean squared error (MSE) between the performance patterns predicted by a model, and a real human performance—, or on a so-called likelihood function (which gives the probability of observing a given (human) performance, given a particular model). What all these approaches have in common is that they base their evaluation on a comparison between a model's output, and a —usually one specific—performance by a human musician (most often additional performances by the same musician(s) from whom the model was learned). This is problematic for several reasons:

- Comparison to a single "target" performance is highly arbitrary, given that there are many valid ways to perform a piece. A good fit may at least indicate that a model has the capacity of encoding and describing the specific performances by a specific performer (with, presumably, a specific style). A poor fit does not necessarily mean that the model's predictions are musically bad.
- What is more, there is no guarantee that higher correlation, or lower MSE implies a musically better performance, nor indeed that a performance that sounds more similar to the target. Especially *outliers* (single errors of great magnitude) can influence these measures. Errors may not be equally salient for all data points. Accounting for this would require a model of perceived saliency of musical positions and errors, which is currently out of reach (or has not been tackled yet).
- A more technical point is that we cannot compare performance models that encode an expressive dimension using different parameters (such as modeling expressive tempo using IBI vs. BPM, or using linear vs. logarithmic parameters), because quantitative correlation or error measures must assume a particular encoding. There are currently no canonical definitions of the expressive dimensions.

These kinds of problems have also been faced in other domains, and have been addressed in the computer vision literature with the introduction of the Structural Similarity Index (Wang et al., 2004), a perception-based metric that considers perceptual phenomena like luminance masking, as well as perceived change in structural information. However, to the best of our knowledge, there has not been any attempt to define similar measures for music, or to propose a custom measure for expressive performance.

Bresin and Friberg (2013) suggest to relate these error metrics to more general perceptual models. An example of this would be reporting the error in terms of just noticeable differences (JNDs) in the expressive parameters. Nevertheless, it is worth noticing that JNDs are highly dependent on the musical context.

4.3.2. Qualitative Evaluation via Listening Tests

The obvious alternative to quantitative, correlation-based evaluation is evaluation by listening: playing human and computer-generated performances to human listeners and asking them to rate various qualities, or to simply rank them according to some musical criteria.

An early initiative that attempted to provide a systematic basis for this was *RenCon (Performance Rendering Contest)*[9], a Japanese initiative that has been organizing a series of contests for computer systems that generate expressive performances (Hiraga et al., 2002, 2003, 2004, 2006; Katayose et al., 2012). At these RenCon Workshops, (piano) performances of different computational models were played to an audience, which then voted for a "winner" (the most "musical" performance). It is currently not clear if this initiative is being continued. (Actually, the last RenCon workshop we are aware of dates back to 2013.)

Also, there are a number of issues with audience-based evaluation, which clearly surfaced also in the RenCon Workshops: the appropriate choice of music; the listeners' limited attention span; the difficulties inherent in comparing different kinds of systems (e.g., fully autonomous vs. interactive), or systems that model different performance parameters (e.g., not all models address the articulation dimension, or autonomously decide on the overall tempo to be chosen). Finally, reproducibility of results is an issue in audience-based evaluation. There is no guarantee that repeating a listening test (even with the same audience) will yield the same results, and it is impossible to compare later models to models that have been evaluated earlier.

[9] www.renconmusic.org

A particular and very subtle problem is the choice, and communication to the human subjects, of the *rating criteria* that should be applied. This can also be exemplified with a recent *"Turing Test"* described in Schubert et al. (2017), where piano performances produced by several systems that had won recent RenCon competitions, along with one performance by a real human pianist, were rated by a human listening panel. The subjects were asked to rate the performances (all rendered on the same piano) according to different dimensions. The question that the analysis in Schubert et al. (2017) then mainly focuses was to what degree the listeners believed that *"[t]he performance was played by a human"*. Without going into the details of the results[10], it is clear that such a question may be interpreted differently by different listeners, or indeed depending on what apparent weaknesses are heard in a performance: a performance with extreme timing irregularities might be (erroneously) classified as "human" because the listener might believe that it was produced by a poor piano student or a child, or that it could not be the output of a computer, because computers would be able to play with perfect regularity. Generally, an inherent problem with qualitative listening evaluations is that one single cue (e.g., "strange," unusual mistake) can give away a given trial as probably computer-generated, independent of how convincing the rest was.

There is plenty of evidence in the music psychology literature showing that the assessment of the quality of a performance depends not only on the quality of its acoustic rendering, but on a number of other factors. Platz and Kopiez (2012) present a meta-analysis of 15 studies from 1985 to 2011 supporting the hypothesis that audio-visual presentation enhances appreciation of music performance. Their results show that the visual component is an important factor in the communication of meaning. Tsay (2013) present controversial results suggesting that visual information alone might be sufficient when determining the winner of a music competition. Wapnick et al. (2009) suggest that certain non-musical attributes like the perceived attractiveness of a performer or the way they behave on stage affects ratings of high-level piano performances, particularly on short performances. Thompson et al. (2007) study the evolution of listeners' assessments of the quality of a performance over the course of the piece. Their results suggest that even while listeners only need a short time to reach a decision on their judgment, there is a significant difference between the initial and final judgments. Wesolowski et al. (2016) present a more critical view of listeners' judgments by examining the precision.

De Poli et al. (2014) and Schubert et al. (2014a) specifically study how the audience judges entire performances of computational models, by analyzing listeners' scores of several aspects including technical accuracy, emotional content and coherence of the performed style. The listeners were categorized into two different cognitive styles: music systemizers (those who judge a performance in technical and formal terms) and music empathizers (describe a performance in terms of its emotive content). Their results suggest that preference for different performances cannot be attributed to these cognitive styles, but the cognitive style does influence the justification for a rating. Schubert et al. (2014b) suggest that the conceptual difference between music empathizers and music systemizers might not be sufficient to capture significant differences in evaluating music performances.

Despite all these problematic aspects, some way of qualitative, expert- or listener-based evaluation of computational performance models seems indispensable, as the quantitative measures described in the previous section definitely fall short of capturing the *musical* (not to mention the emotional) quality of the results. This is a highly challenging open problem for the performance research community—and an essential one.

5. CONCLUSIONS

This work has reviewed some recent developments on the study and generation of expressive musical performances through computational models. Perhaps the most notable trends are a strong focus on data-driven methods for analysis and generation, which mirrors the trend in other areas such as natural language processing and computer vision; and increased interest in interactive systems, which allow us to explore musical human–computer interactions.

