SYMBOLIC AND AUDIO PROCESSING TO CHANGE THE EXPRESSIVE INTENTION OF A RECORDED MUSIC PERFORMANCE

Sergio Canazza *Giovanni De Poli* *Riccardo Di Federico* *Carlo Drioli* *Antonio Rodá*

Centro di Sonologia Computazionale (CSC-DEI), University of Padua, Italy
canazza@dei.unipd.it

ABSTRACT

A framework for real-time expressive modification of audio musical performances is presented. An expressiveness model compute the deviations of the musical parameters which are relevant in terms of control of the expressive intention. The modifications are then realized by the integration of the model with a sound processing engine.

1. INTRODUCTION

In multimedia systems the musical media is essentially based on pre-recorded audio, which can be played by the listener without modification. Recently, a growing interest has been demonstrated in the field of expressiveness control of audio performances [1],[2] and mathematical and perceptual models have been proposed which compute the deviations of some musical parameters in order to add the desired expressive intentions to a neutral performance [3]. These models usually decides their action on the base of a symbolic description of the performance and give the result in term of deviations of parameters of the same description. In this work we propose to perform the expressive musical transformations on digitally recorded performances. The rendering is performed by means of an audio effects framework integrated with the expressiveness model. In this way, the system performs expressive manipulations and high quality signal processing of sound in an organized way, interpreting the symbolic deviations related to expressiveness in terms of sound transformations.

2. ARCHITECTURE

The functional structure of the system is reproduced in Figure 1. The input of the expressiveness model is composed by a description of a neutral musical performance (played without any expressive intention), the nominal score of the performance, and a control of the expressive intention desired by the user. The neutral performance is usually a digitally recorded monophonic part. A symbolic description of the audio performance is derived by analysis of the recorded signal, containing all the musical parameters needed by the model (note onset and offset, time position of the attack, sustain and release portions, information on higher level attributes like vibrato or tremolo). The expressiveness model acts on the symbolic level, computing the deviations of all musical parameters involved in the transformation and finally driving the audio processing engine. The processing engine is made of a given number of basic audio effects. All the sound transformations addressed by the model are realized by means of a frame-rate parametric control of the audio effects.

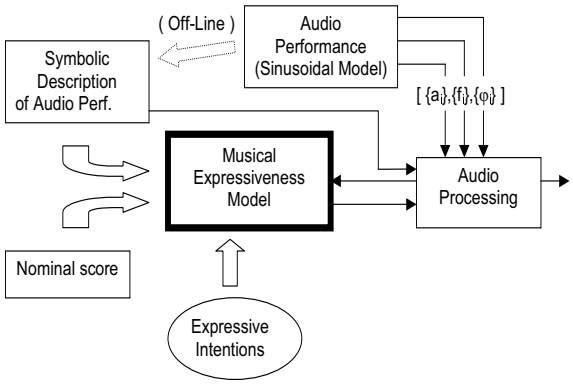

Figure 1: System architecture

3. THE MODEL

The expressiveness model was obtained starting from perceptual and acoustic analysis carried out on several recorded performances with different expressive intentions. Factor analysis on the listeners judgements showed that the performances can be arranged in a bi-dimensional space (Perceptual Parametric Space, PPS), in which the axis are correlated mainly with the kinetics of music and with the energy of sound [4] (see Figure 2).

When the user moves in the space, the model computes the changes of a set of musical parameters, that are next summarized. With reference to Figure 4, IOI is the *Inter Onset Interval* between the current and the successive note, DR is the duration of the current note, DRA is the duration of the attack, the *Intensity* I is the mean envelope energy, the *Envelope Centroid* EC is the time location of the energy envelope center of mass. The definition of *Legato* is

This work was supported by *TELECOM ITALIA* under the research contract *Cantieri Multimediali*

