Modeling and Control of Expressiveness in Music Performance

SERGIO CANAZZA, GIOVANNI DE POLI, MEMBER, IEEE, CARLO DRIOLI, MEMBER, IEEE, ANTONIO RODÀ, AND ALVISE VIDOLIN

Invited Paper

Expression is an important aspect of music performance. It is the added value of a performance and is part of the reason that music is interesting to listen to and sounds alive. Understanding and modeling expressive content communication is important for many engineering applications in information technology. For example, in multimedia products, textual information is enriched by means of graphical and audio objects. In this paper, we present an original approach to modify the expressive content of a performance in a gradual way, both at the symbolic and signal levels. To this purpose, we discuss a model that applies a smooth morphing among performances with different expressive content, adapting the audio expressive character to the user's desires. Morphing can be realized with a wide range of graduality (from abrupt to very smooth), allowing adaptation of the system to different situations. The sound rendering is obtained by interfacing the expressiveness model with a dedicated postprocessing environment, which allows for the transformation of the event cues. The processing is based on the organized control of basic audio effects. Among the basic effects used, an original method for the spectral processing of audio is introduced.

Keywords*—Audio, expression communication, multimedia, music, signal processing.*

Manuscript received February 4, 2003; revised November 8, 2003. This work was supported by Multisensory Expressive Gesture Applications (MEGA) Project IST 1999-20410.

S. Canazza is with Mirage, Department of Scienze Storiche e Documentarie, University of Udine (Polo di Gorizia), Gorizia 34170, Italy, and also with the Centro di Sonologia Computazionale of Padova, Department of Information Engineering, University of Padova, Padova 35131, Italy (e-mail: sergio.canazza@uniud.it, canazza@dei.unipd.it).

S. Canazza, G. De Poli and A. Vidolin are with the Centro di Sonologia Computazionale, Department of Information Engineering, University of Padova, Padova 35131, Italy (e-mail: depoli@dei.unipd.it; vidolin@dei.unipd.it).

C. Drioli is with the Centro di Sonologia Computazionale, Department of Information Engineering, University of Padova, Padova 35131, Italy, and also with the Department of Phonetics and Dialectology, Institute of Cognitive Sciences and Technology, Italian National Research Council (ISTC-CNR), Padova 35121, Italy (e-mail: carlo.drioli@csrf.pd.cnr.it; drioli@csrf.pd.cnr.it).

A. Rodà is with the Centro di Sonologia Computazionale, Department of Information Engineering, University of Padova, Padova 35131, Italy, and also with the Mirage, Dipartimento di Scienze Storiche e Documentarie, University of Udine (sede di Gorizia), Gorizia, Italy (e-mail: rodant@tin.it; ar@csc.unipd.it).

Digital Object Identifier 10.1109/JPROC.2004.825889

I. INTRODUCTION

Understanding and modeling expressive content communication is important for many engineering applications in information technology. In multimedia products, textual information is enriched by means of graphical and audio objects. A correct combination of these elements is extremely effective for the communication between author and user. Usually, attention is put on the visual part rather than sound, which is merely used as a realistic complement to image or as a musical comment to text and graphics. With increasing interaction, the visual part has evolved consequently while the paradigm of the use of audio has not changed adequately, resulting in a choice among different objects rather than in a continuous transformation on these. A more intensive use of digital audio effects will allow us to interactively adapt sounds to different situations, leading to a deeper fruition of the multimedia product. It is advisable that the evolution of audio interaction leads to the involvement of expressive content. Such an interaction should allow a gradual transition (morphing) between different expressive intentions. Recent researches have demonstrated that it is possible to communicate expressive content at an abstract level, so as to change the interpretation of a musical piece [1].

In human musical performance, acoustical or perceptual changes in sound are organized in a complex way by the performer in order to communicate different emotions to the listener. The same piece of music can be performed trying to convey different specific interpretations of the score, by adding mutable expressive intentions. A textual or musical document can assume different meanings and nuances depending on how it is performed; see [2] for an overview of models of expressiveness in speech. In multimedia, when a human performer is not present, it is necessary to have models and tools that allow the modification of a performance by changing its expressive intention. The aim of this paper is to address this problem by proposing a model

for continuous transformation of expressive intentions of a music performance.

Research on music performance carried out in these last decades have analyzed the rules imposed by musical praxis. In fact, audio content normally is represented by a musical score. A mechanical performance (played with the exact values indicated in the score) of that score is, however, lacking of musical meaning and is perceived dull as a text read without any prosodic inflexion. Indeed, human performers never respect tempo, timing, and loudness notations in a mechanical way when they play a score; some deviations are always introduced [3]. These deviations change with the music style, instrument, and musician [4]. A performance which is played accordingly to appropriate rules imposed by a specific musical praxis, will be called *natural*. Moreover, Clynes [5] evidenced the existence of composer's pulses consisting of combined amplitude and timing warps, and specific to each composer. There are also some implicit rules that are related to different musical styles and musical epoch that are verbally handed on and used in the musical practice. Furthermore, a musician has his own performance style and his own interpretation of the musical structure, resulting in a high degree of deviation from the notation of the score. Repp [6] deeply analyzed a lot of professional pianists' performances, measuring deviations in timing and articulation; his results showed the presence of deviation patterns related to musical structure.

Studies in music performance use the word *expressiveness* to indicate the systematic presence of deviations from the musical notation as a communication means between musician and listener [7]. The analysis of these systematic deviations has led to the formulation of several models that try to describe their structures and aim at explaining where, how, and why a performer modifies, sometimes in an unconscious way, what is indicated by the notation of the score. It should be noticed that although deviations are only the external surface of something deeper and not directly accessible, they are quite easily measurable, and, thus, it is useful in developing computational models in scientific research. Some models based on an analysis-by-measurement method have been proposed [8]–[12]. This method is based on the analysis of deviations measured in recorded human performances. The analysis aims at recognizing regularities in the deviation patterns and to describe them by means of mathematical relationships. Another approach derives models, which are described with a collection of parametric rules, using an analysis-by-synthesis method. The most important is the KTH rule system [13]–[15]. Other rules were developed by De Poli [16]. Rules describe quantitatively the deviations to be applied to a musical score, in order to produce a more attractive and humanlike performance than the mechanical one that results from a literal playing of the score. Every rule tries to predict (and to explain with musical or psychoacoustic principles) some deviations that a human performer inserts. Machine learning performance rules is another active research stream. Widmer [17], [18] and Katayose [19] used some artificial intelligence (AI) inductive algorithms to infer performance rules from recorded performances. Similar approaches with AI algorithms using case-based reasoning were proposed by Arcos [20] and Suzuki [21]. Several methodologies of approximation of human performances were developed using neural network techniques [22], a fuzzy logic approach [23], [24] or a multiple regression analysis [25]. Most systems act at symbolic (note-description) level; only Arcos [26] combined it with sound processing techniques for changing a recorded musical performance.

All the above researches aim at explaining and modeling the natural performance. However, the same piece of music can be performed trying to convey different expressive intentions [7], changing the natural style of the performance. One approach for modeling different expressive intentions is being carried out by Bresin and Friberg [27]. Starting from the above-mentioned KTH rules, they developed some macro rules for selecting appropriate values for the parameters in order to convey different emotions.

In this paper, we present a different approach to modify the expressive content of a performance in a gradual way both at symbolic and signal level. The paper is organized as follows. Section II introduces the schema of our system for interactive control of expressiveness; in Section III, a general overview of the expressiveness model and its different levels are given; Section IV discusses the rendering of expressive deviations in prerecorded audio performances by appropriate expressive processing techniques. In Section V, we present the results and some practical examples of the proposed methodology and the assessment based on perceptual tests.

II. System Overview

A musical interpretation is often the result of a wide range of requirements on expressiveness rendering and technical skills. The understanding of why certain choices are, often unconsciously, preferred to others by the musician, is a problem related to cultural aspects and is beyond the scope of this work. However, it is still possible to extrapolate significant relations between some aspects of the musical language and a class of systematic deviations. For our purposes, it is sufficient to introduce two sources of expression. The first one deals with aspects of musical structures such as phrasing, hierarchical structure of phrase, harmonic structure and so on [4], [6], [11], [12]. The second involves those aspects that are referred to with the term *expressive intention*, and that relate to the communication of moods and feelings. In order to emphasize some elements of the music structure (i.e., phrases, accents, etc.), the musician changes his performance by means of expressive patterns as crescendo, decrescendo, sforzando, rallentando, etc.; otherwise, the performance would not sound musical. Many papers analyzed the relation or, more correctly, the possible relations between music structure and expressive patterns [28], [29].

Let us call *neutral* performance a human performance played without any specific expressive intention, in a scholastic way and without any artistic aim. Our model is based on the hypothesis that when we ask a musician to

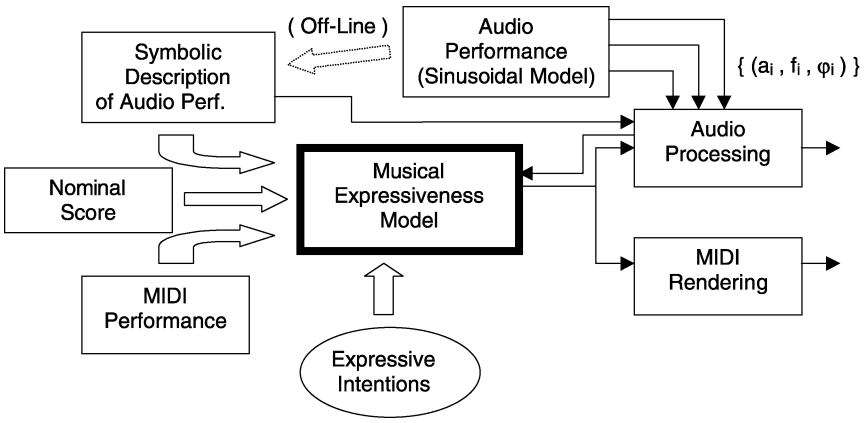


Fig. 1. Scheme of the system. The input of the expressiveness model is composed of a musical score and a description of a neutral musical performance. Depending on the expressive intention desired by the user, the expressiveness model acts on the symbolic level, computing the deviations of all musical cues involved in the transformation. The rendering can be done by a MIDI synthesizer and/or driving the audio processing engine. The audio processing engine performs the transformations on the prerecorded audio in order to realize the symbolic variations computed by the model.

play in accordance with a particular expressive intention, he acts on the available freedom degrees, without destroying the relation between music structure and expressive patterns [11]. Already in the neutral performance, the performer introduces a phrasing that translates into time and intensity deviations respecting the music structure. In fact, our studies demonstrate [30] that by suitably modifying the systematic deviations introduced by the musician in the neutral performance, the general characteristics of the phrasing are retained (thus keeping the musical meaning of the piece), and different expressive intentions can be conveyed.