In their current state, computational models of performance provide support for a number of musically and cognitively plausible hypotheses, such as the existence of certain patterns in performance, the importance of attending to the local context in the score (Marchini et al., 2014; Kosta et al., 2016; Grachten and Cancino-Chacón, 2017), and Kendall and Carterette (1990)'s communication model for the role of composer, performer and listener (Gingras et al., 2016). Nevertheless, most approaches focus on mapping local syntactic structures to performance gestures, but are not able to model the longer hierarchical relationships that might be relevant for a full understanding of the music, and its dramatic structure.

There remains much to be done in order to advance the state of the art, and to improve the utility of such computational models—both as vehicles for exploring this complex art, and as practical tools. We would like to end this article by briefly highlighting four aspects (out of many) to consider for further research in the immediate future:

1. **Dataset creation**. The power of data-driven methods comes at the cost of requiring large amounts of data, in this case specifically, performances aligned to their scores. As pointed out by Juslin (2003), this might be an issue preventing the advance in computational models of music expression. As discussed in section 3.4, currently available datasets do not yet reach the size (in terms of amount of data and variety) that has been able to boost other domains such as computer vision. Still, progress is being made, with initiatives like those

[10]Briefly: it turned out that on this "perceived humanness" rating scale, several computational models scored at a level that was statistically indistinguishable from the human pianist, with the linear Basis Function model (Grachten and Widmer, 2012) achieving the highest "humanness" ratings (higher than even the human performance).

by the CrestMuse group in Japan (Hashida et al., 2017). We would like to encourage the community at large to focus on developing more datasets, in a joint effort.

2. **Expressive parameters.** As discussed in section 4.1, there is no consensus regarding the encoding of expressive dimensions. Efforts should be made to investigate the effects of the choice of performance parameters encoding, as well as joint estimation of parameters. An interesting direction would be to search for representations that are cognitively plausible (in terms of human real-time perception and memory).
3. **Models of music understanding and embodiment.** As pointed out by Widmer (2017), it is necessary to develop models and features that better capture the long term semantic and emotive relationships that appear in music. This might require to develop better features, including learned features, as well as reframing the computational tasks in terms of approaches like reinforcement learning. Furthermore, more research efforts into developing computational models that include aspects of embodied music interaction might be required.
4. **Evaluation** Having well-established and valid criteria for evaluating different models, and comparing their performance to that of humans, is essential to making progress. In terms of quantitative measures, more work will be required to conduct research that studies the effects and biases involved in the choice of evaluation metrics. Furthermore, it would be interesting to evaluate computational models of expression as models of cognition, not only focusing on how well they reproduce the observed data, but also if the predictions of the model are cognitively plausible (Honing, 2006). Ideally, quantitative measures should relate to perceptually relevant aspects of performances, as perceived by musical listeners. In terms of qualitative, truly musical evaluation, which we consider indispensable, we need more efforts toward establishing venues for systematic evaluation and comparison, like the RenCon workshop and similar initiatives. And again, studies that give us a better understanding of how humans evaluate performances, would be extremely useful.

While at this stage (and perhaps forever) it is more than uncertain whether computational models of performance will ever successfully beat humans in high-profile competitions, as stated as a goal by the RenCon initiative (Hiraga et al., 2002), there is no doubt that understanding the way humans create and enjoy expressive performances is of great value. It is our hope that the field of research we have attempted to portray here can contribute to such an understanding, and develop useful musical tools along the way.

AUTHOR CONTRIBUTIONS

This work was developed as part of the Ph.D. research of CC-C, and under the supervision of GW. All authors contributed to the conception and structure of the manuscript. CC-C conducted the literature research and wrote the first draft of the manuscript. MG, WG, and GW reviewed the draft and wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

Funding for this work was provided by the European Research Council (ERC) under the EU's Horizon 2020 Framework Programme (ERC Grant Agreement number 670035, project Con Espressione) and by the Austrian Research Fund FWF under project number P29427.

REFERENCES

Al Kasimi, A., Nichols, E., and Raphael, C. (2007). "A simple algorithm for automatic generation of polyphonic piano fingerings," in *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR 2007)* (Vienna), 355–356.

Baba, T., Hashida, M., and Katayose, H. (2010). ""VirtualPhilharmony": a conducting system with heuristics of conducting an orchestra," in *Proceedings of the 10th International Conference on New Interfaces for Musical Expression, NIME 2010* (Sydney, NSW), 263–270.

Balliauw, M., Herremans, D., Palhazi Cuervo, D., and Sörensen, K. (2015). "Generating fingerings for polyphonic piano music with a Tabu search algorithm," in *Proceedings of the 5th International Conference on Mathematics and Computation in Music (MCM 2015)* (London), 149–160.

Bantula, H., Giraldo, S., and Ramírez, R. (2016). "Jazz ensemble expressive performance modeling," in *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)* (New York, NY), 674–680.

Bella, S. D., and Palmer, C. (2011). Rate effects on timing, key velocity, and finger kinematics in piano performance. *PLoS ONE* 6:e20518. doi: 10.1371/journal.pone.0020518

Bergeron, V., and Lopes, D. M. (2009). Hearing and seeing musical expression. *Philos. Phenomenol. Res.* 78, 1–16. doi: 10.1111/j.1933-1592.2008.00230.x

Binet, A., and Courtier, J. (1896). Recherches graphiques sur la musique. *L'année Psychol.* 2, 201–222.

Bisesi, E., Parncutt, R., and Friberg, A. (2011). "An accent-based approach to performance rendering: music theory meets music psychology," in *Proceedings of the International Symposium on Performance Science 2011 (ISPS 2011)* (Toronto, ON), 27–32.

Bishop, L., and Goebl, W. (2017). Communication for coordination: gesture kinematics and conventionality affect synchronization success in piano duos. *Psychol. Res.* 82, 1177–1194. doi: 10.1007/s00426-017-0893-3

Bishop, L., and Goebl, W. (2018). Beating time: how ensemble musicians' cueing gestures communicate beat position and tempo. *Psychol. Music* 46, 84–106. doi: 10.1177/0305735617702971

Boltz, M. G. (2011). Illusory tempo changes due to musical characteristics. *Music Percept.* 28, 367–386. doi: 10.1525/mp.2011.28.4.367

Bresin, R. (1998). Artificial neural networks based models for automatic performance of musical scores. *J. New Music Res.* 27, 239–270. doi: 10.1080/09298219808570748

Bresin, R., and Friberg, A. (2000). Emotional coloring of computer-controlled music performances. *Comput. Music J.* 24, 44–63. doi: 10.1162/014892600559515

Bresin, R., and Friberg, A. (2011). Emotion rendering in music: range and characteristic values of seven musical variables. *Cortex* 47, 1068–1081. doi: 10.1016/j.cortex.2011.05.009

Bresin, R., and Friberg, A. (2013). "Evaluation of computer systems for expressive music performance," in *Guide to Computing for Expressive Music Performance*, eds A. Kirke and E. R. Miranda (London: Springer-Verlag), 181–203.