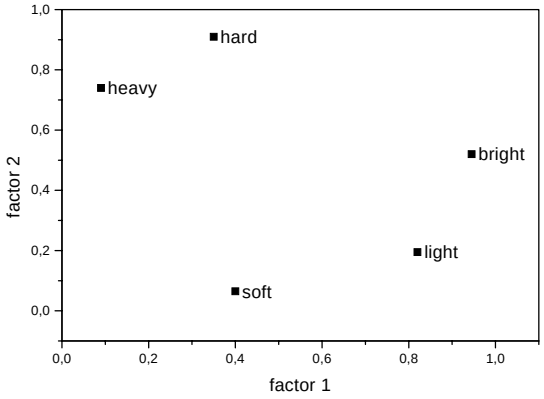

Figure 2: Factor analysis using performance as variables. The first factor (75.2%) is correlated with the kinetics of the music; the second factor is correlated with the energy of the sound.

assumed to be $L = DR/IOI$. If a pause is written in the nominal score, the IOI of the note is estimated with the the formula $IOI = IOI' DR_n/(DR_n + DR_{n,p})$, where IOI' is the interval between the current note and the next note (including the pause), DR_n is the nominal duration of the current note and $DR_{n,p}$ is the nominal duration of the pause. With this definition, *Legato* between two notes gives $L > 1$, while different degrees of staccato gives $L < 1$ and are determined by the *micropause* between the two notes.

For each expressive intention, the deviations of the acoustic parameters are calculated using the following equation

$$\delta P(n) = \frac{P_{out}(n)}{P_{in}(n)} = k_p \frac{\bar{P}_{in}}{P_{in}(n)} + m_p \left(1 - \frac{\bar{P}_{in}}{P_{in}(n)}\right) \quad (1)$$

where n is the cardinal number of the nth note of the score; P is the generic parameter modified by the model; the subscript in indicates that the value of the parameter P is calculated from the inputs. The subscript out indicates the value of the expressive performance, that is the output of the system; $\bar{P}_{in}$ is the mean of the values measured in the input performance (for the parameter P); k_P and m_P are coefficients that carry out two different transformations of the parameter P; The first one performs a translation and the second performs a stretching of the values.

For each parameter P, starting from PPS (see paragraph above) the k and m coefficients are calculated by means of the equations

$$\begin{aligned} k(x,y) &= a_0 + a_1 \cdot x + a_2 \cdot y \\ m(x,y) &= b_0 + b_1 \cdot x + b_2 \cdot y \end{aligned} \quad (2)$$

	N	Ha	S	He	L	B	D
δIOI		$>$			$<<$	$>$	$>>$
δL				$>$	$<$	$<<$	
δDRA		$<<$			$>$	$<$	$<$
δI		$>>$	$<$	$>$	$<$		$<<$
δEC			$>>$		$<$		$>$

Figure 3: Qualitative parameters changes for Neutral, Hard, Soft, Light, Bright and Dark performances. The deviations, referred to 1, are expressed by means of multiplicative factors, so that $\delta DRA = 1.2$ describes a 20% lengthening of the attack

where x and y are the coordinates of the PPS; a_i and b_i are coefficients that represents the relation between the transformations of the acoustic parameters and the axes of the PPS. These coefficients are estimated carrying out a multiple regression on the recorded expressive performances.

Figure 3 summarizes the qualitative rules of the model (here, δ stands for a deviation of the corresponding musical parameter).

Each time a different point (user's expressive intention) is selected in the PPS, the equation 2 computes the coefficients m and k for each acoustic parameter, . Basically, these coefficients are equivalent to instructions like "play louder", "play faster" or "amplify rubato", "play a tempo", and so on. Then, the equation (1) computes the expressive deviation of the parameter P, which is the input of the rendering step of the system. Generally, during the same performance, a trajectory that moves from a region to another one of the PPS can be drawn. The k and m coefficients, in that case, are functions of time ($k(t)$, $m(t)$) and the performance will be characterized by changeable expressive features.

More details on the research aspects and on the realization of the expressiveness model can be found in [3].

4. THE SOUND ANALYSIS AND PROCESSING FRAMEWORK

Audio processing is often aimed to change the acoustical or perceptual characteristics of the sounds. To modify the expressive intention of a recorded performance, however, an organized control of the audio effects is necessary which provides a task-specific high-level control. This approach is found, for example, in [1], where a case-based reasoning system is responsible for the evaluation of transformations realized in terms of basic audio effects.

