Audio in Multimodal Applications

By Francis Rumsey
Staff Technical Writer

INTRODUCTION

A papers session held last year at the 127th Convention in New York, "Audio in Multimodal Applications," highlighted the ways in which audio interacts with other sensory inputs such as visual and tactile ones. Interactive applications for audio within medical environments, games, and installations were discussed, along with novel approaches to spatialization and stereo image correction for displays.

LISTENING WITHIN AND WITHOUT YOU

Inspired by research in the visual domain, Santoro et al. looked into the cognitive performance of individuals when listening to sounds inside and outside their heads in paper 7865, "Listening within You and without You: Center-Surround Listening in Multimodal Displays." They define center sounds as those perceived "within you" and surround sounds as those perceived "without you," which is slightly different from the terminology normally employed with loudspeaker surround sound. Spatialized auditory displays that use binaural stimulation based on HRTFs (head-related transfer functions) can present either the same sound to each ear (known as diotic presentation) or different sounds to each ear (so-called dichotic presentation). The former tends to give rise to sounds perceived inside the head and the latter to sounds that are more externalized. When investigating these they were interested in the speed of decisions related to stimuli of each type, and in comparing the difference between stimuli presented together and separately.

One of the primary motivators for the study was the idea that there is evidence of parallel processing in the visual field for images in the focal or central area of vision and images in the peripheral area. In other words, it is suggested that the brain can attend to activity in these regions independently and simultaneously. While the auditory sense has generally been assumed to have relatively uniform performance in all directions for external free-field sources, it is possible that one might exploit center versus surround listening for the processing of independent information connected to unrelated tasks. The authors therefore attempted to compare the results of a dual auditory task with a dual visual task. In the dual auditory task listeners were asked to speak one of three similar words presented over a headset in the center mode and to identify the direction of one of three pink noise bursts presented left, right. or on the line of sight in the surround mode. Surround mode stimuli were panned using HRTFs based on the KEMAR manikin. In the dual visual task, subjects were required to speak a word presented at a central point in the visual field or locate a large diamond-shaped object presented in the left, right, or upper visual periphery. The tasks were undertaken by five university students who were told that both speedy and accurate performance was required and that they should attend to both tasks equally. Since the response method was spoken for one mode and manual (push button) for the other, it was possible to evaluate situations in which either one or both responses were required together.

Overall conclusions suggested that a manual button press can be done faster than a verbalization response and that responses to auditory stimuli were faster than the corresponding responses to visual stimuli, no matter what the mode of presentation. The authors determined "interference times," which are the difference in response times between concurrent center/surround presentations and separate presentations (the time penalty for trying to do two tasks at once). The increase in performance time in the dual mode averaged 51 ms, while the highest value was 200 ms. It was found that for some individuals dual tasks could be performed with very little additional time compared with the single tasks (interference times of 10 to 20 ms), suggesting the possibility of independent cognitive processing for center and surround stimuli. Overall there was less interference for auditory stimuli than for visual ones, leading the authors to propose the consideration of the auditory sense as a strong candidate for inclusion in the design of human interfaces for decision support systems.

MULTIMODAL AURALIZATION USING AN AUDIO-TACTILE DISPLAY

Abercrombie and Braasch analyze the factors influencing the design of an audio-tactile display for conducting experiments on the dual modalities of hearing and vibration sensing. Their paper, "A Method for Multimodal Auralization of Audio-Tactile Stimuli from Acoustic and Structural Measurements" (paper 7867), shows how it is necessary to measure source and environment conditions independently. In the case of their system, the audio signal from a musical source can be recorded using a Head Acoustics dummy head and the structural vibration generated by the source is recorded using an accelerometer. In order to decouple the musical instrument ➥

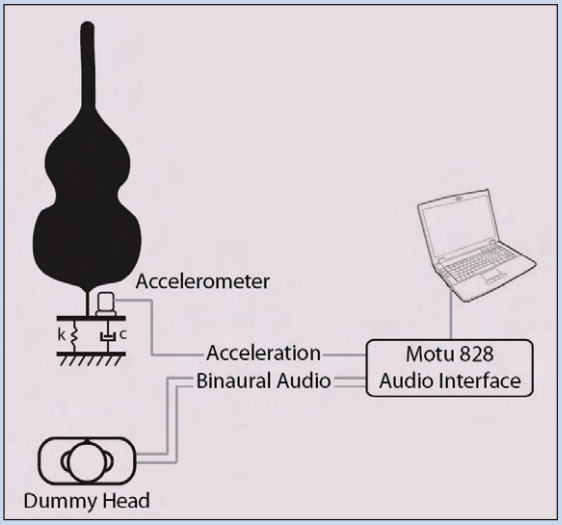

Fig. 1. Configuration of a system for capturing sound and vibration information from a musical instrument (double bass in this case), using a dummy head and accelerometer. The platform is decoupled from the floor using a spring/damper combination. (Figs. 1 and 2 courtesy Abercrombie and Braasch)