Cambouropoulos, E. (1997). "Musical rhythm: a formal model for determining local boundaries, accents and metre in a melodic surface," in *Music, Gestalt and Computing*, ed M. Leman (Berlin; Heidelberg: Springer Berlin Heidelberg), 277–293.

Canazza, S., De Poli, G., and Rodà, A. (2015). CaRo 2.0: an interactive system for expressive music rendering. *Adv. Hum. Comput. Interact.* 2015, 1–13. doi: 10.1155/2015/850474

Cancino Chacón, C. E., and Grachten, M. (2015). "An evaluation of score descriptors combined with non-linear models of expressive dynamics in music," in *Proceedings of the 18th International Conference on Discovery Science (DS 2015)* (Banff, AB), 48–62.

Cancino Chacón, C. E., and Grachten, M. (2016). "The basis mixer: a computational romantic pianist," in *Late Breaking/ Demo, 17th International Society for Music Information Retrieval Conference (ISMIR 2016)* (New York, NY).

Cancino-Chacón, C., Bonev, M., Durand, A., Grachten, M., Arzt, A., Bishop, L., Goebl, W., et al. (2017a). "The ACCompanion v0.1: an expressive accompaniment system," in *Late Breaking/ Demo, 18th International Society for Music Information Retrieval Conference (ISMIR 2017)* (Suzhou).

Cancino-Chacón, C., and Grachten, M. (2018). "A computational study of the role of tonal tension in expressive piano performance," in *Proceedings of the 15th International Conference on Music Perception and Cognition (ICMPC15 ESCOM10)* (Graz).

Cancino-Chacón, C., Grachten, M., Sears, D. R. W., and Widmer, G. (2017b). "What were you expecting? Using expectancy features to predict expressive performances of classical piano music," in *Proceedings of the 10th International Workshop on Machine Learning and Music (MML 2017)* (Barcelona).

Cancino-Chacón, C. E., Gadermaier, T., Widmer, G., and Grachten, M. (2017). An evaluation of linear and non-linear models of expressive dynamics in classical piano and symphonic music. *Mach. Learn.* 106, 887–909. doi: 10.1007/s10994-017-5631-y

Caramiaux, B., Bevilacqua, F., Palmer, C., and Wanderley, M. (2017). "Individuality in piano performance depends on skill learning," in *Proceedings of the 4th International Conference on Movement Computing (MOCO'17)* (London: ACM).

Cheng, E., and Chew, E. (2008). Quantitative analysis of phrasing strategies in expressive performance: computational methods and analysis of performances of unaccompanied bach for solo violin. *J. New Mus. Res.* 37, 325–338. doi: 10.1080/09298210802711660

Chew, E. (2012). About time: strategies of performance revealed in graphs. *Vis. Res. Mus. Educ.* 20. Available online at: http://www-usr.rider.edu/%7Evrme/v20n1/index.htm

Chew, E. (2016). Playing with the edge: tipping points and the role of tonality. *Mus. Percept.* 33, 344–366. doi: 10.1525/mp.2016.33.3.344

Chew, E., and Callender, C. (2013). "Conceptual and experiential representations of tempo: effects on expressive performance comparisons," in *Proceedings of the 4th International Conference on Mathematics and Computation in Music (MCM 2013)* (Montreal, QC), 76–87.

Chew, E., François, A., Liu, J., and Yang, A. (2005). "ESP: a driving interface for expression synthesis," in *Proceedings of the 2005 Conference on New Interfaces for Musical Expression, NIME 2005* (Vancouver, BC), 224–227.

Chew, E., Liu, J., and François, A. R. J. (2006). "ESP: roadmaps as constructed interpretations and guides to expressive performance," in *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia* (New York, NY: ACM), 137–145.

Chew, E., and McPherson, A. (2017). "PERFORMING MUSIC: humans, computers and electronics," in *The Routledge Companion to Music Cognition*, eds R. Ashley and R. Timmers (New York, NY: Routledge), 301–312.

Chuan, C.-H., and Chew, E. (2007). "A dynamic programming approach to the extraction of phrase boundaries from tempo variations in expressive performances," in *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR 2007)* (Vienna), 305–308.

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

Clynes, M. (1969). *Toward a Theory of Man: Precision of Essentic Form in Living Communication.* Berlin; Heidelberg: Springer Berlin Heidelberg.

Clynes, M. (1986). Generative principles of musical thought integration of microstructure with structure. *J. Integr. Study Artif. Intell. Cogn. Sci. Appl. Epistemol.* 3, 185–223.

Clynes, M. (1987). "What can a musician learn about music performance from newly discovered microstruture principles (pm or pas)?," in *Action and Perception in Rhythm and Music, Vol. 55*, ed A. Gabrielsson (Stockholm: Royal Swedish Academy of Music), 201–233.

Clynes, M. (2005). *Automatic Expressive Intonation Tuning System*. U.S. Patent 6,924,426B2.

Cont, A. (2008). "Antescofo: anticipatory synchronization and Control of interactive parameters in computer music," in *Proceedings of the 2008 International Computer Music Conference (ICMC 2008)* (Belfast), 33–40.

Cont, A., Echeveste, J., Giavitto, J.-L., and Jacquemard, F. (2012). "Correct automatic accompaniment despite machine listening or human errors in antescofo," in *Proceedings of the 38th International Computer Music Conference (ICMC 2012)* (Ljubljana).

Coorevits, E., Moelants, D., Maes, P.-J., and Leman, M. (2015). "The influence of tempo on expressive timing: A multimodal approach," in *Proceedings of the Ninth Triennal Conference of the European Society for the Cognitive Sciences of Music (ESCOM 2015)* (Manchester), 17–22.

Coorevits, E., Moelants, D., Maes, P.-J., and Leman, M. (2017). Exploring the effect of tempo changes on violinists' body movements. *Music. Sci.* 1–24. doi: 10.1177/1029864917714609

Dannenberg, R. B. (1984). "An on-line algorithm for real-time accompaniment," in *Proceedings of the 1984 International Computer Music Conference* (Paris), 193–198.

