REAL-TIME MORPHING AMONG DIFFERENT EXPRESSIVE INTENTIONS IN AUDIO PLAYBACK

Sergio Canazza, Giovanni De Poli, Carlo Drioli, Antonio Rodà, Federico Zamperini

University Of Padova, Center of Computational Sonology (CSC)
Dept. of Electronic and Informatics – Via Gradenigo 6/a – 35100 Padova, Italy
canazza@dei.unipd.it

Abstract
In this work we present a method for controlling the mutable musical expressiveness in an abstract way, not necessarily related to the acoustic parameters. Our approach integrates symbolic and audio processing in a real-time networked environment. We present a mapping strategy allowing to move in the control space and varying the expressiveness in a coherent way. All the variations computed by the model on the symbolic representation are successively managed by the signal processing sub-system which coherently transforms the sound by combination of different audio effects. The signal processing engine works in real time to render the desired expressive audio variations. The model is used interactively in a networked environment. The user controls the expressive character of the performance by moving within an appropriate control space.

1 Introduction
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, designer’s attention is put on graphics 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 in multimedia systems, while the visual part has evolved consequently, the paradigm of the use of audio has not changed adequately. In the usual interaction with audio media, transformations on audio objects are not allowed. A more intensive use of digital audio effects, will allow 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 by changing the interpretation of a musical piece. 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 a specific interpretation of the score or the situation, by adding mutable expressive intentions. In a similar way we could be interested in having models and tools which allow to modify a performance by changing its expressive intention. In (Canazza et al. 1998, 1999) we presented models which are able to modify the expressive content of a performance in a gradual way.

We aim at using these models in applications of multimedia products; in particular, an extension of these models is presented, focused on auralization and web based applications.

Fig. 1: Scheme of the system

2 Controlling expressiveness
For an effective control on expressiveness, three different levels of abstraction are needed. The control space is 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. In order to realize morphing among different expressive intentions we developed two abstract control spaces: the first, perceptual parametric space (PPS), was derived by multidimensional analysis of various professionally performed pieces ranging from western classical to popular music. The second, synthetic expressive space, allows the artist to organize his own abstract space by defining expressive points and positioning them in the space. In this way, a certain musical expressive intention can be associated...
to the various multimedia objects. Therefore, audio is modified in its expressive intention, both when the user focuses on a particular multimedia object (by moving with a suitable pointer) and when the object itself enters the scene. One can also take into account the possibility that the multimedia object tell the system information about its state, which can be exploited for expressive control on audio. For instance, in a virtual environment the avatar can tell the system its intentions which will be used to control audio expressiveness; one can thus gain a direct mapping between the intentions of the avatar and audio expressiveness, or a behavior chosen by the artist in the design step.

Using suitable mapping strategies allows the user to vary coherently and gradually the expressiveness (i.e. morphing among happy, solemn and dark), by moving inside the control space. Morphing can be realized with a wide range of graduality (from abrupt to very smooth), allowing to adapt the system to different situations. Analysis-by-synthesis method was applied to estimate which kind of morphing technique ensures the best perceptual result. The computer-generated performances showed appropriate expressive meaning in all the points of control space, computing intermediate points of the space using a quadratic interpolation.

It has to be noticed that expressive content of a performance is revealed on a time scale which is longer of that of a single event; therefore, in order to obtain a fruition which is coherent with the artist's intentions, the system allows to slow down the movements of the user, so to avoid unwanted "expressive discontinuities" in correspondence of abrupt movements. To this end, suitable smoothing strategies have been developed for movement data coming form the pointer.

The expressive parameter layer translates high-level information of the control space in order to modify the acoustical parameters used by the models for the expressive morphing. On the basis of performance analysis (Canazza et al. 1997) some of the parameters have turned out to be particularly important for the reproduction of expressive intentions, for instance Tempo, Legato, Intensity, Phrasing, etc. The models make use of these parameters in order to determine the deviation which have to be applied to the score for the reproduction of the desired expressive intention. The user can define an expressive intention by means of a suitable set of parameters, or can use a mapping Intentions-parameters derived by acoustical analysis.