The purpose of this research is to control in an automatic way the expressive content of a neutral (prerecorded) performance. The model adds an expressive intention to a neutral performance in order to communicate different moods, without destroying the musical structure of the score. The functional structure of the system used as a testbed for this research is shown in Fig. 1.

In multimedia systems, musical performance are normally stored as a Musical Instrument Digital Interface (MIDI) score or audio signal. The MIDI protocol allows electronic devices to interact and work in synchronization with other MIDI compatible devices. It does not send the actual musical note, but the information about the note. It can send messages to synthesizers telling it to change sounds, master volume, modulation devices, which note was depressed, and even how long to sustain the note [31], [32]. Our approach can deal with a melody in both representations. The input of the expressiveness model is composed of a description of a neutral musical performance and a control on the expressive intention desired by the user. The expressiveness model acts on the symbolic level, computing the deviations of all musical cues involved in the transformation. The rendering can be done by a MIDI synthesizer and/or driving the audio processing engine. The audio processing engine performs the transformations on the prerecorded audio in order to realize the symbolic variations computed by the model. The system allows the user to interactively change the expressive intention of a performance by specifying its own preferences through a graphical interface.

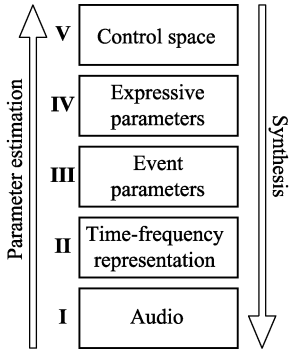


Fig. 2. Multilevel representation.

III. Multilevel Representation

To expressively process a performance, a multilevel representation of musical information is proposed and the relation between adjacent levels is outlined (Fig. 2). The first level is the 44.1-kHz, 16-b digital *audio signal*.

The second level is the *time-frequency* (TF) representation of the signal which is required for analysis and transformation purposes. TF representations are appreciated in the field of musical signal processing because they provide a reliable representation of musical sounds as well as an effective and robust set of transformation tools [33]. The specific TF representation adopted here relies on the well-known sinusoidal model of the signal [34], [35], which has been previously used in the field of musical signal processing with convincing results (see, e.g., [26]), and for which a software tool is freely available (SMS, [36]). The analysis algorithm acts on windowed portions (here called *frames*) of the signal, and produces a time-varying representation as sum of sinusoids (here called *partials*), which frequencies, amplitudes, and phases slowly vary over time. Thus, the ith frame of the sinusoidal modeling is a set $\{(f_h(i), a_h(i), \phi_h(i))\}_{h=1}^{H}$

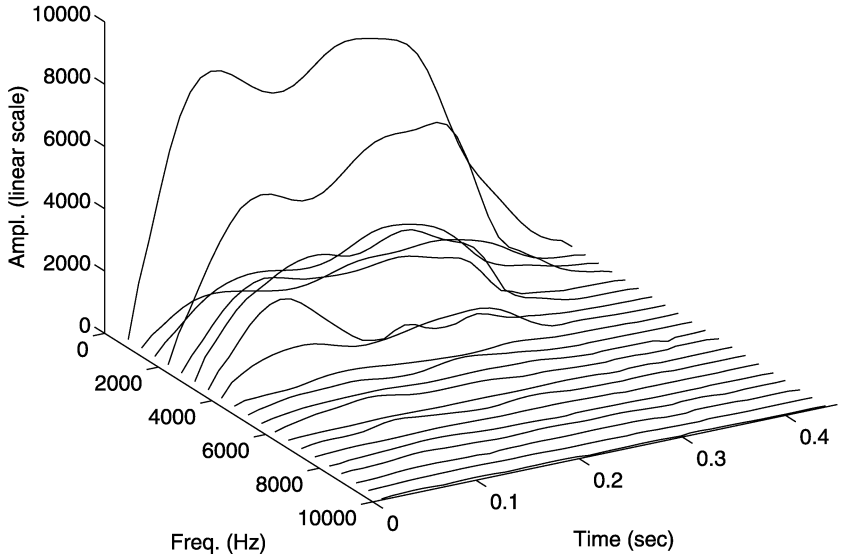


Fig. 3. TF representation of a violin tone: frequencies and amplitudes (only 20 partials are shown).

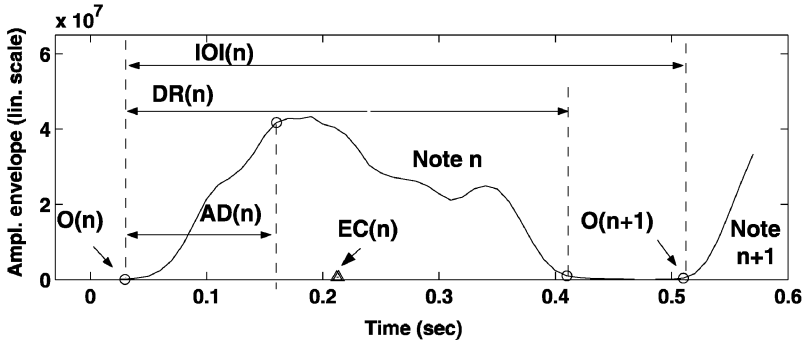


Fig. 4. Musical parameters involved in the control of expressiveness.

of triples of frequency, amplitude, and phase parameters describing each partial. H, the number of partials, is taken high enough to provide the maximum needed bandwidth. The noisy (or *stochastic*) part of the sound, i.e., the difference between the original signal and the sinusoidal reconstruction, is sometimes modeled as an autoregressive (AR) stochastic process. However, we will not consider this component here, and we use the sinusoidal signal representation to model string- and windlike nonpercussive musical instruments. Looking at the TF representation, Fig. 3, the signal appears extremely rich in microvariations, which are responsible for the aliveness and naturalness of the sound.

The third level represents the knowledge on the musical performance as *events*. This level corresponds to the same level of abstraction of the MIDI representation of the performance, e.g., as obtained from a sequencer (MIDI list events). A similar event description can be obtained from an audio performance. A performance can be considered as a sequence of notes. The nth note is described by the pitch value FR(n), the Onset time $O(n)$, and Duration DR(n) (which are time-related parameters), and by a set of timbre-related parameters: Intensity $I(n)$, Brightness BR(n) (measured as the centroid of the spectral envelope [37]), and energy envelope, described by Attack Duration AD(n) and Envelope Centroid EC(n) (i.e., the temporal centroid of the dynamic profile of the note). This representation can be obtained from the TF representation by a semiautomatic segmentation. From the time-related parameters, the Inter Onset Interval IOI$(n) = O(n+1) - O(n)$ and the Legato $L(n) =$ DR(n)/IOI(n) parameters are derived. Fig. 4 and Table 1 show the principal parameters introduced. A more detailed description of musical and acoustical parameters involved in the analysis of expressiveness can be found in [11]. The parameters $P(n)$ (from now on, P-parameters) that will be modified by the model are $L(n)$, IOI(n), and the timbre-related parameters key velocity for MIDI

Table 1
P-Parameters at the Third-Level Representation

$FR(n)$	pitch value
$O(n)$	onset time
$DR(n)$	duration
$IOI(n)$	inter onset interval
$L(n)$	legato
$I(n)$	intensity
$BR(n)$	brightness
$AD(n)$	attack duration
$EC(n)$	envelope centroid

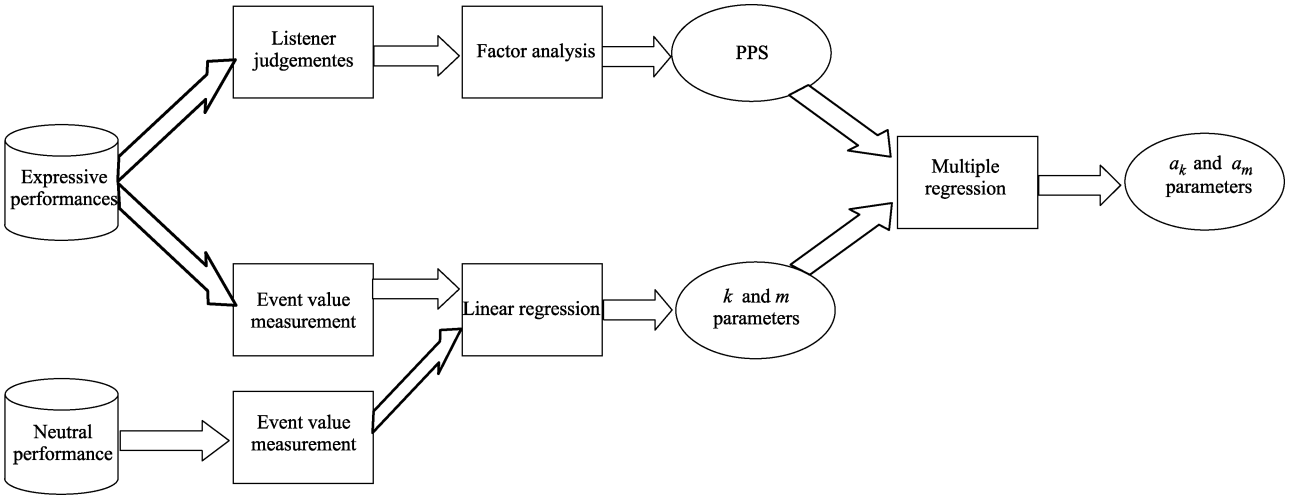


Fig. 5. Computation of the parameters of the model.

performance or $I(n)$, BR(n), AD(n), and EC(n) for audio performance.

The fourth level represents the internal parameters of the expressiveness model. We will use, as *expressive* representation a couple of values $E = \{k, m\}$ for every P-parameter. The meaning of these values will be explained in the next section.

The last level is the *control space* (i.e., the user interface), which controls, at an abstract level, the expressive content and the interaction between the user and the audio object of the multimedia product.

A. The Expressiveness Model

The model is based on the hypothesis, introduced in Section II, that different expressive intentions can be obtained by suitable modifications of a neutral performance. The transformations realized by the model should satisfy some conditions: 1) they have to maintain the relation between structure and expressive patterns and 2) they should introduce as few parameters as possible to keep the model simple. In order to represent the main characteristics of the performances, we used only two transformations: shift and range expansion/compression. Different strategies were tested. Good results were obtained [30] by a linear instantaneous mapping that, for every P-parameter and a given expressive intention e, is formally represented by

$$P_e(n) = k_e \overline{P}_0 + m_e \left(P_0(n) - \overline{P}_0\right) \tag{1}$$

where $P_e(n)$ is the estimated profile of the performance related to expressive intention e, $P_0(n)$ is the value of the P-parameter of the nth note of the neutral performance, $\overline{P}_0$ is the mean of the profile $P_0(n)$ computed over the entire vector, k_e and m_e are, respectively, the coefficients of shift and expansion/compression related to expressive intention. We verified that these parameters are very robust in the modification of expressive intentions [38]. Thus, (1) can be generalized to obtain, for every P-parameter, a morphing among different expressive intentions as

$$P(n) = k(x,y)\overline{P}_0 + m(x,y)\left(P_0(n) - \overline{P}_0\right) \tag{2}$$

This equation relates every P-parameter with a generic expressive intention represented by the expressive parameters k and m that constitute the fourth-level representation and that can be put in relation to the position (x, y) of the control space.

