A Centralized Audio Presentation System

Albert L. Papp III and Meera M. Blattner
University of California, Davis, and
Lawrence Livermore National Laboratory
P.O. Box 808
Livermore, CA 94551

Abstract

The centralized audio presentation system addresses the problems which occur when multiple programs running simultaneously attempt to use the audio output of a computer system. Concurrent audio output can lead to perceptual problems due to psychoacoustic phenomena. The presentation system receives abstract parameterized message requests from the currently running programs, and attempts to create and present a sonic representation in the most perceptually manner through the use of a theoretically and empirically designed rule set. Furthermore, the combination of speech and nonspeech audio is examined; each presents its own problems of perceivability in an acoustic environment composed of multiple auditory streams.

1 Introduction

User interfaces which support concurrent program executions have little, if any, audio management. Typically, some number of audio channels exists as a resource, and programs request the number of audio channels required and the operating system either grants or denies the request. Therefore, in environments where multiple programs output sound, each individual program has no overall context of the auditory system state with the possible exception of how many audio channels the operating system has allocated for other applications. Research has been conducted in creating audio servers [1, 2] to aid in the resource management issue of audio presentation.

However, concurrent sound presentations can potentially lead to numerous problems in the perception of the various sources of sound. Programs typically play audio without regard for the overall auditory environment, which can cause sound masking and perceptual unintelligibility. When numerous applications output audio, there may be confusion in knowing which application produced what sounds.

These perceptual problems can be addressed by continual examination of the global auditory state of the computer system. If an audio presentation is able to be altered in a well understood manner, perceptibility can be improved. The centralized audio presentation system attempts to do just that. It receives descriptive messages which contain information about system activities and program states, as specified by the user or application programmer. The sonic output of the set of running programs and the overall auditory system state is controlled by the presentation system. It chooses how the information is to be represented in sound, within the constraints of the descriptive message. The concept of dynamic representations of information in different forms, or *multimodal objects*, is discussed in [3]. The presentation system must choose the form with consideration for other current sonic output. The appropriate auditory form must not only convey the informational content of the message but also keep all current audible information

3 Interaction of Different Sound Media

In order to display the various auditory media simultaneously in an effective way, the presentation system must have some knowledge of how the different media interact with one another.

3.1 Historical Precedents

There are few precedents for this type of interaction in sounds generated by real world objects, so we have to look elsewhere to begin examining the problem. As has been done in other work in nonspeech audio [12, 13], we examined the historical precedents based on music to examine how related problems were solved. In this specific case, we examined how composers had intertwined voice and nonspeech audio. We specifically did not want to consider singing as voice, although there is a good deal to be learned from song that can be applied to the problem. There are a number of musical precedents for spoken dialogue with background music. It must be emphasized in our examples, that both the dialogue and the music were composed as one piece, rather than an alternation of voice and music as in Peter and the Wolf.

Our first example is from melodrama, a genre of musical theater found in 19th century opera. *Melodrama* is "a genre of musical theater that combined spoken dialogue with background music" [14]. One of the most famous of these is by Carl Maria von Weber in the Finale of *Der Freischutz.* The intention of the finale was to create a diabolically eerie scene with the speaker, Casper, evoking nature-pictures of frightening scenes. Diminished and augmented intervals as well as chromaticism are used in the melody and harmony. Another example we considered were two pieces written in the 20th century by Arnold Schoenberg, *Ode to Napoleon* and *Survivor from Warsaw.* We examined the interlacing of the voice of a narrator in both of these, but the music itself has elements of atonality or his 12-tone method, which we did not consider in our analysis. Lastly, we looked at an example from Rap music, Young M.C.'s *Bust a Move.*

3.2 The Analysis

The material can be analyzed from many aspects. There are two that are particularly interesting which we discuss briefly. The first is the semantic content of the dialogue and background and the second is the use of simultaneous or sequential voice and background sounds. At some points there are real-world sounds accompanying both voice and music. This occurs more often in *Der Freischutz* because it is an opera in which singers are acting out roles. Casper pounds and hammers as he works. Whips crack, horses neigh, and dogs bark in one of his images. The narrator in *Survivor from Warsaw*, recalls horrifying images as well, and in the background one can hear soldiers marching. Both of these pieces have choruses that suddenly begin during the narration. The mood of the voice and the music are reflected in one another.