The system can handle three different models for expressive morphing, which make use of different levels of information from the score. The first one (Canazza et al., 1998) has three inputs: the score, a description of its musical structure, and a "neutral" performance (i.e. human performance without any expressive intention). A second model (Canazza & Rodà, 1999) needs the score together with a neutral performance. Finally, a third model needs only a performance (nominal or neutral). The last two models work in real time, so in the next we will refer to them. By means of the net-protocol, the model is used interactively in a networked environment. The user controls the expressive character of the performance by moving within the control space using suitable control devices.

3 Model

The expressiveness model acts on an abstract level (morphing level), computing the deviations of all musical parameters involved in the transformation. For each expressive intention, the deviations of the acoustic parameters are computed using the following equation

\[
\delta P(n) = k_P \cdot \frac{P_{in}(n)}{P_{out}(n)} + m_P \cdot \left(1 - \frac{P_{in}(n)}{P_{out}(n)}\right)
\]

where:

- \( n \) is the cardinal number of the \( n \)-th note of the score.
- \( P \) stands for the different parameters modified by the model. Up to now, the model can process the Inter Onset Interval between the current and the successive note (IOI), the duration of the current note (DR), the duration of the attack (DRA), the mean envelope energy (E), the time location of the energy envelope center of mass (EC).
- The subscript \( in \) indicates that the value of the parameter \( P \) is computed from the inputs. In fact, some parameters (e. g. IOI and L) were computed starting from both score and neutral performance.
- The subscript \( out \) indicates the value of the expressive performance, that is the output of the system.
- \( P \) stands for 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 one performs a stretching of the values.

For each parameter \( P \), the \( k \) and \( m \) coefficients are computed by means of a mapping strategy (Canazza et al., 1998, 1999) obtained by processing results from acoustical and perceptual analyses (Canazza et al. 1997). By means of these strategies, different points of the control space are associated to sonological parameters ("mid-level" acoustical parameters) such as intensity, tempo, articulation, phrasing, and so on. In Fig. 2 we show a mapping surface between points of a two-dimensional control space (PPS) and the parameter \( K_{tempo} \), (a value greater than one stands for a rallentando, while a value lower than one stands for an accelerando). On the basis of movements of the pointer on the xy plane, the variations of the parameter \( K_{tempo} \) to be applied to the performance are thus computed.
4 Audio Processing

All the variations computed by the model on the symbolic representation are successively managed by the signal processing sub-system which coherently transforms the sound by combination of different audio effects.

The signal processing engine works in real time to render the desired expressive audio variations. It is based on a sinusoidal model analysis/resynthesis framework, a technique suitable for high quality sound transformation and higher level attribute recognition such as notes, grace notes and staccato-legato or vibrato. The input of the system consists of a digitally recorded performance with neutral expressive intention, and of a symbolic segmentation of the performance representing the musical-level attributes like note onsets, and offsets. This allows the definition of a joint description for sound and performance levels, and modifications in the symbolic level are reflected to the signal level through the audio processing techniques. To this purpose, a finite state automata (here called “Effect Manager”), responsible for the selection, combination and time scheduling of the basic audio effects, computes in real time low-level frame-rate control curves related to the desired expressive intentions. These time-varying curves are then used to simultaneously control the different audio effects during morphing.

The basic audio effects involved in the expressive processing are: Time stretching, Pitch shifting, Envelope scaling, and Spectral processing. The rendering of the deviations computed by the expressiveness model (“high-level” effects) may imply the use of just one of the basic sound effects, or the combination of two or more of these effects, as shown in Table 2 and as detailed in the following: Local tempo processing (variation of IOI and DRA) is realized by computing different time-stretching factors for the attack, sustain, and release of notes. Legato control relies on time stretching for changing the duration of the note, and on spectral interpolation between the release of the actual note and the attack of the next note, if overlapping occurs due to time stretching. More details on the management of local tempo and legato processing can be found in (Canazza et al., 1999).