B. The Control Space

The control space level controls the expressive content and the interaction between the user and the final audio performance. In order to realize a morphing among different expressive intentions we developed an abstract control space, called perceptual parametric space (PPS), that is a two-dimensional (2-D) space derived by multidimensional analysis (principal component analysis) of perceptual tests on various professionally performed pieces ranging from Western classical to popular music [29], [39]. This space reflects how the musical performances are organized in the listener's mind. It was found that the axes of PPS are correlated to acoustical and musical values perceived by the listeners themselves [40]. To tie the fifth level to the underlying ones, we make the hypothesis that a linear relation exists between the PPS axes and every couple of expressive parameters $\{k, m\}$

$$\begin{aligned} k(x,y) &= a_{k,0} + a_{k,1}x + a_{k,2}y \\ m(x,y) &= a_{m,0} + a_{m,1}x + a_{m,2}y \end{aligned} \tag{3}$$

where x and y are the coordinates of the PPS.

C. Parameter Estimation

Event, expressive and the control levels are related by (1) and (3). We will now get into the estimation process of the model parameters (see Fig. 5); more details about the relation between x, y, and audio and musical values will be given in Sections IV and V.

The estimation is based on a set of musical performances, each characterized by a different expressive intention. Such recordings are made by asking a professional musician to perform the same musical piece, each time being inspired by a different expressive intention (see Section V for details).

Moreover, a neutral version of the same piece is recorded. Recordings are first judged by a group of listeners, who assign different scores to the performances with respect to a scoring table in which the selectable intentions are reported (see [40] for more details). Results are then processed by a factor analysis. In our case [29], [39], this analysis allowed us to recognize two principal axes explaining at least the 75% of the total variance. The choice of only two principal factors, instead of three or four, is not mandatory. However, this choice results in a good compromise between the completeness of the model and the compactness of the parameter control space (PPS). The visual interface, being the 2-D control space, is effective and easy to realize. Every performance can be projected in the PPS by using its factor loading as x and y coordinates. Let us call (x_e, y_e) the coordinates of the performance e in the PPS. Table 4 in Section V shows the factor loadings obtained from factor analysis. These factor loadings are assumed as coordinates of the expressive performances in the PPS.

An acoustical analysis is then carried out on the expressive performances, in order to measure the deviations' profiles of the P-parameters. For each expressive intention, the profiles are used to perform a linear regression with respect to the corresponding profiles evaluated in the neutral performance, in order to obtain k_e and m_e in the model in (1). The result is a set of expressive parameters E, for each expressive intention and each of the P-parameters. Given x_e, y_e, and k_e, m_e estimated as above, for every P-parameter the corresponding coefficients $a_{k,i}$ and $a_{m,i}$ $(i = 0,\ 1,\ 2)$ of (3) are estimated by multiple linear regression, over expressive intentions.

Up to this point, the schema of Fig. 2 has been covered bottom-up, computing the model parameters from a set of sample performances. Therefore, it is possible to change the expressiveness of the neutral performance by selecting an arbitrary point in the PPS, and computing the deviations of the low-level acoustical parameters. Let us call x_p and y_p the coordinates of a (possibly time-varying) point in the PPS. From (3), for every P-parameter, $k(x, y)$ and $m(x, y)$ values are computed. Then, using (2), the profiles of event-layer cues are obtained. These profiles are used for the MIDI synthesis and as input to the postprocessing engine acting at levels 1 and 2, according to the description in the next section.

IV. Real-Time Rendering

The rendering of expressive variations on digitally recorded audio relies on a sound processing engine based on the sinusoidal representation. The expressiveness model outlined in Section III is adapted to produce the time-varying controls of the sound processing engine, focusing on a wide class of musical signals, namely monophonic and quasi-harmonic sounds such as wind instruments and solo string instruments. All the principal sound effects are obtained by controlling the parameters of the sinusoidal representation, and are briefly summarized. *Time stretching* is obtained by changing the frame rate of resynthesis and by interpolating between the parameters of two frames in case of noninteger step. *Pitch shift* is obtained by scaling the frequencies of the harmonics and by preserving formants with spectral envelope interpolation. *Intensity* and *brightness* control is achieved by scaling the amplitude of partials in an appropriate way, so as to preserve the natural spectral characteristics of the sound when its intensity and brightness are modified. We stress here the fact that spectral modifications can occur mainly as a function of the performance dynamic level, or even as a function of *ad hoc* performance actions influencing the timbre, depending on the degree of control offered by the musical instrument. The nature of the instrument will, thus, determine the degree of independence of the brightness control from the intensity control.

Table 2
Multiplicative Factors of Musical Parameters and Basic Audio Effects

γ_{IOI}	time stretching
γ_{AD}	time stretching
γ_{DR}	time stretching
γ_L	time stretching & spectral processing
γ_I	spectral processing & envelope scaling
γ_{EC}	envelope scaling
γ_{BR}	spectral processing

To the purpose of modeling these spectral cues in expressive musical performances, an original spectral processing method is introduced. This permits the reproduction of the spectral behavior exhibited by a discrete set of sound examples, whose intensity or brightness varies in the desired interval depending on the expressive intention of the performance.

Let us introduce a set of multiplicative factors γ_{IOI}, γ_{AD}, γ_{DR}, γ_L, γ_I, γ_{EC}, γ_{BR}, representing the changes of the musical parameters under the control of the audio processing engine. The first three factors are the time-stretching factors of the IOI interval, the attack duration, and the duration of the whole note, respectively. The *Legato* variation factor is related to the variations of the note duration and of IOI, and can be expressed as $\gamma_L = \gamma_{\mathrm{DR}}/\gamma_{\mathrm{IOI}}$. The intensity factor γ_I specifies a uniform change of the dynamic level over the whole note. The factor γ_{EC} specifies a change in the temporal position of the dynamic profile centroid of the note, and is related to a nonuniform scaling of the dynamic profile over the note duration. The factor γ_{BR} specifies a modification of the spectral centroid over the whole note, and is related to a reshaping of the original short-time spectral envelopes over the note duration. The rendering of the deviations computed by the model may, thus, imply the use of just one of the basic sound effects seen above, or the combination of two or more of these effects (see Table 2), with the following general rules.

Local Tempo: Time stretching is applied to each note. It is well known that in strings and winds, the duration of the attack is perceptually relevant for the characterization of the conveyed expressive intention. For this reason, a specific time-stretching factor is computed for the attack segment and is directly related to the γ_{AD} indicated by the model. The computation of the time stretch control on the

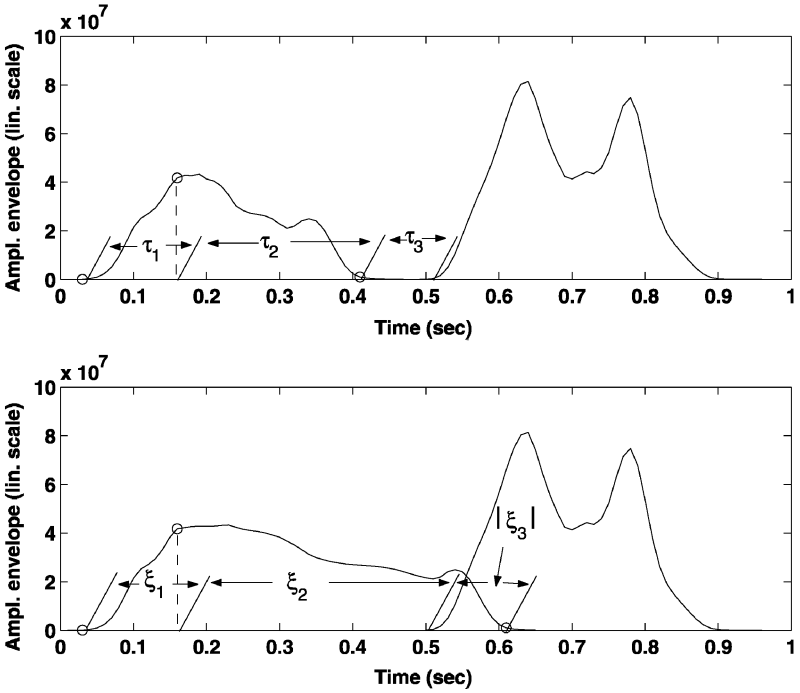


Fig. 6. Energy envelope of two violin notes. Upper panel: original natural performance. Lower panel: overlapping adjacent notes after a modification of the *Legato* parameter.

note relies on the cumulative information given by the γ_{AD} and γ_{IOI} factors, and on the γ_{DR} deviation induced by the *Legato* control considered in the next item.

Legato: This musical feature is recognized to have great importance in the expressive characterization of wind and string instruments performances. However, the processing of *Legato* is a critical task that would imply the reconstruction of a note release and a note attack if the notes are originally tied in a *Legato*, or the reconstruction of the transient if the notes are originally separated by a *micropause*. In both cases, a correct reconstruction requires a deep knowledge of the instrument dynamic behavior, and a dedicated synthesis framework would be necessary. Our approach to this task is to approximate the reconstruction of transients by interpolation of amplitudes and frequency tracks.