Often when words are to be emphasized, the background sounds cease entirely and the voice is heard alone. Similarly, in Rap music, lack of a musical element functions to shift more emphasis to the spoken word. In *Bust a Move*, the bass guitar carries the melody line of the underlying music through much of the composition. However, toward the end of a chorus, the bass abruptly stops as the narrator speaks the emphasized phrase "bust a move" over the soft rhythm.

When the voice stops and music starts, a new idea may begin and the music is used to bridge the spoken sections. Music or real-world sounds often precede the voice as to introduce it. Short auditory messages or motifs often function to indicate a character's presence or recurring theme. If the voice is speaking nearly alone (sometimes soft, barely audible background sounds can be heard) as the emotional content increases, the music gets louder and definite themes are heard.

3.3 Application to Computer Interfaces

How can this be applied to the computer interface? We are doing some experimental work to examine the various effects mentioned above. These effects may be simple to apply in many different cases. The principles found in pieces such as the ones mentioned above can be carried over almost in their entirety. A theme which captures some image either through real-world sounds or with earcons can be heard before a voice is heard. This sets the scene for the message and alerts the listener to the spoken words. Sounds used during speech can emphasize important parts of the message or even impart information that is not spoken but that can be conveyed quickly through emotional response, such as the urgency of the message. Sounds structured using known musical constructions, such as hierarchical structure or transformations can quickly identify a series of related musical fragments to the listener [5]. On the other hand, perception of a missing nonspeech audio element can serve to emphasize the spoken word. Finally, sound that follows a message can be used to bridge messages and convey continuity between them. Real-world sounds can be used between messages to reinforce their content.

4 Example Application: Heart Rate Monitor

This program monitors the user's heart rate and presents a constant auditory message which indicates that rate. Furthermore, it allows the user to set a desired "target" heart rate zone and an upper bound which it would be dangerous to allow the heart rate to exceed. The heart rate indicator and target zone messages are sent only once to the presentation system since they are forms of persistent messages. The other discrete messages are sent by the application at the appropriate times.

1. The *Heart Rate Indicator*
   This message indicates the current heart rate, and could be represented as a parameter in an algorithmically generated music background or as a tone whose volume or pitch is changed as the rate changes.

   - Perception Style: *Persistent*
   - Type: *Abstract Earcon, Voice*
     The heart rate (in beats per minute) is the synthesizer parameter.
   - Priority: *Medium-Low*
   - Latency: *100*
   - Precision: *Medium*

2. The *Target Zone Indicator*
   The message indicates that the user is in the specified rate zone. The message should be present whenever in the zone, and absent otherwise.

   - Perception Style: *Conditionally Persistent*
     The heart rate must be in the user specified zone for this message to be displayed.
   - Type: *Abstract Earcon, Representational Earcon, Voice*
   - Priority: *Medium-High*
   - Latency: *200*
   - Precision: *Medium-High*

3. The *Below-Zone Event Marker*
   This message indicates that the user's heart rate has just dropped below the target zone.
   - Perception Style: *Discrete*
   - Type: *Abstract Earcon, Representational Earcon, Voice*
   - Priority: *Medium-High*
   - Latency: *50*
   - Precision: *Medium*
4. The *Above-Zone Event Marker*
   This message indicates that the user's heart rate has just gone above the target zone.
   - Perception Style: *Discrete*
   - Type: *Abstract Earcon, Representational Earcon, Voice*
   - Priority: *Medium-High*
   - Latency: *50*
   - Precision: *Medium*
5. The *Over Maximum Threshold Event Marker*
   This message indicates that the user's heart rate has exceeded the maximum value. Harm will come to the user unless the heart rate is dropped immediately.
   - Perception Style: *Discrete*
   - Type: *Voice*
   - Priority: *High*
   - Latency: *0*
   - Precision: *High*