from the supporting floor, an isolation platform is used that is designed to ensure that vibrations above the instrument's fundamental frequency of 41 Hz are not affected by the floor structure. The basic configuration is shown in Fig. 1. A parameter known as accelerance is measured for the structural environment within which the instrument is played. Essentially this is measured by hitting the floor at a point where the instrument would be played and then measuring the acceleration output at the point where the listener would be situated. A laser doppler vibrometer is suspended from a step-ladder with a bungee cord and the output is time-integrated using postprocessing. A frequency range of 35 to 100 Hz is considered feasible (and fulfils the requirements for whole-body vibration standards), although extension above this range will be needed for future research on structural radiation of sound.

Tactile signals are reproduced using a motion platform made of two layers of 20-mm plywood laminated together and supported by a frame, as shown in Fig. 2. Challenges are associated with the avoidance of resonances and bending waves in the platform, and this can be compensated to some extent by calibrating it in the frequency domain. The authors claim that attention to the frequency response of the tactile part of the display is as important as, for example, the calibration of a critical listening room. An electrodynamic shaker is used, similar to that used in experiments by others on whole-body vibrations in cars, known as the Buttkicker LFE. One of the problems with audio–tactile displays is that the vibration platform on which the listener sits can also radiate sound energy when it is moving, and the actuating mechanism can make some noises as well. For controlled experiments one would wish to isolate the variables involved, so the authors propose to use closed-ear headphones for audio replay (Sennheiser HD280) because of their high levels of sound isolation.

In conclusion, we find that tactile signals in real environments vary widely between listener positions in the room, much more than the corresponding acoustic conditions. It is said that a listener who walks across a single floor construction may experience up to 26 dB difference in frequency-weighted acceleration response to the same source. Tactile signals tend to lag audio signals with delays of up to 74 ms.

MUSICAL BIOFEEDBACK FOR BREATHING REGULATION

Musical biofeedback was studied by Siwiak et al. and reported in "Catch Your Breath—Musical Biofeedback for Breathing Regulation" (paper 7870). Motivated by the challenge of capturing useful data in medical 4D-CT (four-dimensional computational tomography) scans, the authors looked for ways of steadying a patient's breathing using musical tempo and harmonic rhythm. Apparently breathing irregularity is one main reason why patient scans can be disrupted, leading to poor quality images. This can be particularly important when imaging human lungs in order to diagnose lung cancer, as patient nervousness can interfere with the imaging process. In the prototype developed by the authors, breathing is monitored by a camera and a reflective cube that optically scans the chest motion. A musical accompaniment is synchronized to the breathing rate, which adapts continuously so that the accompaniment and principal parts of the music align in a harmonious way when the patient's breathing aligns with a target breathing pattern. Results were good, providing a 55% reduction in respiratory variations and a 70% reduction in period variations.

Fig. 2. Prototype vibration platform used for an audio-tactile display

Artistic applications of the system were also attempted in public spaces such as an art museum, conference center, and trade fair. They also plan to examine the potential for use with 4D PET imaging in the medical domain, as well as to investigate whether or not the system can be useful in methods used to promote relaxation and reduce blood pressure. They also mention potential applications in training young musicians.

AN AUGMENTED REALITY AUDIO GAME PROTOTYPE

Augmented reality audio (ARA) is the term coined for systems that integrate natural sound and reproduced sound so as to enhance or complement natural listening. Synthetic or reproduced sounds can be heard along with natural sounds, usually using binaural technology and headphones, in such a way that the natural sounds are either elec-

tronically mixed with the artificial ones or that the headphones are acoustically transparent to natural sounds. In "Eidola: An Interactive Augmented Reality Audio-Game Prototype" (paper 7872), Moustakas et al. show how ARA can be used to introduce 3-D audio into game play that does not rely on computer-generated visual information. Eidola is effectively an audio-based game in which the player has to enter a room and defeat a number of creatures that move within a 3-D space. In order to provide some markers and boundaries, as well as a more natural game environment, within the game room are placed a number of real physical objects, which emit sounds themselves and constrain the movement of the virtual creatures by their sound volume.