Dannenberg, R. B., Gold, N. E., Liang, D., and Xia, G. (2014). Methods and prospects for human–computer performance of popular music. *Comput. Music J.* 38, 36–50. doi: 10.1162/COMJ_a_00238

Dannenberg, R. B., and Mohan, S. (2011). "Characterizing tempo change in musical performances," in *Proceedings of the International Computer Music Conference (ICMC 2011)* (Huddersfield, UK), 650–656.

Davies, S. (1994). *Musical Meaning and Expression*. Ithaca, NY: Cornell University Press.

Davies, S. (2001). "Philosophical perspectives on music's expressiveness," in *Music and Emotion: Theory and Research*, eds P. N. Juslin and J. A. Sloboda (Oxford, UK: Oxford University Press), 23–44.

De Poli, G. (2004). Methodologies for expressiveness modelling of and for music performance. *J. New Music Res.* 33, 189–202. doi: 10.1080/0929821042000317796

De Poli, G., Canazza, S., Rodà, A., and Schubert, E. (2014). The role of individual difference in judging expressiveness of computer-assisted music performances by experts. *ACM Trans. Appl. Percept.* 11, 1–20. doi: 10.1145/2668124

Desain, P., and Honing, H. (1994). Does expressive timing in music performance scale proportionally with tempo? *Psychol. Res.* 56, 285–292. doi: 10.1007/BF00419658

Di Carlo, D., and Rodà, A. (2014). "Automatic music "listening" for automatic music performance: a grandpiano dynamics classifier," in *Proceedings of the 1st International Workshop on Computer and Robotic Systems for Automatic Music Performance (SAMP 14)* (Venice), 1–8.

Dixon, S., Goebl, W., and Cambouropoulos, E. (2006). Perceptual smoothness of tempo in expressively performed music. *Music Percept.* 23, 195–214. doi: 10.1525/mp.2006.23.3.195

Dixon, S., Goebl, W., and Widmer, G. (2005). "The "Air Worm": an interface for real-time manipulation of expressive music performance," in *Proceedings of the 2005 International Computer Music Conference (ICMC 2005)* (Barcelona).

Eerola, T., Friberg, A., and Bresin, R. (2013). Emotional expression in music: contribution, linearity, and additivity of primary musical cues. *Front. Psychol.* 4:487. doi: 10.3389/fpsyg.2013.00487

Elowsson, A., and Friberg, A. (2017). Predicting the perception of performed dynamics in music audio with ensemble learning. *J. Acoust. Soc. Am.* 141, 2224–2242. doi: 10.1121/1.4978245

Fabiani, M. (2011). *Interactive Computer-Aided Expressive Music Performance: Analysis, Control, Modification and Synthesis*. Ph.D. thesis, KTH Royal Institute of Technology.

Fabiani, M., Friberg, A., and Bresin, R. (2013). "Systems for interactive control of computer generated music performance," in *Guide to Computing for Expressive*

Music Performance, eds A. Kirke and E. R. Miranda (London: Springer-Verlag), 49–73.

Farbood, M. M. (2012). A parametric, temporal model of musical tension. *Music Percept.* 29, 387–428. doi: 10.1525/mp.2012.29.4.387

Flossmann, S., Goebl, W., Grachten, M., Niedermayer, B., and Widmer, G. (2010). The magaloff project: an interim report. *J. New Music Res.* 39, 363–377. doi: 10.1080/09298215.2010.523469

Flossmann, S., Grachten, M., and Widmer, G. (2011). "Expressive performance with Bayesian networks and linear basis models," in *Rencon Workshop Musical Performance Rendering Competition for Computer Systems (SMC-Rencon)* (Padova).

Flossmann, S., Grachten, M., and Widmer, G. (2013). "Expressive performance rendering with probabilistic models," in *Guide to Computing for Expressive Music Performance*, eds A. Kirke and E. R. Miranda (London: Springer), 75–98.

Friberg, A. (2005). "Home conducting-control the Overall Musical expression with gestures," in *Proceedings of the 2005 International Computer Music Conference (ICMC 2005)* (Barcelona).

Friberg, A. (2006). pDM: an expressive sequencer with real-time control of the KTH music-performance rules. *Comput. Music J.* 30, 37–48. doi: 10.1162/comj.2006.30.1.37

Friberg, A., and Bisesi, E. (2014). "Using computational models of music performance to model stylistic variations," in *Expressiveness in Music Performance: Empirical Approaches Across Styles and Cultures*, eds D. Fabian, R. Timmers, and E. Schubert (Oxford, UK: Oxford University Press), 240–259. doi: 10.1093/acprof:oso/9780199659647.003.0014

Friberg, A., Bresin, R., and Sundberg, J. (2006). Overview of the KTH rule system for musical performance. *Adv. Cogn. Psychol.* 2, 145–161. doi: 10.2478/v10053-008-0052-x

Friberg, A., Colombo, V., Frydén, L., and Sundberg, J. (2000). Generating musical performances with director musices. *Comput. Music J.* 24, 23–29. doi: 10.1162/014892600559407

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

Fu, M., Xia, G., Dannenberg, R., and Wasserman, L. (2015). "A statistical view on the expressive timing of piano rolled chords," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)* (Malaga), 578–584.

Gabrielsson, A. (1974). Performance of rhythm patterns. *Scand. J. Psychol.* 15, 63–72.

Gabrielsson, A. (1999). "The performance of music," in *The Psychology of Music*, ed D. Deutsch (San Diego, CA: Academic Press), 501–602.

Gabrielsson, A. (2003). Music performance research at the millennium. *Psychol. Music* 31, 221–272. doi: 10.1177/03057356030313002

Gabrielsson, A., and Lindström, E. (2010). "The role of structure in the musical expression of emotions," in *Handbook of Music and Emotion: Theory, Research, Applications* (Oxford University Press), 367–400.

Gingras, B., Pearce, M. T., Goodchild, M., Dean, R. T., Wiggins, G., and McAdams, S. (2016). Linking melodic expectation to expressive performance timing and perceived musical tension. *J. Exp. Psychol. Hum. Percept. Perform.* 42, 594–609. doi: 10.1037/xhp0000141

Giraldo, S., and Ramírez, R. (2016). A machine learning approach to ornamentation modeling and synthesis in jazz guitar. *J. Math. Mus.* 10, 107–126. doi: 10.1080/17459737.2016.1207814

Giraldo, S. I., and Ramirez, R. (2016). A machine learning approach to discover rules for expressive performance actions in Jazz guitar music. *Front. Psychol.* 7:1965. doi: 10.3389/fpsyg.2016.01965

Goebl, W. (1999). *The Vienna 4x22 Piano Corpus*. doi: 10.21939/4X22

Goebl, W., Dixon, S., De Poli, G., Friberg, A., Bresin, R., and Widmer, G. (2008). "'Sense' in expressive music performance: data acquisition, computational studies, and models," in *Sound to Sense – Sense to Sound: A State of the Art in Sound and Music Computing*, eds P. Polotti and D. Rocchesso (Berlin: Logos), 195–242.