<table>
<thead>
<tr>
<th>Effect</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>δIOI</td>
<td>Time Stretching</td>
</tr>
<tr>
<td>δL</td>
<td>Time Stretching &amp; Spectral Processing</td>
</tr>
<tr>
<td>δDRA</td>
<td>Time Stretching</td>
</tr>
<tr>
<td>δS</td>
<td>Spectral Processing &amp; Envelope Scaling</td>
</tr>
<tr>
<td>δEC</td>
<td>Envelope Scaling</td>
</tr>
</tbody>
</table>

Fig. 4: Basic audio effects involved in each high level audio effect.

The musical accent of the note, which is related to the amplitude envelope centroid, is usually located on the attack of notes for expressive intentions like light or heavy, or on the release for intentions like soft or dark. This parameter can be controlled by changing the position of the apex of a triangle fitted to the original energy envelope, and by scaling the energy of frames to fit the new target triangle.

Intensity control can be obtained by scaling of amplitudes of the harmonics. 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 method based on accurate interpolation of spectra relatives to different intensities is used. From the analysis of performances with different expressive intentions, he mean spectral envelopes for maximum intensity \( I_{max} \), minimum intensity \( I_{min} \), and intensity of the natural performance \( I_{nat} \), are obtained for each note (Figure 5, upper frame, shows a single note). An interpolation schema is then designed in terms of a map \( \mathcal{F} : \mathbb{R} \rightarrow \mathbb{R}^H \), where \( H \) is the number of harmonics, so that for a desired intensity change \( \delta I \), \( \mathcal{F}(\delta I) = [\delta f_1 \ldots \delta f_h] \) gives the magnitude deviations \( \delta f_h, h=1..H \), to be applied to each harmonic of the reference spectrum (i.e., the one relative to the intensity \( I_{nat} \)). For the purpose of having a parametric model to represent the behavior of the sound spectrum, Radial Basis Function Networks are used (Drioli, 1999). To resume the principal characteristics,
RBFNs can learn from examples, have fast training procedures, and have good interpolation properties. The proposed interpolation schema is used to compute the spectral envelope related to a desired intensity in $[I_{\min}, I_{\max}]$ (Figure 5, lower frame). The analysis step is due for each musical note in the score, in order to have different maps for notes with different pitch.

![Fig. 5](image1.png)

**Fig. 5**: Control of the intensity of sound, with spectral preserving interpolation. Upper figure: analysis step. Lower figure: synthesis of spectra relatives to a desired intensity $I$, with $I$ in $[I_{\min}, I_{\max}]$.

5 Once Upon a time

An example is given in this section, where an application for the fruition of fairy-tales in a remote multimedia environment is designed. This kind of applications allowed us to validate the system. Moreover, it allowed to select the sonological parameters that best agreed with the expressive morphing. In Table 1, the parameters used in this application are shown. With the selected parameters, an expressive character can be assigned to each individual in the tale and to the different multimedia objects of the virtual environment. In this way, the character of each individual of the tale become stronger due to the expressive content of audio.

![Table 1](image2.png)

**Table 1**: The sonological parameters used to define the expressive characters in the tale called “Once upon a time...”

The application has been realized as an applet. The expressive content of audio is gradually modified with respect to the position and movements of the mouse pointer, using the mapping strategies described above.

6 Conclusions

In multimedia products, audio is becoming an active part of the non-verbal communication process. The proposed system gives to the user the possibility to build multimedia applications in which the expressiveness of audio is modified with respect to the user actions. The system is based on three levels and musical information is processed on an abstract level.

Acknowledgements

This work has been supported by Telecom Italia under research contract *Cantieri Multimediali*.

References