The deviations of the *Legato* parameter are processed by means of two synchronized actions: the first effect of a *Legato* change is a change in the duration of the note by γ_{DR}, since $L_e = L_0\ \gamma_L = (\mathrm{DR}_0\ \gamma_{\mathrm{DR}})/(\mathrm{IOI}_0\ \gamma_{\mathrm{IOI}})$, where L_0 is the original *Legato* degree and L_e is the *Legato* for the new expressive intention. This time-stretching action must be added to the one considered for the *Local Tempo* variation, as we will see in detail. Three different time-stretching zones are recognized within each note (with reference to Fig. 6): attack, sustain and release, and micropause. The time-stretching deviations must satisfy the following relations:

$$\begin{cases} \xi_1 = \tau_1\ \gamma_{\mathrm{AD}} \\ \xi_1 + \xi_2 = (\tau_1 + \tau_2)\gamma_{\mathrm{DR}} \\ \xi_1 + \xi_2 + \xi_3 = (\tau_1 + \tau_2 + \tau_3)\gamma_{\mathrm{IOI}} \end{cases}$$

where τ_1, τ_2, and τ_3, are the duration of the attack, sustain–release, and micropause segment, respectively, and ξ_1, ξ_2, and ξ_3, are the new duration of these segments. Each region will be processed with a time stretch coefficient computed from the above equations

$$\begin{cases} \gamma_1 = \frac{\xi_1}{\tau_1} = \gamma_{\mathrm{AD}} \\ \gamma_2 = \frac{\xi_2}{\tau_2} = \frac{(\tau_1+\tau_2)\gamma_L\ \gamma_{\mathrm{IOI}} - \gamma_1\tau_1}{\tau_2} \\ \gamma_3 = \frac{\xi_3}{\tau_3} = \frac{-(\tau_1+\tau_2)\gamma_L\ \gamma_{\mathrm{IOI}} + (\tau_1+\tau_2+\tau_3)\gamma_{\mathrm{IOI}}}{\tau_3} \end{cases} \tag{4}$$

where γ_1, γ_2, and γ_3 are the time-stretching factors of the attack, sustain–release, and micropause segment, respectively. If an overlap occurs due to the lengthening of a note, the time stretch coefficient γ_3 in (4) becomes negative. In this case, the second action involved is a spectral linear interpolation between the release and attack segments of two adjacent notes over the overlapping region (see Fig. 6). The length of the overlapping region is determined by the *Legato* degree, and the interpolation within partial amplitude will be performed over the whole range. The frequency tracks of the sinusoidal representation are lengthened to reach the pitch transition point. Here, a 10- to 15-ms transition is generated by interpolating the tracks of the actual note with those of the successive note. In this way, a transition without *glissando* is generated. Glissando effects can be controlled by varying the number of interpolated frames. This procedure, used to reproduce the smooth transition when the stretched note overlaps with the following note, is a severe simplification of instruments transients, but is general and efficient enough for real-time purposes.

Envelope Shape: The center of mass of the energy envelope is related to the musical *accent* of the note, which is usually located on the attack for *Light* or *Heavy* intentions, or close to the end of note for *Soft* or *Dark* intentions. To change the position of the center of mass, a triangular-shaped func-

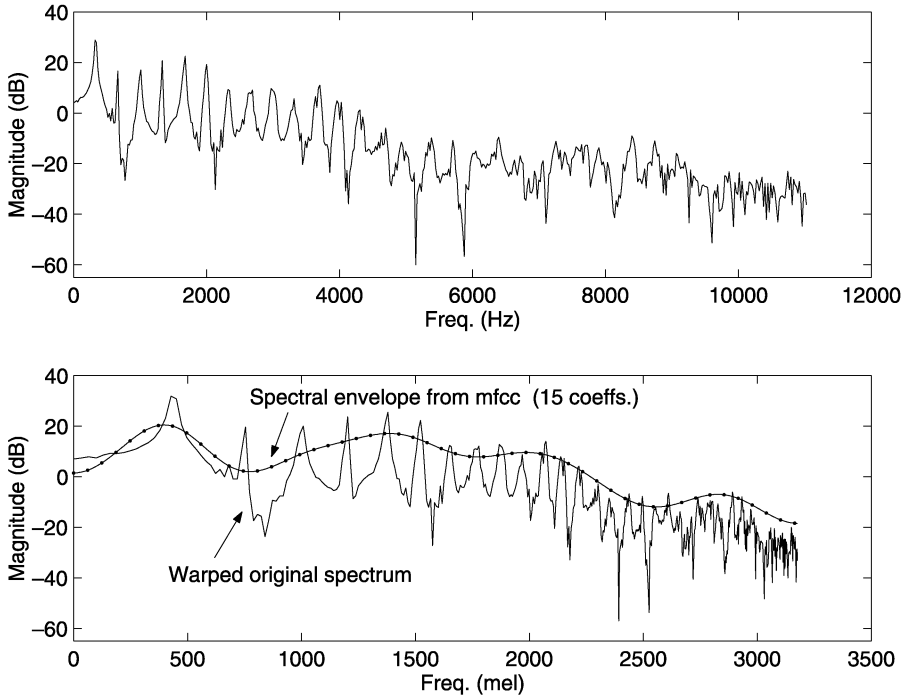


Fig. 7. Spectral envelope representation by mel-cepstrum coefficients. Upper panel: original spectrum (frequency axis in hertz). Lower panel: warped and smoothed version of the original spectrum, and the spectral envelope obtained by using 15 mel-cepstrum coefficients (frequency axis in mel).

tion is applied to the energy envelope, where the apex of the triangle corresponds to the new position of the accent.

Intensity and Brightness Control: The intensity and brightness of the sound frame are controlled by means of a spectral processing model relying on learning from real data the spectral transformations which occur when such a musical parameter changes. First, a perceptually weighted representation of spectral envelopes is introduced, so that the perceptually relevant differences are exploited in the comparison of spectral envelopes. Next, the parametric model used to represent spectral changes is outlined. Finally, the proposed method is applied to the purpose of modeling the intensity and brightness deviation for the control of expressiveness.

A. Representation of Spectral Envelopes

To switch from the original sinusoidal description to a perceptual domain, the original spectrum is turned to the *mel-cepstrum* spectral representation. The mel-frequency cepstral coefficients (mfcc) for a given sound frame are defined as the discrete-cosine transform (DCT) of the frequency domain logarithmic output of a mel-spaced filter bank. The first $M+1$ mel-cepstrum coefficients $\{b_i\}_{i=0}^{M}$, where M is usually in the range 10–30, represent a smooth and warped version of the spectrum, as the inversion of the DCT leads to

$$|C(\lambda)| = b_0 + 2\sum_{i=1}^{M} b_i \cos\left(\frac{\pi\lambda i}{2B_H}\right) \tag{5}$$

where λ is the frequency in mel, b_0 is the frame energy, and $B_H = \min\{mel(f_H), mel(F_s/2)\}$ with F_s being the sampling frequency. The normalization factor B_H is introduced to ensure that the upper limit of the band corresponds to a value of one on the normalized warped frequency axis. The conversion from hertz to mel is given by the analytical formula $\lambda = mel(f) \approx 1127\log(1+f/700)$ [41]. Fig. 7 shows an example of a mel-cepstrum spectral envelope.

The above definition of mel-cepstrum coefficients usually applies for a short sound buffer in the time-domain. To convert from a sinusoidal representation, alternative methods such as the discrete cepstrum method [42] are preferred: for a given sinusoidal parametrization, the magnitudes $\{a_h\}_{h=1}^{H}$ of the partials are expressed in the log domain and the frequencies $\{f_h\}_{h=1}^{H}$ in hertz are converted to mel frequencies $\{\lambda_h\}_{h=1}^{H}$. The real mel-cepstrum parameters $\{b_i\}_{i=0}^{M}$ are finally computed by minimizing the following least-squares (LS) criterion:

$$\sum_{h=1}^{H} \left(|C(\lambda_h)| - 20\log_{10}(a_h)\right)^2. \tag{6}$$

The aim of the mel-cepstrum transformation in our framework is to capture the perceptually meaningful differences between spectra by comparing the smoothed and warped versions of spectral envelopes.

We call now $c_h = |C(\lambda_h)| = |C(mel(f_h))|$ the hth partial magnitude (in dB) of the mel-cep\-strum spectral envelope, and $\Delta C = \{\Delta C_h\}_{h=1}^{H}$, with $\Delta C_h = (c_h^{(2)} - c_h^{(1)})$, the difference between two mel-cepstrum spectral envelopes. By comparison of two different spectral envelopes, it is possible to express the deviation of each partial in the multiplicative form $r_h = 10\exp[\Delta C_h/20]$, and we call *conversion pattern* the set $\{r_h\}_{h=1}^{H}$ computed by the comparison of two spectral envelopes.

B. Spectral Conversion Functions

In this section, the parametric model for the spectral conversion functions and the parameter identification principles are presented. The conversion is expressed in terms of deviations of magnitudes, normalized with respect to the frame energy b_0, from the normalized magnitudes of a reference spectral envelope. The reference spectral envelope can be taken from one of the tones in the data set. If the tones in the data set are notes from a musical instrument, with a simple attack–sustain–release structure, we will always consider the sustain average spectral envelopes, where the average is generally taken on a sufficient number of frames of the sustained part of the tones. Once the spectrum conversion function has been identified, the reference tone can be seen as a source for the synthesis of tones with different pitch or intensity, and correct spectral behavior. Moreover, we are interested in keeping also the natural time variance of the source tone, as well as its attack–sustain–release structure. To this purpose, we make the simplifying hypothesis that the conversion function identified with respect to the sustained part of notes can be used to process every frame of the source note. We further make the following assumptions on the structure of the conversion function [43].

- Due to the changing nature of the spectrum with the pitch λ_0 of the tone, the conversion function is dependent on the pitch of the note. From the above consideration, the function will then be a map $\mathcal{F}:\mathbb{R}\rightarrow\mathbb{R}^H$, where H is the maximum number of partials in the SMS representation.
- We adopt the following parametric form for the generic conversion function:

$$\mathcal{F}(\lambda_0)=\begin{bmatrix}\mathcal{F}_1(\lambda_0)\\ \vdots \\ \mathcal{F}_H(\lambda_0)\end{bmatrix} \tag{7}$$

with

$$\mathcal{F}_j(\lambda_0)=\sum_{i=1}^{U} w_{i,j}\cdot G(\lambda_0;\mathbf{q}_i) \tag{8}$$

where $G(\lambda_0;\mathbf{q}_i)$ denotes a radial basis function with parameter vector $\mathbf{q}_i$, U is the number of radial basis units used, and $\mathbf{W}=\{w_{i,j}\}_{i=1\cdots U,j=1\cdots H}$ is a $U\times H$ matrix of output weights. The jth component of the conversion function, $\mathcal{F}_j(\lambda_0)$, describes how the magnitude of the jth partial will adapt with respect to the desired fundamental frequency λ_0.

The parametric model introduced in (8) is known in the literature as a *radial basis function network* (RBFN) and is a special case of feedforward neural networks which exhibit high performances in nonlinear curve-fitting (approximation) problems [44]. Curve fitting of data points is equivalent to finding the surface in a multidimensional space that provides a best fit to the training data, and generalization is the equivalent to the use of that surface to interpolate the data. Their interpolation properties have proven to be effective in signal processing tasks relating to our application, e.g., for voice spectral processing aimed at speaker conversion [45]. The radial functions $G(\cdot;\cdot)$ in (7) can be of various kinds. Typical choices are Gaussian, cubic, or sigmoidal functions. Here, a cubic form $G(\lambda;\mathbf{q})=(\|\lambda-\mu\|)^3$, with $\mathbf{q}=\{\mu\}$, is used. This may not be the best choice as for the final dimension and efficiency of the network, e.g., RBFNs with normalized Gaussian kernels (NRBF nets) can result in smaller and more compact networks. However, a simpler implementation with a reduced set of parameters per kernel and with essentially the same curve-fitting capabilities was preferred here.