Goebl, W., and Palmer, C. (2009). Synchronization of timing and motion among performing musicians. *Music Percept.* 26, 427–438. doi: 10.1525/mp.2009.26.5.427

Goebl, W., and Widmer, G. (2009). "On the use of computational methods for expressive music performance," in *Modern Methods for Musicology: Prospects, Proposals, and Realities*, eds T. Crawford, and L. Gibson (London: Ashgate), 93–113.

Goodchild, M., Gingras, B., and McAdams, S. (2016). Analysis, performance, and tension perception of an unmeasured prelude for harpsichord. *Music Percept.* 34, 1–20. doi: 10.1525/mp.2016.34.1.1

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Goto, M. (2007). Active music listening interfaces based on signal processing. in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007)* (Honolulu, HI), 1441–1444.

Grachten, M., and Cancino-Chacón, C. E. (2017). "Temporal dependencies in the expressive timing of classical piano performances," in *The Routledge Companion to Embodied Music Interaction*, eds M. Lessafre, P. J. Maes, and M. Leman (New York, NY: Routledge), 360–369.

Grachten, M., Cancino-Chacón, C. E., Gadermaier, T., and Widmer, G. (2017). Towards computer-assisted understanding of dynamics in symphonic music. *IEEE Multimedia* 24, 36–46. doi: 10.1109/MMUL.2017.4

Grachten, M., Goebl, W., Flossmann, S., and Widmer, G. (2009). Phase-plane representation and visualization of gestural structure in expressive timing. *J. New Music Res.* 38, 183–195. doi: 10.1080/09298210903171160

Grachten, M., and Krebs, F. (2014). An assessment of learned score features for modeling expressive dynamics in music. *IEEE Trans. Multimedia* 16, 1211–1218. doi: 10.1109/TMM.2014.2311013

Grachten, M., and Widmer, G. (2009). "Who is who in the end? Recognizing pianists by their final ritardandi," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)* (Kobe), 51–56.

Grachten, M., and Widmer, G. (2012). Linear basis models for prediction and analysis of musical expression. *J. New Music Res.* 41, 311–322. doi: 10.1080/09298215.2012.731071

Grindlay, G., and Helmbold, D. (2006). Modeling, analyzing, and synthesizing expressive piano performance with graphical models. *Mach. Learn.* 65, 361–387. doi: 10.1007/s10994-006-8751-3

Gu, Y., and Raphael, C. (2012). "Modeling piano interpretation using switching kalman filter," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)* (Porto), 145–150.

Hashida, M., Matsui, T., and Katayose, H. (2008). "A new music database describing deviation information of performance expression," in *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR 2008)* (Philadelphia, PA), 489–495.

Hashida, M., Nakamura, E., and Katayose, H. (2017). "Constructing PEDB 2nd Edition: a music performance database with phrase information," in *Proceedings of the 14th Sound and Music Computing Conference (SMC 2017)* (Espoo), 359–364.

Herremans, D., and Chew, E. (2016). "Tension ribbons: quantifying and visualising tonal tension," in *Proceedings of the Second International Conference on Technologies for Music Notation and Representation TENOR* (Cambridge, UK).

Herremans, D., Chuan, C.-H., and Chew, E. (2017). A functional taxonomy of music generation systems. *ACM Comput. Surveys* 50, 1–30. doi: 10.1145/3108242

Hiraga, R., Bresin, R., Hirata, K., and Katayose, H. (2003). "After the first year of Rencon," in *Proceedings of the 2003 International Computer Music Conference, ICMC 2003, September 29 - October 4, 2003* (Singapore).

Hiraga, R., Bresin, R., Hirata, K., and Katayose, H. (2004). "Rencon 2004: turing test for musical expression," in *Proceedings of the 2004 International Conference on New Interfaces for Musical Expression (NIME-04)* (Hamamatsu), 120–123.

Hiraga, R., Bresin, R., and Katayose, H. (2006). "Rencon 2005," in *Proceedings of the 20th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI2006)* (Tokyo).

Hiraga, R., Hashida, M., Hirata, K., Katayose, H., and Noike, K. (2002). "RENCON: toward a new evaluation system for performance rendering systems," in *Proceedings of the 2002 International Computer Music Conference (ICMC 2002)* (Gothenburg).

Hoffman, G., and Weinberg, G. (2011). Interactive improvisation with a robotic marimba player. *Auton. Robots* 31, 133–153. doi: 10.1007/s10514-011-9237-0

Honing, H. (2005). "Timing is tempo-specific," in *Proceedings of the 2005 International Computer Music Conference (ICMC 2005)* (Barcelona).

Honing, H. (2006). Computational modeling of music cognition: a case study on model selection. *Music Percept.* 23, 365–376. doi: 10.1525/mp.2006.23.5.365

Humphrey, E. J., Bello, J. P., and LeCun, Y. (2012). "Moving beyond feature design: deep architectures and automatic feature learning in music informatics," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)* (Porto), 403–408.

Huron, D., and Fantini, D. A. (1989). The avoidance of inner-voice entries: perceptual evidence and musical practice. *Music Percept.* 7, 43–47.

Juslin, P. (2003). Five facets of musical expression: a psychologist's perspective on music performance. *Psychol. Music* 31, 273–302. doi: 10.1177/03057356030313003

Juslin, P. N. (2001). "Communicating emotion in music performance: a review and theoretical framework," in *Music and Emotion: Theory and Research,* eds P. N. Juslin and J. A. Sloboda (Oxford, UK: Oxford University Press), 309–337.

Juslin, P. N., and Sloboda, J. (2011). *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press.

Katayose, H., Hashida, M., De Poli, G., and Hirata, K. (2012). On evaluating systems for generating expressive music performance: the rencon experience. *J. New Music Res.* 41, 299–310. doi: 10.1080/09298215.2012.745579

Kendall, R. A., and Carterette, E. C. (1990). The communication of musical expression. *Music Percept.* 8, 129–163.