Our expressiveness model is suited 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.

The audio processing techniques involved in this stage

are based on a *Sinusoidal* model of the input signal, which in the last decade showed to be one of the most flexible tools for sound analysis and resynthesis [5]. The estimation of frequencies, amplitudes and phases of sinusoidal components involves a Short Time Fourier Transform (STFT) analysis, followed by a peak picking algorithm. The result of this procedure is the identification of time evolving parameters.

Resynthesis of sound is performed by inverse FFT to synthesize overlapping output sound frames. This approach presents a much higher computational efficiency if compared to the classical additive synthesis, and therefore is preferred for real-time applications.

All the principal sound effects are obtained by control on 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 non-integer step. *Pitch shift* is obtained by scaling the frequencies of the harmonics and by preserving formants with spectral envelope interpolation. *Intensity* control can be made by scaling of partial amplitudes. However, an equal scale factor for each harmonic would result in a distortion of the instrument characteristics. To preserve the natural relation between intensity and spectral envelope, a map $\mathcal{F} : \mathbb{R} \rightarrow \mathbb{R}^H$, where H is the number of harmonics, is identified for each pitch value in the musical score, in order to have different maps for notes with different pitch. The map, based on a neural network, can learn the spectral behaviour of the sound from a set of sound examples [6].

Finally, some other higher level sound processing, such as vibrato and tremolo control, is performed with specific algorithms, as detailed in [2]. This processing proved to be necessary whenever time stretching must be applied to sound segments originally characterized by vibrato or tremolo, in order to avoid the alteration of the vibrato and tremolo rate. This procedure relies on a first step which identifies and subtract the feature on the sinusoidal representation, and on a second step which applies back the feature with the correct parameters on the stretched segment.

Among its properties, the sinusoidal representation of the sound offers a powerful analysis framework for the extraction of musical features at various level. A symbolic description of the performance is created with an off-line procedure, containing for each note all the musical parameters involved in the control of expressiveness.

The rendering of the deviations computed by the model may imply the use of just one of the basic sound effects seen above, or the combination of two or more of these effects, with the following general rules:

Local Tempo: time stretching is applied to each note. The analysis conducted on real performances with different expressive intentions, revealed that for strings and winds the

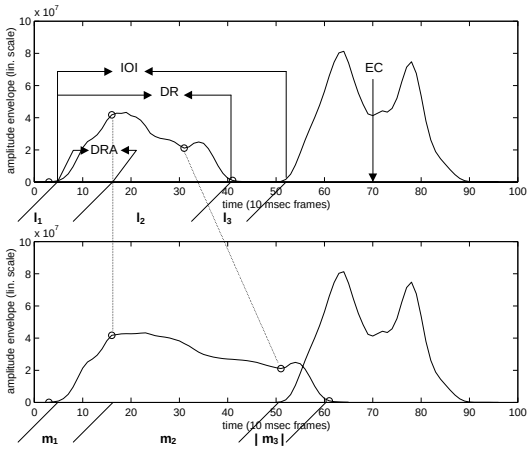

Figure 4: Musical parameters involved in the control of expressiveness. A modification of the *Legato* parameter is also shown

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 δDRA indicated by the model. The computation of the time stretch control on the note relies on the cumulative information given by the δDRA and δIOI factors, and on the δDR 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 behaviour, and a 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 δDR of the note, while is $L' = L\ \delta L = (DR\ \delta DR)/(IOI\ \delta IOI)$ and $\delta L = \delta DR/\delta IOI$. This time stretching action must be added to the one considered for the *Local Tempo* variation, as we can see in detail. Three different time stretching zones are distinguished within each note (with reference to Figure 4): attack, sustain and release, micropause. The time-stretching deviations must satisfy the relations $m_1 = l_1 \cdot \delta DRA$, $m_1 + m_2 = (l_1 + l_2) \cdot \delta DR$, $m_1 + m_2 + m_3 = (l_1 + l_2 + l_3) \cdot \delta IOI$, and each region will be processed with a time stretch factor K_i ($i = 1, 2, 3$) computed from the above relations: $K_1 = m_1/l_1 = \delta DRA$,

