Editors: Tizania Catarci and Thomas D.C. Little

Audio Morphing Different Expressive Intentions for Multimedia Systems

Sergio Canazza, Giovanni De Poli, Carlo Drioli, Antonio Rodà, and Alvise Vidolin
University of Padova, Italy

In multimedia products, graphical and audio objects enrich textual information. Combining these elements correctly makes it easier for users to interact with the software. Usually, multimedia authors focus on the graphical objects rather than sound, which complements an image or provides a musical comment to text and graphics. While the demand for increased interaction with programs has helped the visual part evolve, the paradigm of using audio hasn't changed adequately, resulting in several different objects rather than a continuous transformation of these objects. A more intensive use of digital audio effects will let sounds adapt interactively to different situations, allowing users to enjoy the product even more.

We believe that the evolution of audio interaction promotes the use of expressive content. Such an interaction should allow a gradual transition (morphing) between different expressive intentions.

Recent studies have demonstrated that it's possible to communicate expressive content at an abstract level, which can change the interpretation of a musical piece.[1-4] In fact, in human musical performance, the performer organizes acoustical or perceptual changes in sound to communicate different emotions to the listener. The same piece of music can be performed by differently by the same musician—trying to convey a specific interpretation of the score or the situation—by adding mutable expressive intentions. Similarly, we're interested in having models and tools that let us modify a performance to change its expressive intention.

Our study of the expressive content of audio led us to develop a system for the auralization of multimedia objects for Web-based applications.

The audio authoring tool

Authoring tools are frequently used in multimedia object development. These tools handle audio as a comment to video objects. They have simple synchronizing tools and, sometimes, audio editors to process (in a static way) prerecorded sounds. In this way, different musical events can be associated to relative objects. Conversely, we aim to vary the expressive content of music—much like a soundtrack that uses the same musical theme in different contexts, adapting its expressive content to highlight a situation's ambiance.

Our audio authoring tool manages audio expressive content. Authors can apply a smooth morphing among different expressive intentions in music performances and adapt the audio-expressive character to their taste. The audio-authoring tool lets you associate different expressive characters (rather than musical melodies) to various multimedia objects. Cinema achieves this by varying the expressive content of the musical theme, which adds emotion (such as sadness, excitement, and so on) not directly communicated by the video.

However, this approach doesn't work well in the multimedia field, where users interact with audio-visual events. On the contrary, we believe that having a smoothly varying musical comment can augment the user's emotional involvement, compared to using an abrupt succession of different sound comments.

Figure 1 shows the structure of our audio authoring tool. The input consists of a description

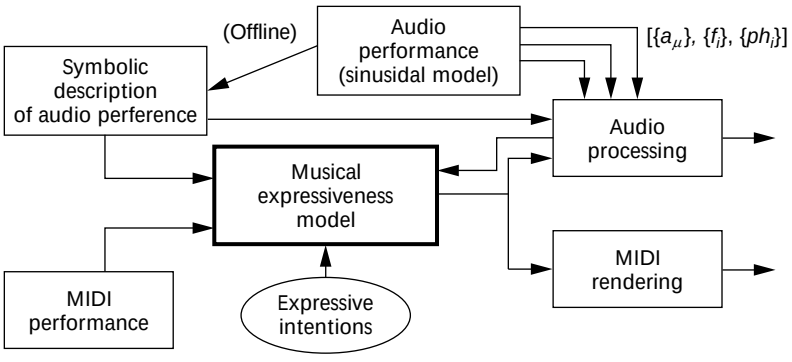

Figure 1. System architecture.

1070-986X/00/$10.00 © 2000 IEEE

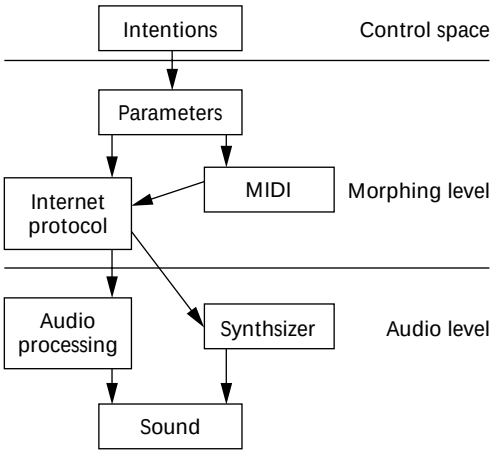

Figure 2. Scheme of the system showing three levels of abstraction: control space, morphing level, and audio level.

Figure 3. Example of a set of parameters associated to the adjectives of the control space: duro (hard), brillante (bright), leggero (light), morbido (soft), and pesante (heavy). The left window presents the perceptual expressive space. The right window shows the design window for the parameters associated to the expressive intentions.

of a neutral musical performance (played without any expressive intention), the nominal score of the performance, and an interactive control facility for the expressive intention the user wants. The neutral performance consists of a polyphonic accompaniment stored as a Musical Instrument Digital Interface (MIDI) file and by a digitally recorded monophonic part. We derive a symbolic (MIDI-like) description of the audio performance offline by analyzing the recorded signal, which contains all the musical parameters the system needs (such as note onset and offset; time definitions of attacking, sustaining, and releasing notes; and information on higher level attributes). The system acts on the symbolic level, computing the deviations of all musical parameters involved in the transformation, and finally drives the MIDI synthesizer section or the audio processing engine. In the last case, a resynthesis procedure reproduces the sound from its sinusoidal representation. All the sound transformations the system addresses occur at resynthesis time by a frame-rate parametric control of the audio effects.