Kim, T. H., Fukayama, S., Nishimoto, T., and Sagayama, S. (2010). "Performance rendering for polyphonic piano music with a combination of probabilistic models for melody and harmony," in *Proceedings of the 7th International Conference on Sound and Music Computing (SMC 2010)* (Barcelona), 23–30.

Kim, T. H., Fukayama, S., Nishimoto, T., and Sagayama, S. (2011). "Polyhymnia: an automatic piano performance system with statistical modeling of polyphonic expression and musical symbol interpretation," in *Proceedings of the 11th International Conference on New Interfaces for Musical Expression (NIME 2011)* (Oslo), 96–99.

Kim, T. H., Fukuyama, S., Nishimoto, T., and Sagayama, S. (2013). "Statistical approach to automatic expressive rendition of polyphonic piano music," in *Guide to Computing for Expressive Music Performance*, eds A. Kirke and E. R. Miranda (London: Springer), 145–179.

Kirke, A., and Miranda, E. R. (eds.). (2013). "An overview of computer systems for expressive music performance," in *Guide to Computing for Expressive Music Performance* (London: Springer-Verlag), 1–48.

Kosta, K., Bandtlow, O. F., and Chew, E. (2014). "Practical implications of dynamic markings in the score: is piano always piano?," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio* (London).

Kosta, K., Bandtlow, O. F., and Chew, E. (2015). "A change-point approach towards representing musical dynamics," in *Proceedings of the 5th International Conference on Mathematics and Computation in Music (MCM 2015)* (London), 179–184.

Kosta, K., Ramírez, R., Bandtlow, O. F., and Chew, E. (2016). Mapping between dynamic markings and performed loudness: a machine learning approach. *J. Math. Music* 10, 149–172. doi: 10.1080/17459737.2016.1193237

Krebs, F., and Grachten, M. (2012). "Combining score and filter based models to predict tempo fluctuations in expressive music performances," in *Proceedings of the Ninth Sound and Music Computing Conference (SMC 2012)* (Copenhagen).

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. New York, NY: Oxford University Press.

Langner, J., and Goebl, W. (2003). Visualizing expressive performance in tempo–loudness space. *Comput. Music J.* 27, 69–83. doi: 10.1162/014892603322730514

Leman, M., Lesaffre, M., and Maes, P.-J. (2017a). "Introduction: what is embodied music interaction?," in *The Routledge Companion to Embodied Music Interaction*, eds M. Lessafre, P. J. Maes, and M. Leman (New York, NY: Routledge), 1–10.

Leman, M., Nijs, L., and Di Stefano, N. (2017b). "On the role of the hand in the expression of music," in *The Hand: Perception, Cognition, Action*, eds M. Bertolaso and N. Di Stefano (Cham: Springer International Publishing), 175–192.

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

Li, S., Black, D., Chew, E., and Plumbley, M. D. (2014). "Evidence that phrase-level tempo variation may be represented using a limited dictionary," in *Proceedings of the 13th International Conference for Music Perception and Cognition (ICMPC13-APSCOM5)* (Seoul), 405–411.

Li, S., Black, D. A., and Plumbley, M. D. (2015). "The clustering of expressive timing within a phrase in classical piano performances by Gaussian Mixture Models," in *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR 2015)* (Plymouth), 322–345.

Li, S., Dixon, S., Black, D. A., and Plumbley, M. D. (2016). "A model selection test on effective factors of the choice of expressive timing clusters for a phrase," in *Proceedings of the 13th Sound and Music Conference (SMC 2016)* (Hamburg).

Li, S., Dixon, S., and Plumbley, M. D. (2017). "Clustering expressive timing with regressed polynomial coefficients demonstrated by a model selection test," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)* (Suzhou).

Liebman, E., Ornoy, E., and Chor, B. (2012). A phylogenetic approach to music performance analysis. *J. New Music Res.* 41, 195–222. doi: 10.1080/09298215.2012.668194

Liem, C., Hanjalic, A., and Sapp, C. (2011). "Expressivity in musical timing in relation to musical structure and interpretation: a cross-performance, audio-based approach," in *Audio Engineering Society Conference: 42nd International Conference: Semantic Audio* (Ilmenau).

Liem, C. C., Gómez, E., and Schedl, M. (2015). "PHENICX: innovating the classical music experience," in *Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (Turin).

Liem, C. C. S., and Hanjalic, A. (2011). "Expressive timing from cross-performance and audio-based alignment patterns: an extended case study," in *Proceedings of the 12th International Society for Music Information Retrieval (ISMIR 2011)* (Miami, FL), 519–525.

Liem, C. C. S., and Hanjalic, A. (2015). "Comparative analysis of orchestral performance recordings: an image-based approach," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)* (Malaga).

Lim, A., Mizumoto, T., Ogata, T., and Okuno, H. G. (2012). A musical robot that synchronizes with a coplayer using non-verbal cues. *Adv. Robot.* 26, 363–381. doi: 10.1163/156855311X614626

Longuet-Higgins, H. C., and Lee, C. S. (1982). The perception of musical rhythms. *Perception* 11, 115–128.

Longuet-Higgins, H. C., and Lee, C. S. (1984). The rhythmic interpretation of monophonic music. *Music Percept.* 1, 424–441.

Marchini, M., Papiotis, P., and Maestre, E. (2013). "Investigating the relationship between expressivity and synchronization in ensemble performance: an exploratory study," in *Proceedings of the International Symposium on Performance Science 2013 (ISPS 2013)* (Vienna), 217–222.

Marchini, M., Ramírez, R., Papiotis, P., and Maestre, E. (2014). The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *J. New Music Res.* 43, 303–317. doi: 10.1080/09298215.2014.922999

Masko, J., Friberg, J. F., and Friberg, A. (2014). "Software tools for automatic music performance," in *Proceedings of the 1st International Workshop on Computer and Robotic Systems for Automatic Music Performance (SAMP 14)* (Venice), 537–544.

Moelants, D., Demey, M., Grachten, M., Wu, C.-F., and Leman, M. (2012). The influence of an audience on performers: a comparison between rehearsal and concert using audio, video and movement data. *J. New Music Res.* 41, 67–78. doi: 10.1080/09298215.2011.642392

Molina-Solana, M., Arcos, J.-L., and Gómez, E. (2008). "Using expressive trends for identifying violin performers," in *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR 2008)* (Philadelphia, PA), 495–500.