1) Identification of the RBFN Parameters: As usually needed by the neural networks' learning procedures, the original data is organized in a training set. In our case, the pitch values of the training set notes are stored in the input training vector $\mathbf{T}_{in}=[\lambda_0^{(1)},\ldots,\lambda_0^{(N)}]$, where each component corresponds to a row of the output matrix $\mathbf{T}_{\text{out}}=\mathbf{R}$, with $\mathbf{R}=\{r_{i,j}\}_{i=1\cdots N,j=1\cdots H}$ being a matrix whose rows are the spectral envelope conversion patterns coming from the comparisons among the spectral envelopes from the source data and those from the target data. The way spectra are selected from both data sets depends on the final high-level transformation to be realized. In the next section, a practical case will be treated to exemplify the training set generation procedure. Here, we make the hypothesis that the training set has been computed with some strategy, and we summarize the RBFN parametric identification procedure. The centers μ of the radial basis functions are iteratively selected with the OLS algorithm [46], which places the desired number U of units (with $U\leq N$) in the positions that best explains the data. Once the radial units with centers $\mu_1,\ldots,\mu_U$ have been selected, the image of $\mathbf{T}_{in}$ through the radial basis layer can be computed as $\mathbf{G}=[\mathbf{G}_1\cdots\mathbf{G}_U]$, $\mathbf{G}_i=[G(\lambda_0^{(1)},\mu_i)\cdots G(\lambda_0^{(N)},\mu_i)]^T(i=1,\ldots,U)$. The problem of identifying the parameters $w_{i,j}$ of (8) can, thus, be given in the closed form $\mathbf{T}_{\text{out}}=\mathbf{GW}$, the LS solution of which is known to be $\mathbf{W}=\mathbf{T}_{\text{out}}\mathbf{G}^+$ with $\mathbf{G}^+$ pseudo-inverse of $\mathbf{G}$. As can be seen, this parametric model relies on a fast-learning algorithm, if compared to other well-known neural network models whose iterative learning algorithms are quite slow (e.g., backpropagation or gradient descent algorithms).

To summarize the principal motivations why we adopted the RBFN model, we emphasize that the RBFNs can learn from examples, have fast training procedures, and have good generalizing properties, meaning that if we use a training set of N tones having pitch values of $\lambda_0^{(1)}<\lambda_0^{(2)}<\cdots<\lambda_0^{(N)}$, the resulting conversion function will provide a coherent result in the whole interval $[\lambda_0^{(1)},\lambda_0^{(N)}]$.

2) Training Set Generation for the Control of Intensity: The spectral modeling method will be now used to realize intensity transformations which preserve the spectral identity of a musical instrument.

Let $\mathcal{D}(\lambda_0)$ be a conversion function identified following the procedure described. The synthesis formula is then

$$\bar{a}_h=\mathcal{D}_h(\lambda_0)\cdot a_h \tag{9}$$

where a_h is the magnitude of the hth partial of a source tone. Let us say now that, given a source tone with intensity level I_r (e.g., a note from the neutral performance), we are

interested in rising or lowering the original intensity. The analysis of the same notes taken from musical performances with different expressive intentions allows us to determine, for each note, the two tones having the minimum and the maximum intensity, here called, respectively, I_m and I_M. Say $\mathcal{D}_{I_M}(\lambda_0) = [\mathcal{D}_{I_M,1}(\lambda_0) \cdots \mathcal{D}_{I_M,H}(\lambda_0)]^T$ is the conversion function that allows to switch from I_r to I_M, and say $\mathcal{D}_{I_m}(\lambda_0) = [\mathcal{D}_{I_m,1}(\lambda_0) \cdots \mathcal{D}_{I_m,H}(\lambda_0)]^T$ is the conversion function that allows to switch from I_r to I_m. Note that $\mathcal{D}_{I_m}(\lambda_0)$ and $\mathcal{D}_{I_M}(\lambda_0)$ are still functions of fundamental frequency and not of intensity; we are in fact assuming that they turn the original note with intensity level I_r into a note with intensity level I_m or I_M, respectively. A simple way to produce a tone with intensity level between I_r and I_m or between I_r and I_M is, thus, to weight the effect of the conversion functions.[1] To this purpose, let us define $\mathcal{D}'_{I_{m(M)}}(\lambda, I) = \mathcal{D}_{I_{m(M)}}(\lambda) \cdot \alpha(I)$, where the function $\alpha(I)$, ranging from $1/\mathcal{D}_{I_{m(M)}}(\lambda_0)$, for $I = I_r$, to one, for $I = I_{m(M)}$, weights the effect of the conversion function. Then, the resynthesis formula that computes the new amplitudes for the intensity level $I \in [I_m, I_M]$ is

$$\begin{aligned} \bar{a}_h &= \mathcal{D}'_{I_m,h}(\lambda_0, I) \cdot a_h, \quad \text{if } I \leq I_r \\ \bar{a}_h &= \mathcal{D}'_{I_M,h}(\lambda_0, I) \cdot a_h, \quad \text{if } I > I_r. \end{aligned} \tag{10}$$

A logarithmic function for the function $\alpha(I)$ has shown to be suitable to perform an effective control on the range $[I_m, I_M]$.

An alternative solution, slightly more complicated, would have been to design the conversion function $\mathcal{D}(\lambda, I)$ adopting bivariate radial functions $G(\lambda, I; \mathbf{q})$. The design of the training set in this case would have required the selection, for each note in the performance, of a minimum number of sound frames with intensities spanning the range $[I_m, I_M]$.

As a final remark, we stress the fact that this spectral processing method is based on a learning-from-data approach and is highly dependent on the training data. As a consequence, with the present setup it is not possible to apply a given conversion function on a neutral performance which is not the one used during the training, and a different conversion function will be necessary for each new neutral performance to be processed.

V. Results and Applications

We applied the proposed methodology on a variety of digitally recorded monophonic melodies from classic and popular music pieces. Professional musicians were asked to perform excerpts from various musical scores, inspired by the following adjectives: light, heavy, soft, hard, bright, and dark. The neutral performance was also added and used as a reference in the acoustic analysis of the various interpretations. Uncoded adjectives in the musical field were deliberately chosen to give the performer the greatest possible freedom of expression. The recordings were carried out in three sessions, each session consisting of the seven different interpretations. The musician then chose the performances that, in his opinion, best corresponded to the proposed adjectives. This procedure is intended to minimize the influence that the order of execution might have on the performer. The performances were recorded at the the Centro di Sonologia Computazionale (CSC), University of Padova, Padova, Italy, in monophonic digital format at 16 b and 44.1 kHz. In total, 12 scores were considered, played with different instruments (violin, clarinet, piano, flute, voice, saxophone) and by various musicians (up to five for each melody). Only short melodies (between 10 and 20 s) were selected, allowing us to assume that the underlying process is stationary (the musician does not change the expressive content in such a short time window).

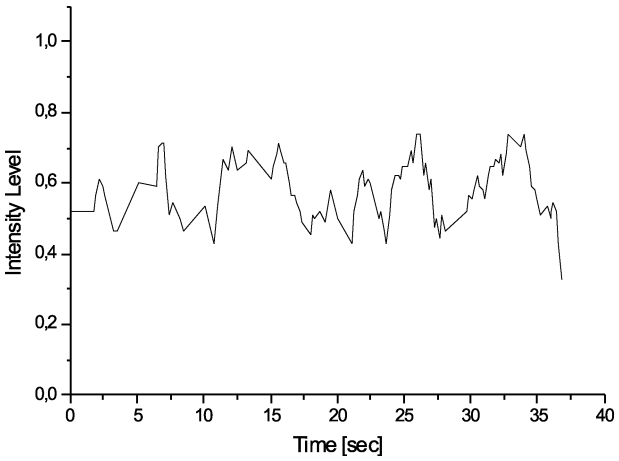


Fig. 8. Analysis: normalized intensity level of neutral performance of Mozart's sonata K545.

Semiautomatic acoustic analyses were then performed in order to estimate the expressive time- and timbre-related cues IOI, L, AD, I, EC, and BR. Fig. 8 shows the time evolution of one of the considered cues, the intensity level I, normalized in respect to maximum Key Velocity, for the neutral performance of an excerpt of Mozart's sonata K545 (piano solo).

Table 3 reports the values of the k and m parameters computed for Mozart's sonata K545, using the procedure described in Section III-C. For example, it can be noticed that the $k$value of the Legato (L) parameter is important for distinguishing *hard* ($k = 0$, 92 means quite staccato) and *soft* ($k = 1$, 43 means very legato) expressive intentions; considering the Intensity (I) parameter, *heavy* and *bright* have a very similar k value, but a different m value; that is, in *heavy* each note is played with a high Intensity ($m = 0$, 70), on the contrary *bright* is played with a high variance of Intensity ($m = 1$, 06).

The factor loadings obtained from factor analysis carried out on the results of the perceptual test are shown in Table 4. These factor loadings are assumed as coordinates of the expressive performances in the PPS. It can be noticed that factor 1 distinguishes *bright* (0.8) from *dark* (-0.8) and *heavy* (-0.75), factor 2 differentiates *hard* (0.6) and *heavy* (0.5) from *soft* (-0.7) and *light* (-0.5). From the

[1]Although one is not guaranteed on whether the model will reproduce or not the original spectral behavior of the instrument with respect to changes of the intensity level, this approach has proven to be satisfactory for the class of sounds considered here.

Table 3
Expressive Parameters Estimated From Performances of Mozart's Sonata K545

	IOI		L		DRA		I		EC		BR	
	k	m	k	m	k	m	k	m	k	m	k	m
Bright	0.87	0.98	0.68	0.95	0.76	0.96	1.07	1.06	0.90	0.79	1.13	0.80
Dark	1.05	1.01	1.09	1.02	0.93	1.12	0.87	1.05	1.12	1.06	0.67	0.72
Hard	0.95	0.86	0.92	1.06	0.73	0.84	1.06	0.76	0.98	1.04	1.17	0.96
Soft	1.03	1.08	1.43	0.89	1.06	1.02	0.92	1.03	1.18	1.11	0.74	1.05
Heavy	1.16	0.91	1.35	0.98	0.97	1.05	1.06	0.70	0.98	1.06	1.10	0.99
Light	0.90	0.96	0.79	1.12	1.13	1.10	0.97	1.12	0.84	0.84	0.82	1.03

Table 4
Factor Loadings Are Assumed as Coordinates of the Expressive Performances in the PPS

	Factor 1	Factor 2
Bright	0.8	0.1
Dark	-0.8	0.28
Hard	-0.4	0.6
Soft	-0.35	-0.7
Heavy	-0.75	0.5
Light	0.6	-0.5

data such as the ones in Table 3 and the positions in the PPS, the parameters of (3) are estimated. Then the model of expressiveness can be used to change interactively the expressive cues of the neutral performance by moving in the 2-D control space. The user is allowed to draw any trajectory which fits his own feeling of the changing of expressiveness as time evolves, morphing among expressive intentions (Fig. 9).