5 Conclusion

Many of the problems introduced by allowing multiple applications to output audio can be solved with a centralized approach. By giving the application programmer a higher level abstraction to encode information into sound, the burden of designing audio for the interface is lessened. Furthermore, by allowing a particular auditory event to have multiple representations, the presentation system can attempt to choose the "best" one based on the current auditory state. This dynamic decision leads to an improved auditory perception since the application programmer would not have been able to statically design the sonic output with any knowledge of the overall auditory context to which the new application was being added.

6 Acknowledgments

This work was performed with partial support of NSF Grant #IRI-9213823 and under auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

Meera Blattner is also with the Department of Biomathematics, M.D. Anderson Cancer Research Hospital, University of Texas Medical Center, Houston.

References

[1] Arons, B. "Tools for Building Asynchronous Servers to Support Speech and Audio Applications." In *Proceedings of the ACM Symposium on User Interface Software and Technology*, 1992.

[2] Reichbach, J. D., and R. A. Kemmerer. "Soundworks: An Object-Oriented Distributed System for Digital Sound." *IEEE Computer* (March 1992): 25–37.

[3] Glinert, E. P., and M. M. Blattner. "Programming the Multimodal Interface." In *ACM Multimedia '93 Proceedings* (1993): 189–206.

[4] Bregman, A. S. *Auditory Scene Analysis.* Cambridge, MA: MIT Press, 1990.

[5] Blattner, M. M., D. A. Sumikawa, and R. M. Greenberg. ' "Earcons and Icons: Their Structure and Common Design Principles." *Human–Computer Interaction* **4(1)** (1989): 11–44.

[6] Brewster, S. A., P. C. Wright, and A. D. N. Edwards. "A Detailed Investigation into the Effectiveness of Earcons." In *Auditory Display: Sonification, Audification, and Auditory Interfaces*, edited by G. Kramer, 471–498. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XVIII. Reading, MA: Addison Wesley, 1994.

[7] Williams, S. M. "Perceptual Principles in Sound Grouping." In *Auditory Display: Sonification, Audification, and Auditory Interfaces*, edited by G. Kramer, 95–126. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XVIII. Reading, MA: Addison Wesley, 1994.

[8] Gaver, W. W. "Auditory Icons: Using Sound in Computer Interfaces" *Human–Comp. Interaction* **2(2)** (1986): 167–177.

[9] Brewster, S. A. "Providing a Model for the Use of Sound in User Interfaces." Technical Report #169, University of York, Heslington, York, Y01 5DD, June 1991.

[10] Langston, P. S. "Img/1: An Incidental Music Generator." *Comp. Mus. J.* **15(1)** (1991): 28–39.

[11] Langston, P. S. "(201) 644-2332 or Eedie & Eddie on the Wire, an Experiment in Music Generation." In *Proceedings of the Usenix Summer '86 Conference*, 1986.

[12] Blattner, M. M., and R. M. Greenberg. "Communicating and Learning Through Non-speech Audio." In *Multimedia Interface Design in Education*, edited by A. Edwards and S. Holland, 133–143. NATO ASI Series F. Berlin: Springer-Verlag, 1992.

[13] Blattner, M. M., R. M. Greenberg, and M. Kamegi. "Listening to Turbulence: An Example of Scientific Audiolization." In *Multimedia Interface Design*, edited by M. Blattner and R. Dannenberg, 87–102. Reading, MA: ACM Press/Addison-Wesley, 1992.

[14] Grout, D. J. and C. V. Palisca. *A History of Western Music.* New York: W. W. Norton, 1988.