Molina-Solana, M., Arcos, J.-L., and Gómez, E. (2010a). Identifying violin performers by their expressive trends. *Intell. Data Anal.* 14, 555–571. doi: 10.3233/IDA-2010-0439

Molina-Solana, M., Grachten, M., and Widmer, G. (2010b). "Evidence for pianist specific rubato style in chopin nocturnes," in *Proceedings of the 11th International Society for Music Information Retrieval (ISMIR 2010)* (Utrecht).

Moulieras, S., and Pachet, F. (2016). Maximum entropy models for generation of expressive music. *arXiv:1610.03606v1*. Available online at: https://arxiv.org/abs/1610.03606v1

Nakamura, E., Cuvillier, P., Cont, A., Ono, N., and Sagayama, S. (2015a). "Autoregressive hidden semi-markov model of symbolic music performance for score following," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)* (Málaga).

Nakamura, E., Nakamura, T., Saito, Y., Ono, N., and Sagayama, S. (2014a). Outer-product hidden Markov model and polyphonci MIDI score following. *J. New Music Res.* 43, 183–201. doi: 10.1080/09298215.2014.884145

Nakamura, E., Ono, N., and Sagayama, S. (2014b). "Merged-output HMM for piano fingering of both hands," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (Taipei), 531–536.

Nakamura, E., Ono, N., Sagayama, S., and Watanabe, K. (2015b). A stochastic temporal model of polyphonic MIDI performance with ornaments. *J. New Music Res.* 44, 287–304. doi: 10.1080/09298215.2015.1078819

Nakamura, E., Takeda, H., Yamamoto, R., Saito, Y., Sako, S., and Sagayama, S. (2013). Score following handling performances with arbitrary repeats and skips and automatic accompaniment. *J. Inform. Process. Soc. Jpn.* 54, 1338–1349. Available online at: http://id.nii.ac.jp/1001/00091563/

Nakamura, E., Yoshii, K., and Katayose, H. (2017). "Performance error detection and post-processing for fast and accurate symbolic music alignment," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2018)* (Suzhou), 347–353.

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures : The Implication-Realization Model*. Chicago, IL: University of Chicago Press.

Novembre, G., and Keller, P. E. (2014). A conceptual review on action-perception coupling in the musicians' brain: what is it good for? *Front. Hum. Neurosci.* 8:603. doi: 10.3389/fnhum.2014.00603

Ohishi, Y., Mochihashi, D., Kameoka, H., and Kashino, K. (2014). "Mixture of Gaussian process experts for predicting sung melodic contour with expressive dynamic fluctuations," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)* (Florence), 3714–3718.

Okumura, K., Sako, S., and Kitamura, T. (2011). "Stochastic modeling of a musical performance with expressive representations from the musical score," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)* (Miami, FL), 531–536.

Okumura, K., Sako, S., and Kitamura, T. (2014). "Laminae: a stochastic modeling-based autonomous performance rendering system that elucidates performer characteristics," in *Joint Proceedings of the 40th International Computer Music Conference (ICMC 2014) and the 11th Sound and Music Computing Conference (SMC 2014)* (Athens), 1271–1276.

Palmer, C. (1996). Anatomy of a performance: sources of musical expression. *Music Percept.* 13, 433–453.

Palmer, C. (1997). Music performance. *Annu. Rev. Psychol.* 48, 115–138.

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. Ph.D. thesis, City University London, London.

Peng, L., and Gerhard, D. (2009). "A gestural interface for orchestral conducting education," in *Proceedings of the First International Conference on Computer Supported Education (CSEDU 2009)* (Lisboa), 406–409.

Peperkamp, J., Hildebrandt, K., and Liem, C. C. S. (2017). "A formalization of relative local tempo variations in collections of performances," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)* (Suzhou).

Platz, F., and Kopiez, R. (2012). When the eye listens: a meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Percept.* 30, 71–83. doi: 10.1525/mp.2012.30.1.71

Ramírez, R., Maestre, E., Pertusa, A., Gómez, E., and Serra, X. (2007). Performance-based interpreter identification in saxophone audio recordings. *IEEE Trans. Circ. Syst. Video Technol.* 17, 356–364. doi: 10.1109/TCSVT.2007.890862

Raphael, C. (2001a). "Music plus one: a system for flexible and expressive musical accompaniment," in *Proceedings of the 2001 International Computer Music Conference (ICMC 2001)* (Havana).

Raphael, C. (2001b). Synthesizing musical accompaniments with Bayesian belief networks. *J. New Music Res.* 30, 59–67. doi: 10.1076/jnmr.30.1.59.7121

Raphael, C. (2009). "Symbolic and structural representation of melodic expression," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2010)* (Kobe), 555–560.

Raphael, C. (2010). "Music plus one and machine learning," in *Proceedings of the 30th International Conference on Machine Learning (ICML 2010)* (Haifa).

Repp, B. H. (1996). The art of inaccuracy: why pianists' errors are difficult to hear. *Music Percept.* 14, 161–184.

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

Repp, B. H., Windsor, L., and Desain, P. (2002). Effects of tempo on the timing of simple musical rhythms. *Music Percept.* 19, 565–593. doi: 10.1525/mp.2002.19.4.565

Rink, J. (ed.). (1995). *The Practice of Performance: Studies in Musical Interpretation*. Cambridge, UK: Cambridge University Press.

Rink, J. (ed.). (2002). *Musical Performance. A Guide to Understanding*. Cambridge, UK: Cambridge University Press.

Rink, J. (2003). In respect of performance: the view from musicology. *Psychol. Music* 31, 303–323. doi: 10.1177/03057356030313004

Rowe, R. (1992). *Interactive Music Systems: Machine Listening and Composing*. Cambridge, MA: MIT Press.

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178.

Sapp, C. S. (2007). "Comparative analysis of multiple musical performances," in *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR 2007)* (Vienna).

Sapp, C. S. (2008). "Hybrid numeric/rank similarity metrics for musical performance analysis," in *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR 2008)* (Philadelphia, PA), 501–506.

Sarasúa, Á., Melenhorst, M., Julià, C. F., and Gómez, E. (2016). "Becoming the maestro - a game to enhance curiosity for classical music," in *Proceedings of the 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES 2016)* (Barcelona), 1–4.

Saunders, C., Hardoon, D. R., Shawe-Taylor, J., and Widmer, G. (2008). Using string kernels to identify famous performers from their playing style. *Intell. Data Anal.* 12, 425–440. Available online at: https://content.iospress.com/articles/intelligent-data-analysis/ida00338?resultNumber=0&totalResults=229&start=0&q=Using+string+kernels+to+identify+famous+performers+from+their+playing+style&resultsPageSize=10&rows=10

Schlüter, J. (2017). *Deep Learning for Event Detection, Sequence Labelling and Similarity Estimation in Music Signals*. Ph.D. thesis, Johannes Kepler University Linz.