As an example, Fig. 10 shows the effect of the control action described by the trajectory (solid line) in Fig. 9 on the intensity level I (to be compared with the neutral intensity profile show in Fig. 8). It can be seen how the intensity level varies according to the trajectory; for instance, hard and heavy intentions are played louder than the soft one. In fact, from Table 3, the k values are 1.06 (hard), 1.06 (heavy), and 0.92 (soft). On the other hand, we can observe a much wider range of variation for light performance $(m = 1.12)$ than for heavy performance $(m = 0.70)$. The new intensity level curve is used, in its turn, to control the audio processing engine in the final rendering step.

As a further example, an excerpt from the Corelli's sonata op. V is considered (Fig. 11). Figs. 12–14 show the energy envelope and the pitch contour of the original neutral, heavy, and soft performances (violin solo). The model is used to obtain a smooth transition from heavy to soft (dashed trajectory

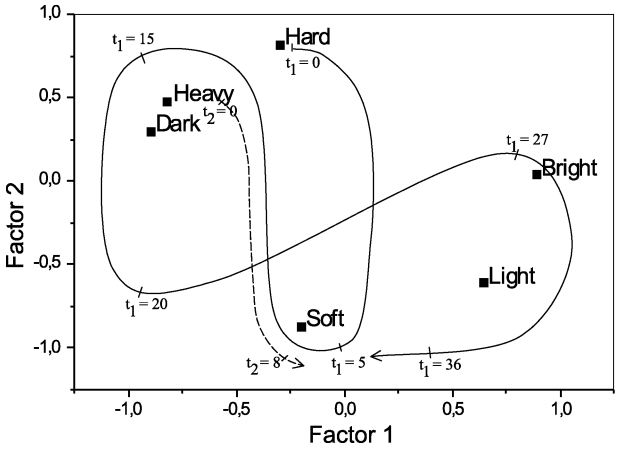


Fig. 9. Control: trajectories in the PPS space corresponding to different time evolution of the expressive intention of the performance. Solid line: the trajectory used on the Mozart theme; dashed line: trajectory used on the Corelli theme.

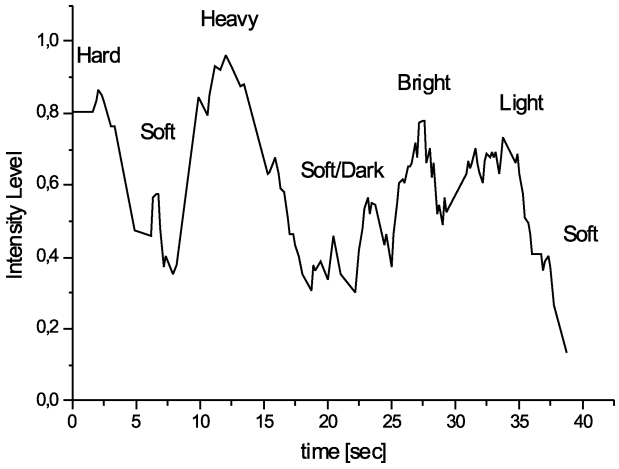


Fig. 10. Synthesis: normalized intensity level corresponding to the trajectory in Fig. 9.

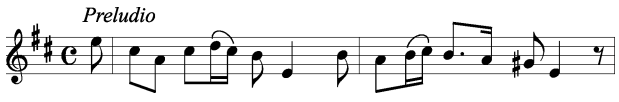


Fig. 11. Score of the theme of Corelli's sonata op. V.

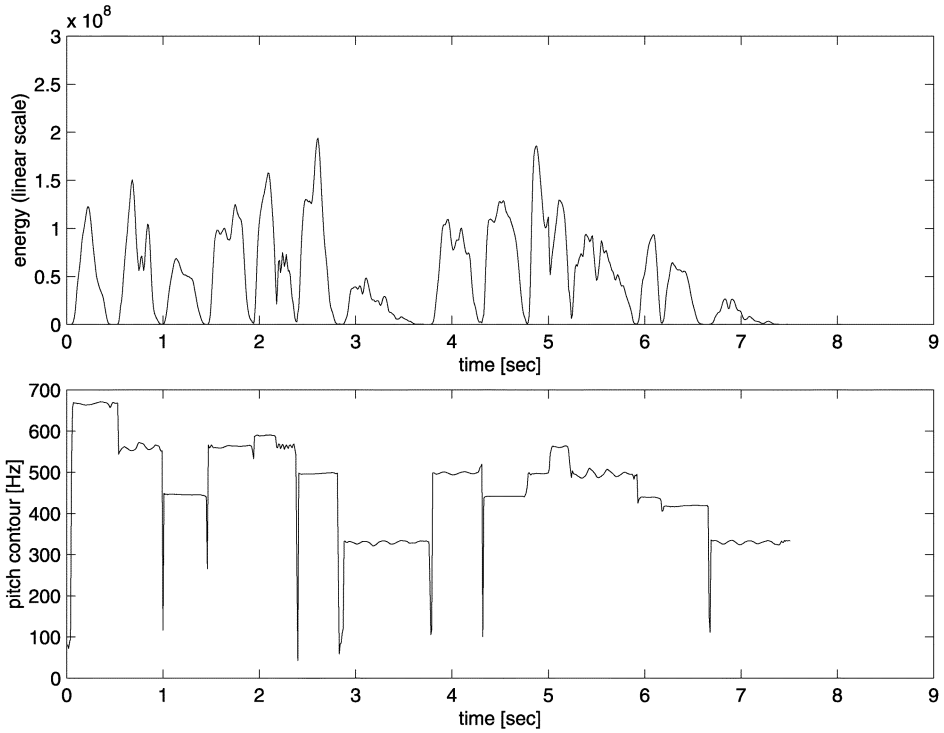


Fig. 12. Analysis: energy envelope and pitch contour of neutral performance of Corelli's sonata op. V.

in Fig. 9) by applying the appropriate transformations on the sinusoidal representation of the neutral version. The result of this transformation is shown in Fig. 15. It can be noticed that the energy envelope changes from high to low values, according to the original performances (heavy and soft). The pitch contour shows the different behavior of the IOI parameter: the soft performance $(k = 1.03)$ is played faster than the heavy performance $(k = 1.16)$. This behavior is preserved in our synthesis example.

We developed an application, released as an applet, for the fruition of fairytales in a remote multimedia environment [38]. In these kinds of applications, an expressive identity can be assigned to each character in the tale and to the different multimedia objects of the virtual environment. Starting from the storyboard of the tale, the different expressive intentions are located in a control spaces defined for the specific contexts of the tale. By suitable interpolation of the expressive parameters, the expressive content of audio is gradually modified in real time with respect to the position and movements of the mouse pointer, using the model describe above.

This application allows a strong interaction between the user and the audiovisual events. Moreover, the possibility of having a smoothly varying musical comment augments the user emotional involvement, in comparison with the participation reachable using rigid concatenation of different sound comments. Sound examples can be found on our Web site [47].

VI. Assessment

A perceptive test was realized to validate the system. A categorical approach was considered. Following the categorical approach [28], we intend to verify if performances synthesized according to the adjectives used in our experiment, are recognized. The main objective of this test is to see if a "static" (i.e., not time-varying) intention can be understood by listeners and if the system can convey the correct expression.

According to Juslin [48], forced-choice judgments and free-labeling judgments give similar results when listeners attempt to decode a performer's intended emotional expression. Therefore, it was considered sufficient to make a forced-choice listening test to assess the efficacy of the emotional communication. A detailed description of the procedure and of the statistical analyses can be found in [49]. In the following, some results are summarized.

A. Material

We synthesized different performances using our model. Given a score and a *neutral* performance, we obtain the five different interpretations from the control space, i.e., *bright*, *hard*, *light*, *soft*, and *heavy*. We did not consider the *dark* one, because in our previous experiments we noticed that it was confused with the *heavy* one, as can be seen in Fig. 9.

It was important to test the system with different scores to understand how high is the correlation between the inherent structure of the piece and the expressive recognition. Three classical pieces for piano with different sonological characteristics were selected in this experiment: "Sonatina in sol" by L. van Beethoven, "Valzer no. 7 op. 64" by F. Chopin, and K545 by W. A. Mozart.

The listeners' panel was composed of 30 subjects: 15 *experts* (musicians and/or conservatory graduated) and 15 *commons* (without any particular musical knowledge). No

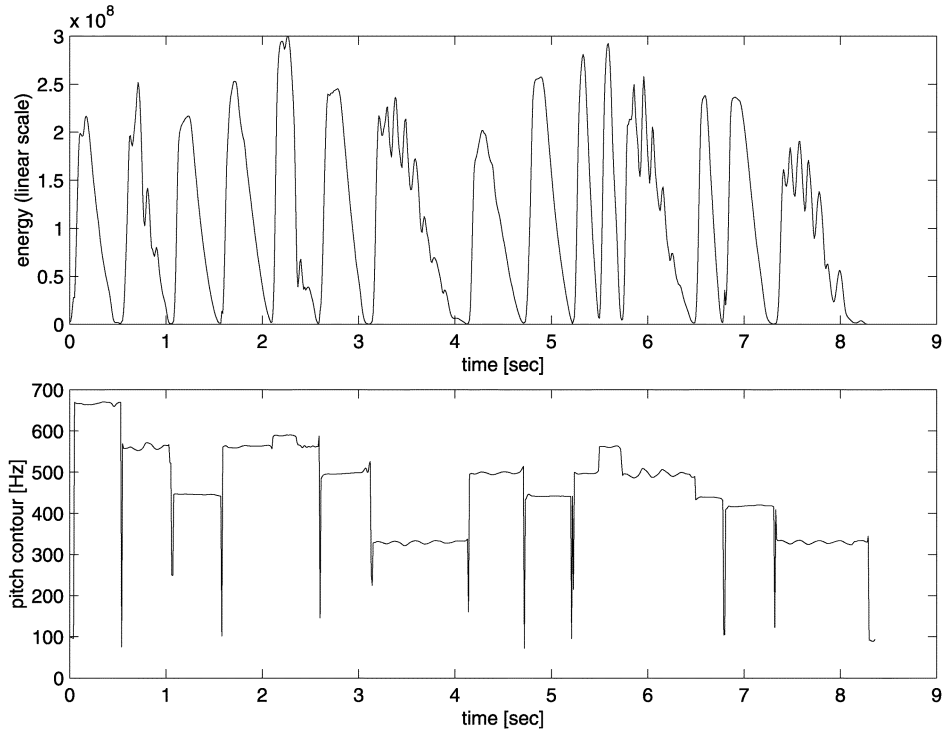


Fig. 13. Analysis: energy envelope and pitch contour of heavy performance of Corelli's sonata op. V.