An ARA headset worn by the player is equipped with wireless audio reception operating in the 2.4-GHz range, as well as head-tracking based on a modified computer joystick, as shown in Fig. 3. The output of the head-tracker was relayed back to the central controller using a Bluetooth link, which was found to have adequate bandwidth and latency for real-time transmission of the data, as well as carrying information about the user's "trigger." Head tracking was used to measure movements of the user's head for use in accurate binaural rendering, and a video camera was used to identify the user's absolute position in the room. In-ear phones are combined with miniature microphones that are intended to capture natural sounds binaurally, so that they can be mixed with the synthetic sounds of the game. The architecture of the prototype system is shown in Fig. 4.

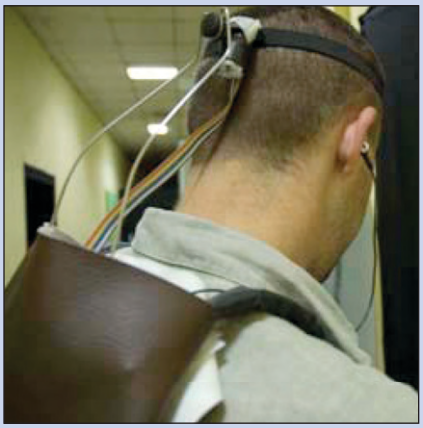

Fig. 3. Head-tracking system used in the Eidola ARA game (Figs. 3 and 4 courtesy Moustakas et al.)

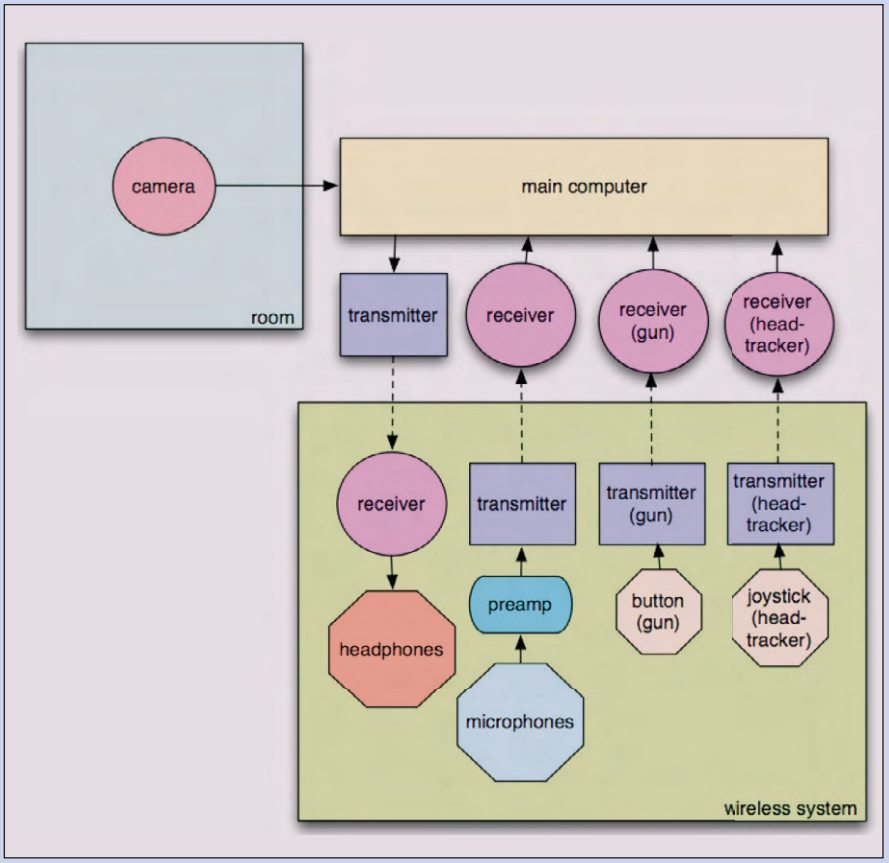

Fig. 4. Architecture of the prototype Eidola ARA game system

Demonstrations and basic tests of the game prototype revealed that users generally found the installation environment suitable for creating a gameplay atmosphere, while the binaural technology was found to be realistic and accurate by experienced listeners. The scores of experienced listeners were found to be higher than those who were unfamiliar with binaural technology. Having to wear the equipment was acceptable but slightly restricted movement, particularly walking and at the start of the session. Players appeared to adapt quite rapidly to the augmented environment and accepted it as an advanced type of computer game.

LOUDSPEAKERS FOR LARGE FLAT DISPLAYS

A problem with large flat displays is that it's hard to get a good stereo image for listeners in various positions when you have loudspeakers mounted on the sides of the display. The image may be adequate for listeners seated in the hot spot, when a strong phantom center can be formed, but for those off to one side the image quickly shifts into the nearest loudspeaker. The main reason for this is the relatively small time delay between the channels needed to shift the image to one side in conventional stereo (not much more than one millisecond), which can be introduced if a listener is only a foot or two off center.