Controlling expressiveness

Effectively controlling expressiveness requires three different levels of abstraction (see Figure 2). The control space is the user interface, which controls, at an abstract level, the expressive content and the interaction between the user and the multimedia product's audio object.

To realize morphing among different expressive intentions, the audio-authoring tool works with two abstract control spaces. The first one, called *perceptual expressive space*, was derived by a multidimensional analysis of various professionally performed pieces ranging from western classical to popular music. The second, called *synthetic expressive space*, lets authors organize their 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, the audio object changes its expressive intention, both when the user focuses on a particular multimedia object (by moving it with a pointer) and when the object itself enters the scene.

Our tool permits receiving information from a multimedia object about its state—the author can exploit it for expressive control of audio. For instance, in a virtual environment the avatar can tell the system its intentions, which controls audio expressiveness. Thus, you can gain a direct mapping between the intentions of the avatar and audio expressiveness or a behavior chosen by the artist in the design step.

With mapping strategies, users can vary the expressiveness (that is, morphing among happy, solemn, and sad), by moving inside the control space. Morphing can be realized with a wide range of graduality (from abrupt to very smooth), allowing the system to adapt to different situations.

As a case study we developed a perceptual expressive space for sensorial adjectives (Figure 3). We applied the analysis-by-synthesis method 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 the control space by computing intermediate points of the space using a quadratic interpolation.

Note that a time scale reveals a performance's expressive content. To obtain results coherent with the artist's intentions, the system can slow down the movements of the pointer the user controls to avoid unwanted "expressive discontinuities" as a result of abrupt movements. To this end, we implemented smoothing strategies for movement data coming from the pointer.

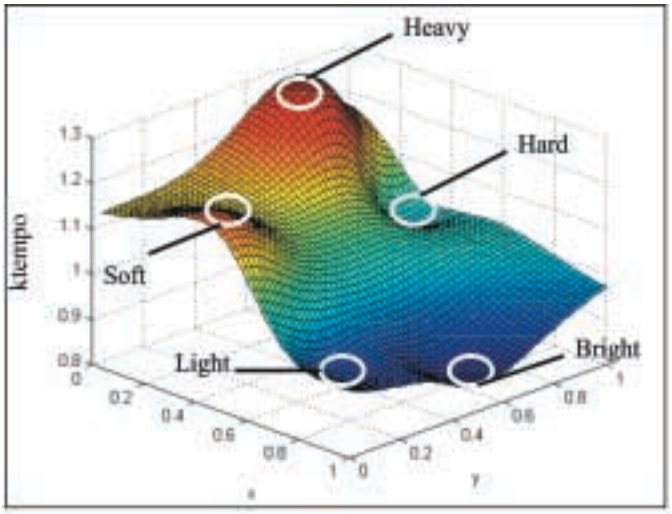

Figure 4. Mapping surface among points of a 2D space (x-y plane) and the parameter Ktempo (z axis).

The morphing layer translates high-level information from the control space to modify the acoustical parameters the system uses for expressive morphing. Based on performance analysis, some of the parameters have turned out to be important for the reproduction of expressive intentions (for instance, tempo, legato or smoothness, intensity, phrasing, and so on). The system uses these parameters and determines the deviations that must be applied to the score for reproducing a particular expressive intention. Authors can define an expressive intention through a set of parameters or use mapping intentions—parameters derived from acoustical analysis.

Through the Internet protocol, the objects authors design can interact in a networked environment. The user controls the performance's expressive character by moving within the control space using multimodal control devices.

Expressiveness morphing module

The system uses different modules for expressive morphing. We developed these modules based on the results of perceptual and acoustic analyses of professional performances. These analyses revealed that to render a particular expressive intention, the performer uses different strategies depending on personal preferences, the score structure, and the musical instrument's expressive controls. To render the expressive content, the morphing models use a reduced set of controls, which have proven to be representative and independent of the instrument and of the score.

Through different morphing strategies, the system relates the regions of the control space to sonological parameters (medium-level acoustic parameters) such as loudness, tempo, articulation, phrasing, and so on. Figure 4 shows a mapping surface among points of a 2D control space and the parameter Ktempo. A value greater than 1 stands for *rallentando* (gradually slackening in tempo), while a value lower than 1 stands for *accelerando* (gradually accelerating in tempo). Based on the movements of the pointer on the *x-y* plane, the system computes the variations of the parameter Ktempo to apply to the performance. In this case, the model computed intermediate points of the space using a quadratic interpolation.