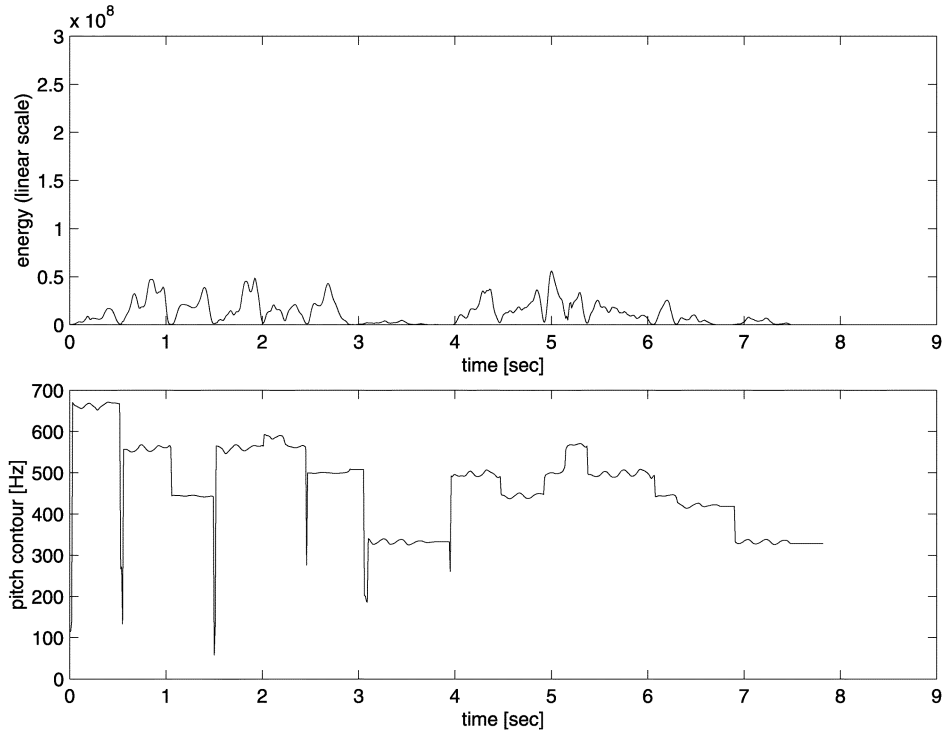


Fig. 14. Analysis: energy envelope and pitch contour of soft performance of Corelli's sonata op. V.

restrictions related to formal training in music listening were used in recruiting subjects. None of the subjects reported having hearing impairments.

B. Procedure

The stimuli were played by a PC. The subjects listened to the stimuli through headphones at a comfortable loudness level. The listeners were allowed to listen the stimuli as many time as they needed, in any order. Assessors were asked to evaluate the grade of *brightness*, *hardness*, *lightness*, *softness*, and *heaviness* of all performances on a graduated scale (0 to 100). Statistical analyses were then conducted in order to determine if the intended expressive intentions were correctly recognized.

C. Data Analysis

Table 5 summarizes the assessors' evaluation. The ANOVA test on the subject's response always yielded a

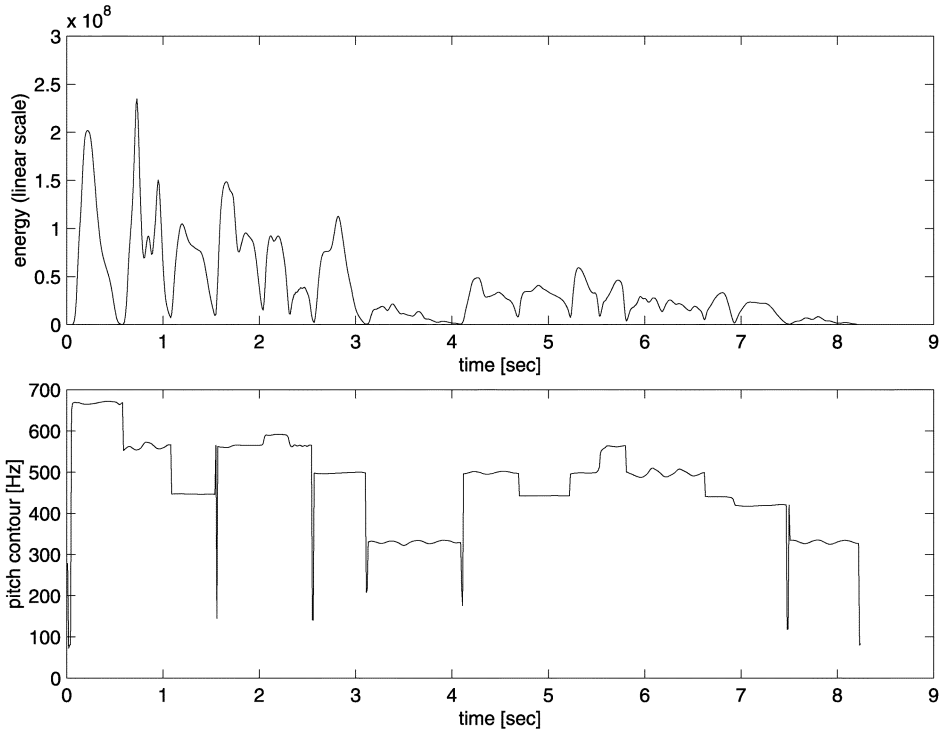


Fig. 15. Synthesis (loop of the 16-note excerpt): energy envelope and pitch contour of an expressive morphing. The expressive intention changes smoothly from heavy to soft. The final rendering is the result of the audio transformations controlled by the model and performed on the neutral performance.

Table 5
Assessors' Evaluation Average (From 0 to 100)

	Sonatina						Valzer						K545					
	B	Hr	L	S	Hv	N	B	Hr	L	S	Hv	N	B	Hr	L	S	Hv	N
B	**73.2**	59.3	51.7	50.7	43.2	**55.3**	**69.8**	50.1	36.6	42.2	35.5	44.7	**71.6**	67.3	35.2	41.7	46.0	33.8
Hr	59.1	**79.2**	28.1	32.6	61.7	33.4	49.8	**74.7**	18.0	25.6	65.2	37.9	56.8	**68.4**	22.5	15.7	80.3	17.1
L	41.9	17.7	**66.6**	57.4	17.6	54.2	42.0	21.1	66.1	57.1	17.1	47.5	34.2	27.6	**75.0**	75.5	12.9	69.3
S	24.1	13.1	59.9	**64.6**	24.7	55.1	29.8	22.7	**72.5**	**66.0**	22.5	**53.0**	22.6	27.7	65.3	**77.2**	13.9	**72.7**
Hv	43.4	69.8	18.2	25.1	**69.8**	25.2	39.0	65.7	22.9	24.7	**78.1**	37.7	37.3	52.8	15.2	15.0	**82.8**	17.8

Rows represent the evaluation labels, and columns show the different stimuli. Legend: B=Bright, Hr=Hard, L=Light, S=Soft, Hv=Heavy, N=Neutral.

p-index less than 0.001: the p values indicate that one or more populations' means differ quite significantly from the others. From data analyses, such as observation of the means and standard deviations, we notice that generally, for a given interpretation, the correct expression obtains the highest mark. One exception is the Valzer, where the *light* interpretation is recognized as *soft*—with a very slight advantage. Moreover, with K545, *heavy* performance was judged near to *hard* expressive intention (82.8 versus 80.3) whereas *hard* performance near to *bright* (68.4.8 versus 67.3) expressive intention, suggesting a slight confusion between these samples.

It is also interesting to note that listeners, in evaluating the neutral performance, did not spread uniformly their evaluation among the adjectives. Even if all the expressions are quite well balanced, we have a predominance of *light* and *soft*. The *bright* expression is also quite high but no more than the average brightness of all performances.

A high correlation between *hard* and *heavy* and between *light* and *soft* can be noticed. Those expressions are well individuated in two groups. On the other hand, *bright* seems to be more complicated to highlight. An exhaustive statistical analysis of the data is discussed in [49], as well as the description of a test carried out by means of a dimensional approach. It is important to notice that the factor analysis returns our PPS. Automatic expressive performances synthesized by the system give a good modeling of expressive performance realized by human performers.

VII. Conclusion

We presented a system to modify the expressive content of a recorded performance in a gradual way both at the

symbolic and the signal levels. To this purpose, our model applies a smooth morphing among different expressive intentions in music performances, adapting the expressive character of the audio/music/sound to the user's desires. Morphing can be realized with a wide range of graduality (from abrupt to very smooth), allowing to adapt the system to different situations. The analyses of many performances allowed us to design a multilevel representation, robust with respect to morphing and rendering of different expressive intentions. The sound rendering is obtained by interfacing the expressiveness model with a dedicated postprocessing environment, which allows for the transformation of the event cues. The processing is based on the organized control of basic audio effects. Among the basic effects used, an original method for the spectral processing of audio is introduced. The system provided interesting results for both the understanding and focusing of topics related to the communication of expressiveness, and the evaluation of new paradigms of interaction in the fruition of multimedia systems.

REFERENCES

[1] A. Gabrielsson, "Expressive intentions and performance," in *Music and the Mind Machine*, R. Steinberg, Ed. Berlin, Germany: Springer-Verlag, 1995, pp. 35–47.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Felenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Mag.*, vol. 18, pp. 32–80, Jan. 2001.

[3] C. Palmer, "Music performance," *Annu. Rev. Psychol.*, vol. 48, pp. 115–138, 1997.

[4] B. H. Repp, "Diversity and commonality in music performance: an analysis of timing microstructure in Schumann's 'Träumerei'," *J. Acoust. Soc. Amer.*, vol. 92, pp. 2546–2568, 1992.

[5] M. Clynes, "Microstructural musical linguistics: composer's pulses are liked best by the best musicians," *Cognition: Int. J. Cogn. Sci.*, vol. 55, pp. 269–310, 1995.

[6] B. H. Repp, "Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists," *J. Acoust. Soc. Amer.*, vol. 88, pp. 622–641, 1990.

[7] A. Gabrielsson, "Music performance, the psychology of music," in *The Psychology of Music*, 2nd ed, D. Deutsch, Ed. New York: Academic, 1997, pp. 35–47.

[8] N. P. Todd, "A model of expressive timing in tonal music," *Music Perception*, vol. 3, no. 1, pp. 33–58, 1985.

[9] ——, "The dynamics of dynamics: a model of musical expression," *J. Acoust. Soc. Amer.*, vol. 91, no. 6, pp. 3540–3550, 1992.

[10] ——, "The kinematics of musical expression," *J. Acoust. Soc. Amer.*, vol. 97, no. 3, pp. 1940–1949, 1995.

[11] G. De Poli, A. Rodá, and A. Vidolin, "Note-by-note analysis of the influence of expressive intentions and musical structure in violin performance," *J. New Music Res.*, vol. 27, no. 3, pp. 293–321, 1998.

[12] M. Clynes, "Some guidelines for the synthesis and testing of pulse microstructure in relation to musical meaning," *Music Perception*, vol. 7, no. 4, pp. 403–422, 1990.