In "A Loudspeaker Design to Enhance the Sound Image Localization on Large Flat Displays" (paper 7866), Nava et al. attempt to remedy this situation for an immersive teleconferencing system known as t-Room by introducing a physical means of modifying the radiation pattern of loudspeakers using rigid barriers attached to the side of large displays. This acts to modify the level of sound reaching a listener's ears according to his position, as shown in Fig. 5. Rather like earlier time–level trading loudspeaker designs introduced in the past, such as Canon's Wide Imaging Stereo, a listener closer to the right-hand loudspeaker will experience an attenuated version of that loudspeaker's output ➥

compared with a listener in the center, at least at high frequencies. The polar pattern is such that it favors listeners on the opposite side of the listening position range to the loudspeaker. In this case it is suggested that Listener A, for example, will experience an earlier sound from the right loudspeaker, which would normally result in the image being shifted to that side, but this is somewhat compensated by the reduction in level caused by the physical barrier. Simulated and measured results suggested that the design was indeed achieving an effect similar to the one intended, with a frequency-dependent polar pattern that favored the opposite side to the sounding loudspeaker. It was noted that the dimensions of the structure will affect the frequency range over which the effect is most prominent, presumably needing to be fairly large to have a significant effect at low frequencies.

Listening tests using panned noise bursts at five different locations from left to right, panned using a combination of time and level difference between the channels, suggested that listeners in the three positions shown in Fig. 6 perceived images in the intended positions, although the correlation between intended and perceived positions was found to be highest in the central listening position. A prototype display employing this arrangement was constructed using a real 65-inch LCD panel and six loudspeakers with rigid barriers, as well as additional loudspeakers top and bottom to improve the low-frequency response. A video representation of a person talking while moving around was aligned with the sound image of the voice using stereoscopic video cameras to track the position of the mouth. The authors suggest that the perception of sound images is as good as that obtained with conventional loudspeaker for those sitting at the center, while those off center are able to localize the images almost as well.

WAVEFIELD SYNTHESIS IN INTERACTIVE SOUND INSTALLATIONS

Wavefield synthesis has the potential to create focused sources, that is virtual sound images with an almost holophonic character, which can be located in front of a loudspeaker array and remain in a constant position when listeners move. This aspect of its capability was used by Leslie et al. in their interactive sound installation known as GrainStick, which was a collaboration between the composer Pierre Jodlowski and the European project SAME (Sound and Music for Everyone Everyday Everywhere Everyway), and discussed in paper 7871, "Wavefield Synthesis for Interactive Sound Installations." Image spatialization of virtual percussive instruments was controlled using the gestures of one or more people tracked by means of a six-camera ARtrack motion camera system. Nintendo Wii controllers were used to record the acceleration of users' gestures in order to enable the identification of "kick" gestures that could be used to hit virtual percussive instruments. The instruments thus hit were rendered by the WFS system at the point in space where the user had gestured, so that the sounds appear to follow the user's hand.

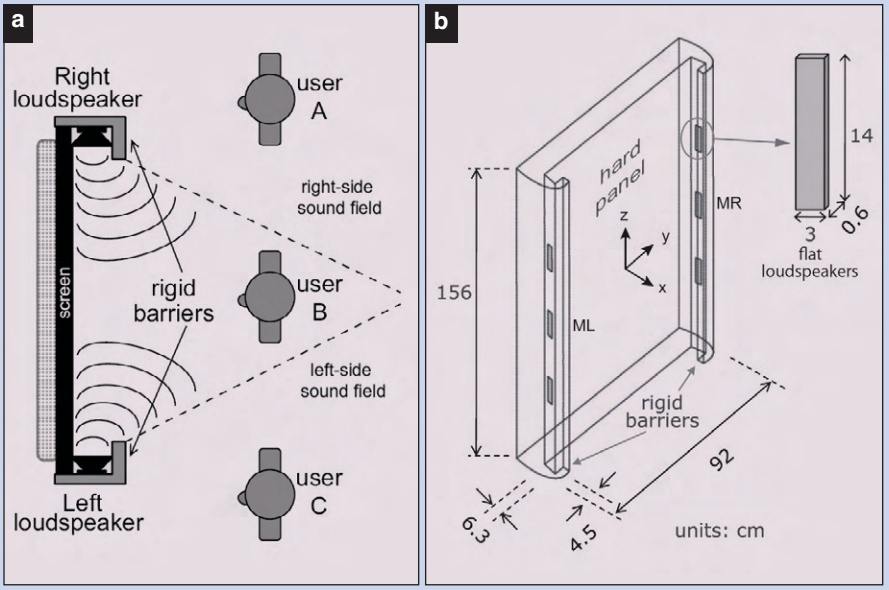

Fig. 5a. Loudspeaker configuration to enhance the localization of sound images on flat panels. 5b. Dimensions of the experimental prototype. (Figs. 5 and 6 courtesy Nava et al.)