$K_2 = m_2/l_2 = [(l_1 + l_2)\delta L \cdot \delta IOI - K_1 l_1]/[l_2]$ and $K_3 = m_3/l_3 = -(l_1+l_2)\delta L \cdot \delta IOI + (l_1+l_2+l_3) \cdot \delta IOI$. Note that the coefficient K_3 can be negative if an overlap occurs due to the lengthening of the actual note. In this case the second action involved is a spectral linear interpolation between the release of the actual note and attack of the next note over the intersection of the two (see figure 5). The overlap region length 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 prolonged to reach the pitch transition point. Here, a $10 - 15\ msec$ transition is generated by interpolating the tracks of the actual note with the ones of the successive. In this way, a transition without *glissando* is generated. Glissando effect 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 sufficiently general and efficient for real-time purposes.

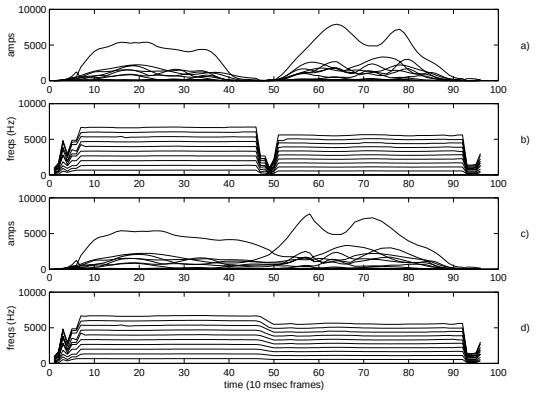

Figure 5: *Legato* of two notes originally separated by a *micropause* (only the first ten partials of the sinusoidal analysis are shown)

Intensity: the intensity and spectral envelope of the sound frame are controlled by means of the map $\mathcal{F}$. For a desired change δI, is $\mathcal{F}(\delta I) = [r_1 r_2 \cdots r_H]$, where r_h is the magnitude deviation (in dB) to be applied to the hth harmonic.

Envelope Shape: The center of mass of the amplitude 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 function is applied to the energy envelope, where the apex of the triangle correspond to the new position of the accent.

5. REAL-TIME IMPLEMENTATION

One of the objective of the system is to give the user the possibility to interact with the model of expressiveness by moving in the perceptual space and feel the changes rendered by the sound processing engine. This real-time approach requires the management of some aspects related to the information exchange and the synchronization between the model and the processing engine. The communication protocol between the model and the sound processing engine must be bi-directional, in order to let the model know the exact time-position of the processing engine, which is constrained to follow the time evolution of the underlying signal. This sometimes produces delayed responses of the system to the user input.

6. CONCLUSIONS

A system for real-time control of expressive intentions in musical performance has been presented. The main result of this research is the integration of an expressiveness model with an audio effects engine, which allows for expressive signal processing. Although simplified transformations on instrument transients where assumed for real time implementation, the system proved the feasibility of interactive transformations of expressiveness in digitally recorded performances.

7. REFERENCES

[1] J. L. Arcos, R. L. de Mántaras, and Xavier Serra, "Saxex: A case-based reasoning system for generating expressive musical performances," *Journal of New Music Research*, pp. 194–210, Sept. 1998.

[2] R. Di Federico and C. Drioli, "An integrated system for analysis-modification-resynthesis of singing," *Proc. of the IEEE SMC98 Conf.*, pp. 1254–1259, Oct. 1998.

[3] S. Canazza and A. Rodá, "A parametric model of expressiveness in musical performance based on perceptual and acoustic analyses," *Proc. of the ICMC99 Conf.*, 1999, To be published.

[4] S. Canazza, A. Rodá, and A. Vidolin, "Analysis and synthesis of expressive intentions in musical performance," *Proc. of the ICMC97 Conf.*, pp. 113–120, 1997.

[5] X. Serra, "Musical sound modeling with sinusoids plus noise," *in Musical Signal Processing*, pp. 497–510, 1997, Swets and Zeitlinger.

[6] C. Drioli, "Radial basis function networks for conversion of sound spectra," *Proc. of the DAFX99 Conf., Trondheim*, Dec. 1999, To be published.