[13] A. Friberg, L. Frydén, L. G. Bodin, and J. Sundberg, "Performance rules for computer-controlled contemporary keyboard music," *Comput. Music J.*, vol. 15, no. 2, pp. 49–55, 1991.

[14] J. Sundberg, "How can music be expressive?," *Speech Commun.*, vol. 13, pp. 239–253, 1993.

[15] A. Friberg, V. Colombo, L. Frydén, and J. Sundberg, "Generating musical performances with director musices," *Comput. Music J.*, vol. 24, no. 3, pp. 23–29, 2000.

[16] G. De Poli, L. Irone, and A. Vidolin, "Music score interpretation using a multilevel knowledge base," *Interface (J. New Music Res.)*, vol. 19, pp. 137–146, 1990.

[17] G. Widmer, "Learning expressive performance: the structure-level approach," *J. New Music Res.*, vol. 25, no. 2, pp. 179–205, 1996.

[18] ——, "Large-scale induction of expressive performance rules: first qualitative results," in *Proc. 2000 Int. Comp. Music Conf.*, Berlin, Germany, 2000, pp. 344–347.

[19] H. Katayose and S. Inokuchi, "Learning performance rules in a music interpretation system," *Comput. Humanities*, vol. 27, pp. 31–40, 1993.

[20] J. L. Arcos and R. L. de Màntaras, "An interactive case-based reasoning approach for generating expressive music," *Appl. Intell.*, vol. 14, no. 1, pp. 115–129, 2001.

[21] T. Suzuki, T. Tokunaga, and H. Tanaka, "A case based approach to the generation of musical expression," in *Proc. 1999 IJCAI*, pp. 642–648.

[22] R. Bresin, "Artificial neural networks based models for automatic performance of musical scores," *J. New Music Res.*, vol. 27, no. 3, pp. 239–270, 1998.

[23] R. Bresin, G. D. Poli, and R. Ghetta, "A fuzzy approach to performance rules," in *Proc. XI Colloquium on Musical Informatics (CIM-95)*, pp. 163–168.

[24] ——, "Fuzzy performance rules," in *Proc. KTH Symp. Grammars for Music Performance*, Stockholm, Sweden, 1995, pp. 15–36.

[25] O. Ishikawa, Y. Aono, H. Katayose, and S. Inokuchi, "Extraction of musical performance rule using a modified algorithm of multiple regression analysis," in *Proc. KTH Symp. Grammars for Music Performance"*, 2000, pp. 348–351.

[26] J. L. Arcos, R. L. de Màntaras, and X. Serra, "Saxex: a case-based reasoning system for generating expressive musical performances," *J. New Music Res.*, pp. 194–210, Sept. 1998.

[27] R. Bresin and A. Friberg, "Emotional coloring of computer controlled music performance," *Comput. Music J.*, vol. 24, no. 4, pp. 44–62, 2000.

[28] P. Juslin and J. Sloboda, Eds., *Music and Emotion: Theory and Research*. Oxford, U.K.: Oxford Univ. Press, 2001.

[29] S. Canazza, G. De Poli, and A. Vidolin, "Perceptual analysis of the musical expressive intention in a clarinet performance," in *Music, Gestalt, and Computing*, M. Leman, Ed. Berlin, Germany: Springer-Verlag, 1997, pp. 441–450.

[30] S. Canazza, A. Rodà, and N. Orio, "A parametric model of expressiveness in musical performance based on perceptual and acoustic analyses," in *Proc. ICMC99 Conf.*, 1999, pp. 379–382.

[31] C. Roads, *The Computer Music Tutorial*. Cambridge, MA: MIT Press, 1996.

[32] B. Vercoe, W. Gardner, and E. Scheirer, "Structured audio: creation, transmission, and rendering of parametric sound representation," *Proc. IEEE*, vol. 86, pp. 922–940, May 1998.

[33] W. J. Pielemeier, G. H. Wakefield, and M. H. Simoni, "Time-frequency analysis of musical signals," *Proc. IEEE*, vol. 84, pp. 1216–1230, Sept. 1996.

[34] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, Aug. 1986.

[35] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*. Lisse, The Netherlands: Swets & Zeitlinger, 1997, pp. 497–510.

[36] Spectral modeling synthesis software [Online]. Available: http://www.iua.upf.es/ sms/

[37] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modification on musical timbres," *J. Acoust. Soc. Am.*, vol. 65, no. 5, pp. 1493–1500, 1978.

[38] S. Canazza, G. De Poli, C. Drioli, A. Rodà, and A. Vidolin, "Audio morphing different expressive intentions for multimedia systems," *IEEE Multimedia*, vol. 7, pp. 79–84, July–Sept. 2000.

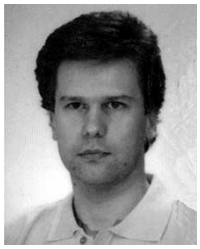
[39] S. Canazza and N. Orio, "The communication of emotions in jazz music: a study on piano and saxophone performances," *Gen. Psychol. (Special Issue on Musical Behavior and Cognition)*, vol. 3/4, pp. 261–276, Mar. 1999.

[40] S. Canazza, G. De Poli, and A. Vidolin, "Perceptual analysis of the musical expressive intention in a clarinet performance," in *Proc. IV Int. Symp. Systematic and Comparative Musicology*, 1996, pp. 31–37.

[41] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Amer.*, vol. 68, no. 5, pp. 1523–1525, November 1980.

[42] O. Cappé, J. Laroche, and E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points," in *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1995, pp. 213–216.

[43] C. Drioli, "Radial basis function networks for conversion of sound spectra," *J. Appl. Signal Process.*, vol. 2001, no. 1, pp. 36–44, 2001.
[44] S. Haykin, *Neural Networks. A Comprehensive Foundation*. New York: Macmillan, 1994.
[45] N. Iwahashi and Y. Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks," *Speech Commun.*, vol. 16, no. 2, pp. 139–151, 1995.
[46] S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis functions networks," *IEEE Trans. Neural Networks*, vol. 2, pp. 302–309, Mar. 1991.
[47] Algorithms and models for sound synthesis: Audio examples [Online]. Available: http://www.dei.unipd.it/ricerca/csc/research_groups/expressive_performance_examples.html
[48] P. Juslin, "Can results from studies of perceived expression in musical performances be generalized across response formats?," *Psychomusicology*, vol. 16, no. 3, pp. 77–101, 1997.
[49] D. Cirotteau, S. Canazza, G. De Poli, and A. Rodà, "Analysis of expressive contents in synthesized music: categorical and dimensional approach," presented at the 5th International Workshop Gesture and Sign Language Based Human-Computer Interaction, Genova, Italy, 2003.



Sergio Canazza received the Laurea degree in electronic engineering from the University of Padova, Padova, Italy, in 1994.

He is currently Assistant Professor at the Department of Scienze Storiche e Documentarie, University of Udine (Polo di Gorizia), Gorizia, Italy, where he teaches classes on informatics and digital signal processing for music. He is also a Staff Member of Centro di Sonologia Computazionale (CSC), University of Padova, and Invited Professor in musical informatics at the Conservatory of Music, Trieste, Italy. His main research interests are in preservation and restoration of audio documents, models for expressiveness in music, multimedia systems and human–computer interaction.

Giovanni De Poli (Member, IEEE) received the Laurea degree in electronic engineering from the University of Padova, Padova, Italy.

He is currently an Associate Professor of computer science at the Department of Electronics and Computer Science, University of Padova, Padova, Italy, where he teaches classes on the fundamentals of informatics and processing systems for music. He is also the Director of the Centro di Sonologia Computazionale (CSC), University of Padova. He is author of several scientific international publications, served on the scientific committees of international conferences, and is Associate Editor of the *Journal of New Music Research*. He is owner of patents on digital music instruments. His main research interests are in algorithms for sound synthesis and analysis, models for expressiveness in music, multimedia systems and human-computer interaction, and preservation and restoration of audio documents. He in involved with several European research projects: COST G6—Digital Audio Effects (National Coordinator); MEGA IST Project—Multisensory Expressive Gesture Applications (Local Coordinator); MOSART IHP Network (Local Coordinator). Systems and research developed in his lab have been exploited in collaboration with the digital musical intruments industry (GeneralMusic, Rimini, Italy).

Dr. De Poli is a Member of the Executive Committee (ExCom) of the IEEE Computer Society Technical Committee on Computer Generated Music, a Member of the Board of Directors of Associazione Italiana di Informatica Musicale (AIMI), a Member of the Board of Directors of Centro Interuniversitario di Acustica e Ricerca Musicale (CIARM), and a Member of the Scientific Committee of the Association pour la Création et la recherche sur les Outils d'Expression (ACROE, Institut National Politechnique Grenoble).

Carlo Drioli (Member, IEEE) received the Laurea degree in electronic engineering and the Ph.D. degree in electronic and telecommunication engineering from the University of Padova, Padova, Italy, in 1996 and 2003, respectively.

Since 1996, he has been a Researcher with the Centro di Sonologia Computazionale (CSC), University of Padova, in the field of sound and voice analysis and processing. From 2001 to 2002, he was also a Visiting Researcher at the Royal Institute of Technology (KTH), Stockholm, Sweden, with the support of the European Community through a Marie Curie Fellowship. He is also currently a Researcher with the Department of Phonetics and Dialectology of the Institute of Cognitive Sciences and Technology, Italian National Research Council (ISTC-CNR), Padova. His current research interests are in the fields of signal processing, sound and voice coding by means of physical modeling, speech synthesis, and neural networks applied to speech and audio.

Antonio Rodà was born in Verona, Italy, in 1971. He received the bachelor's degree from the National Conservatory of Music in Padova, Italy, in 1994 and the Laurea degree in electrical engineering from the University of Padova, Padova, in 1996.

Since 1997, he has been working on music performance analysis at the Centro di Sonologia Computazionale (CSC), University of Padova. Since 1999, he has also been a Teacher in musical informatics at the Conservatory of Music, Trieste, Italy.

Alvise Vidolin was born in Padova, Italy, in 1949. He received the Laurea degree in electronic engineering from the University of Padova, Padova, Italy.

He is Cofounder and Staff Member with the Centro di Sonologia Computazionale (CSC), University of Padova, Padova, where he is also teaching computer music as an Invited Professor and conducting his research activity in the field of computer-assisted composition and real-time performance. He is also currently teaching electronic music at B. Marcello Conservatory of Music, Venezia, Italy. Since 1977, he has often worked with La Biennale di Venezia, Venezia. He has also given his services to several important Italian and foreign institutions, and he worked with several composers, such as C. Ambrosini, G. Battistelli, L. Berio, A. Guarnieri, L. Nono, and S. Sciarrino, on the electronic realization and performance of their works.

Prof. Vidolin is Cofounder of the Italian Computer Music Association (AIMI), where he was President from 1988 to 1990 and is currently a Member of the Board of Birectors. He is also a Member of the Scientific Committee of the Luigi Nono Archive.