Audio processing

To render imitative musical messages with high-quality results (such as for a violin or singing voice, which permit a high expressive control), we included a postprocessing tool. To control the expressiveness of digitally recorded musical performances, the audio authoring tool interfaces with a postprocessing engine, which works in real time to render the desired expressive audio variations. All the variations that the system computes on the symbolic representation are managed by the postprocessing engine, which transforms the sound through different audio effects.

We based the postprocessing tool on a sinusoidal model analysis-resynthesis framework—a technique for high-quality sound transformation. The input of the postprocessing tool consists of a digitally recorded performance with neutral expressive intention and a symbolic description of the musical attributes (such as legato, vibrato or vibration, and so on). This provides a joint description for sound and performance levels. Modifications in the symbolic level are reflected in the signal level through audio processing tech-

Courtesy of Telecom Italia and Univ. of Padova

Figure 5. Screen shot from the application "Once upon a time ...."

niques. For this purpose, the model computes in real time low-level, frame-rate control curves related to the desired expressive intentions. These time-varying curves then simultaneously control the different audio effects during morphing. Expressiveness-oriented, high-level audio effects such as tempo, brightness, or legato variations are realized by an organized control of basic audio effects such as time stretching, pitch shifting, and spectral processing. An organized sound processing control system handles selecting, combining, and time scheduling the basic audio effects.

Once upon a time ...

Figure 5 shows an application of our tool, released as an applet, for fairy tales in a remote multimedia environment. In this kind of application 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, the different expressive intentions are located in synthetic control spaces defined for the specific contexts of the tale. The expressive content of audio gradually changes in response to the position and movements of the mouse pointer, using the mapping strategies described above.

Conclusion

For multimedia products, we believe that audio should become an active part of the nonverbal communication process. Our audio authoring tool lets designers build multimedia applications in which the expressiveness of audio changes based on users' actions.

Our plans for future research include the following:

1. The analysis and recognition of expressive gestures from the users both in the particular modality (such as recognizing expressive information in human movements and musical gestures) and from a multimodal perspective (such as how to use information coming from the analysis of expressive content in human movement to perform a better and deeper analysis of expressive content in music performances and vice versa).

2. The communication of expressive content in the particular output channels (such as communication of expressive content through sound, music, movements, and visual media) as well as in a multimodal perspective—that is, the synthesis of expressive content from combining several nonverbal communication channels. MM

Acknowledgments

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

References

1. S. Canazza, G. De Poli, and A. Vidolin, "Perceptual Analysis of the Musical Expressive Intention in a Clarinet Performance," *Music, Gestalt, and Computing*, M. Leman, ed., Springer-Verlag, Berlin, 1997, pp. 441-450.
2. G. De Poli, A. Rodà, and A. Vidolin, "Note-by-note Analysis of the Influence of Expressive Intentions and Musical Structure in Violin Performance," *J. New Music Research*, Vol. 27, No. 3, 1998, pp. 293-321.
3. A. Gabrielsson, "Expressive Intention and Performance," *Music and the Mind Machine*: *The Psychophysiology and Psychopathology of the Sense of Music*, R. Steinberg, ed., Springer-Verlag, Berlin, 1995, pp. 35-47.

4. C. Palmer, "Anatomy of a Performance: Sources of Musical Expression," *Music Perception*, Vol. 13, No. 3, 1996, pp. 433-453.

Readers may contact De Poli at the Center of Computational Sonology, University of Padova, Via San Francesco 11A 35100, Italy, e-mail depoli@dei.unipd.it, http://www.dei.unipd.it/ricerca/csc/intro.html.

Contact Multimedia at Work editors Tizania Catarci at the Department of Information Systems, University of Rome "La Sapienza," Via Salara 113, 00198 Rome, Italy, e-mail catarci@dis.uniroma1.it; and Thomas Little at the Multimedia Communications Lab, Department of Electrical and Computer Engineering, Boston University, 8 Saint Mary's St., Boston, MA 02215, e-mail tdcl@computer.org.

Author guidelines for IEEE MultiMedia *are available online at http://computer.org/multimedia/author.htm.*

Web Extras

To listen to sample audio files and view a demo of the audio authoring tool, visit http://computer.org/multimedia/mu2000/u3toc.htm and click on the links:

- *Sonatina in sol* (by Beethoven) played neutral (without any expressive intentions): http://computer.org/multimedia/mu2000/extras/u3079x1.mp3
- Expressive performance of *Sonatina in sol* generated by the authoring tool in a symbolic way (that is, as a MIDI file): http://computer.org/multimedia/mu2000/extras/u3079x2.mp3
- *Sonata K545* (by Mozart) played neutral (without any expressive intentions): http://computer.org/multimedia/mu2000/extras/u3079x3.mp3
- Expressive performance of *Sonata K545* generated by the authoring tool in a symbolic way (that is, as a MIDI file): http://computer.org/multimedia/mu2000/extras/u3079x4.mp3
- Expressive performance of *Sonata in A Major Op. V* (by Corelli) generated by the audio authoring tool (using the audio postprocessing tool): http://computer.org/multimedia/mu2000/extras/u3079x5.mpg