Schubert, E., Canazza, S., De Poli, G., and Rodà, A. (2017). Algorithms can mimic human piano performance: the deep blues of music. *J. New Music Res.* 46, 175–186. doi: 10.1080/09298215.2016.1264976

Schubert, E., De Poli, G., Rodà, A., and Canazza, S. (2014a). "Music systemisers and music empathisers – do they rate expressiveness of computer generated performances the same?," in *Joint Proceedings of the 40th International Computer Music Conference (ICMC 2014) and the 11th Sound and Music Computing Conference (SMC 2014)* (Athens), 223–227.

Schubert, E., Kreuz, G., and von Ossietzky, C. (2014b). "Open ended descriptions of computer assisted interpretations of musical performance: an investigation of individual differences," in *Proceedings of the 1st International Workshop on Computer and Robotic Systems for Automatic Music Performance (SAMP 14)* (Venice), 565–573.

Seashore, C. E. (1938). *Psychology of Music*. New York, NY: McGraw-Hill Book Company, Inc.

Sébastien, V., Ralambondrainy, H., Sébastien, O., and Conruyt, N. (2012). "Score analyzer: automatically determining scores difficulty level for instrumental e-Learning," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)* (Porto), 571–576.

Simon, I., Oore, S., Dieleman, S., and Eck, D. (2017). "Learning to create piano performances," in *Proceedings of the NIPS 2017 Workshop on Machine Learning for Creativity and Design* (Long Beach, CA).

Solis, J., and Takanishi, A. (2013). "Anthropomorphic musical robots designed to produce physically embodied expressive performances of music," in *Guide to Computing for Expressive Music Performance*, eds A. Kirke and E. R. Miranda (London, UK: Springer), 235–255.

Stamatatos, E., and Widmer, G. (2005). Automatic identification of music performers with learning ensembles. *Artif. Intell.* 165, 37–56. doi: 10.1016/j.artint.2005.01.007

Sundberg, J., Askenfelt, A., and Frydén, L. (1983). Musical performance: a synthesis-by-rule approach. *Comput. Music J.* 7, 37–43.

Tekman, H. G. (2002). Perceptual integration of timing and intensity variations in the perception of musical accents. *J. Gen. Psychol.* 129, 181–191. doi: 10.1080/00221300209603137

Teramura, K., Okuma, H., Taniguchi, Y., Makimoto, S., and Maeda, S. (2008). "Gaussian process regression for rendering music performance," in *Proceedings of the 10th International Conference on Music Perception and Cognition (ICMPC 10)* (Sapporo).

Thompson, S., Williamon, A., and Valentine, E. (2007). Time-dependent characteristics of performance evaluation. *Music Percept.* 25, 13–29. doi: 10.1525/mp.2007.25.1.13

Tobudic, A., and Widmer, G. (2006). Relational IBL in classical music. *Mach. Learn.* 64, 5–24. doi: 10.1007/s10994-006-8260-4

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

Toiviainen, P., Luck, G., and Thompson, M. R. (2010). Embodied meter: hierarchical eigenmodes in music-induced movement. *Music Percept.* 28, 59–70. doi: 10.1525/mp.2010.28.1.59

Tsay, C.-J. (2013). Sight over sound in the judgment of music performance. *Proc. Natl. Acad. Sci. U.S.A.* 110, 14580–14585. doi: 10.1073/pnas.1221454110

van Herwaarden, S., Grachten, M., and de Haas, W. B. (2014). "Predicting expressive dynamics in piano performances using neural networks," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (Taipei), 47–52.

Vos, J., and Rasch, R. (1981). The perceptual onset of musical tones. *Percept. Psychophys.* 29, 323–335.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Wapnick, J., Campbell, L., Siddell-Strebel, J., and Darrow, A.-A. (2009). Effects of non-musical attributes and excerpt duration on ratings of high-level piano performances. *Music. Sci.* 13, 35–54. doi: 10.1177/1029864909013001002

Wesolowski, B. C., Wind, S. A., and Engelhard, G. (2016). Examining rater precision in music performance assessment: an analysis of rating scale structure using the multifaceted rasch partial credit model. *Music Percept.* 33, 662–678. doi: 10.1525/mp.2016.33.5.662

Widmer, G. (1995). Modeling the rational basis of musical expression. *Comput. Music J.* 19, 76–96.

Widmer, G. (1996). Learning expressive performance: the structure-level approach. *J. New Music Res.* 25, 179–205.

Widmer, G. (2000). "Large-scale induction of expressive performance rules: first quantitative results," in *Proceedings of the 2000 International Computer Music Conference (ICMC 2000)* (Berlin).

Widmer, G. (2003). Discovering simple rules in complex data: a meta-learning algorithm and some surprising musical discoveries. *Artif. Intell.* 146, 129–148. doi: 10.1016/S0004-3702(03)00016-X

Widmer, G. (2017). Getting closer to the essence of music: the *Con Espressione* Manifesto. *ACM Trans. Intell. Syst. Technol.* 8, 1–13. doi: 10.1145/2899004

Widmer, G., Flossmann, S., and Grachten, M. (2009). YQX plays chopin. *AI Mag.* 30, 35–48. doi: 10.1609/aimag.v30i3.2249

Widmer, G., and Goebl, W. (2004). Computational models of expressive music performance: the state of the art. *J. New Music Res.* 33, 203–216. doi: 10.1080/0929821042000317804

Widmer, G., and Tobudic, A. (2002). Playing mozart by analogy: learning multi-level timing and dynamics strategies. *J. New Music Res.* 32, 259–268. doi: 10.1076/jnmr.32.3.259.16860

Wiggins, G. A., Müllensiefen, D., and Pearce, M. T. (2010). On the non-existence of music: why music theory is a figment of the imagination. *Music. Sci.* 14(1 Suppl.), 231–255. doi: 10.1177/10298649100140S110

Xia, G. (2016). *Expressive Collaborative Music Performance via Machine Learning*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Xia, G., and Dannenberg, R. B. (2015). "Duet interaction: learning musicianship for automatic accompaniment," in *Proceedings of the 15th International Conference on New Interfaces for Musical Expression (NIME 2015)* (Baton Rouge, LA).

Xia, G., Wang, Y., Dannenberg, R., and Gordon, G. (2015). "Spectral learning for expressive interactive ensemble music performance," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)* (Málaga), 816–822.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.