One of the virtual percussion instruments concerned was a form of synthetic rainstick, created by triggering small sound samples from a corpus based on the angle between the controllers. The authors explain that the resulting effect is one of a natural-feeling and sounding recreation of small grains spilling from one side of a container to another, which are being rendered as focal sources using the WFS system. "The rendered grains travel along the length of the WFS array as they tumble from one end to the other in the virtual instrument, integrating synthesis control with spatialization control," they say.

A photograph of two users playing with the system is shown in Fig. 7. They can collaborate in controlling the virtual rainstick, the authors suggest, tipping it in various ways and spreading the sound grains and percussive sounds across the soundfield. Customized filters designed by IRCAM enabled the creation of focal sources with a controlled radiation pattern and apparent directivity, under the control of three computers driving a total of 48 loudspeakers spaced 15 cm apart on 12 multiactuator panels. The total array length was 7.2 meters. The system was designed by Sonic

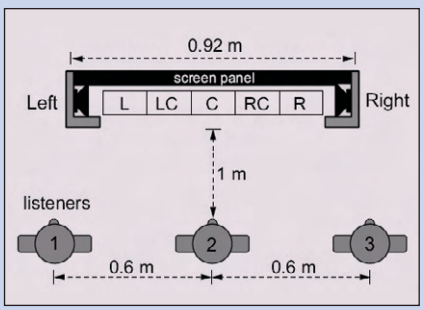

Fig. 6. Listener and sound-image positions for a listening experiment to localize sounds

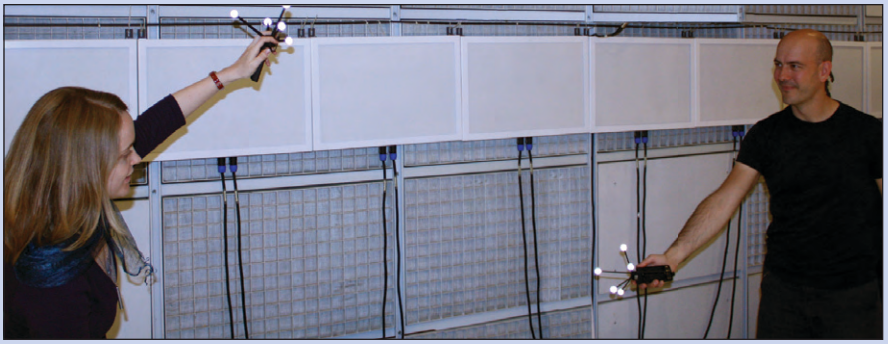

Fig. 7. Two users collaborate in controlling a spatialized virtual instrument. Wavefield synthesis panel loudspeakers are shown in the background. (Courtesy Leslie et al.)

Emotion, and position and directivity information from the handheld controllers were sent using UDP messages to the networked computers. In the setup shown in Fig. 7, focal sources could be rendered within a specified region bounded by the ends of the array and some meters in front of the array. A limiting factor is that the focal sources have to be between the array and the listeners—they cannot be further out into the room beyond the listener for a system with only a linear array on one side.

A primary advantage of this type of rendering approach, the authors claim, lies in the fact that sources remain stable as listeners move around. In this way installations and system designs can be made allocentric, which means that they are designed around a common coordinate system. This is different from more traditional forms of rendering that are inherently egocentric, or based on the subject's head position as a reference point for the coordinate system.

CONCLUSION

Novel approaches to spatialization—coupled with motion tracking, biofeedback, or tactile information—provide new ways to integrate reproduced sound with human activity. It may be possible, it seems, for the brain to process different spatial audio streams concurrently, leading to the potential for use in information display systems using more than one spatial modality. There is the potential for therapeutic applications of audio as well as implementation in new forms of entertainment or art. The seamless integration hoped for between spatialized visual and auditory information in interactive systems is gaining a greater degree of success as research moves forward the boundaries of multimodal technology.

Editor's note: The papers reviewed in this article, and all AES papers, can be purchased online at <www.aes.org/publications/preprints/search.cfm> and <www.aes.org/journal/search.cfm>. AES members also have free access to past technical review articles such as this one and other tutorials from AES conventions and conferences at <www.aes.org/tutorials/>.