AudioStreamer: Exploiting Simultaneity for Listening

Chris Schmandt
MIT Media Lab
20 Ames St.
Cambridge, MA 02139, USA
E-mail: geek@media.mit.edu

Atty Mullins
MIT Media Lab
20 Ames St.
Cambridge, MA 02139, USA
E-mail: mullins@media.mit.edu

INTRODUCTION

AudioStreamer exploits peoples' ability to separate the mix of sounds that arrive at our ears into distinct sources to more effectively browse multiple simultaneous channels of real-time or stored audio. AudioStreamer's listener interface enhances our ability to selectively attend to the source of greatest interest by making it acoustically prominent. It also augments our ability to perceive events in the audio channels which are out of focus by auditorially alerting us to salient events on those channels. The main contributions of AudioStreamer are the use of spatial separation and simultaneous listening for audio document retrieval and modeling listener interest to enhance the effectiveness of simultaneous listening.

SIMULTANEOUS AUDIO PRESENTATION

The user of AudioStreamer listens to three simultaneous sound sources. A number of acoustic cues allows the separation of a mix of sounds into distinct sources: these include location, harmonics and frequency, continuity, volume, and correlation to visual events [2]. The sources used in this project, audio news programs, naturally exhibit some of these cues, as the news anchors are different speakers, possibly of different sex and accents (one source is the BBC World Service). AudioStreamer adds synthetic locational cues to enhance the listener's ability to separate the component sound sources.

The digital audio sources are routed to a Crystal River Beachtron audio card in a server PC. The Beachtron uses a head related transfer function (HRTF) to model many of the acoustic cues we use for spatial localization of sound. The Beachtron can localize and mix up to four monaural audio inputs into a stereo output, which is meant to be listened to with ordinary headphones.

Spatialized audio has been utilized most heavily in virtual reality environments, usually correlated with images viewed in a head-mounted display. AudioStreamer uses this technology to create an audio-only browsing environment, presenting three speech sources arrayed about the listener's head. The sources are arrayed in the horizontal plane of the head, with one source directly in front and the others offset 60 degrees on either side. The angular distance between sources was chosen to be large enough to allow easy perceptual separation of the sources, but still limit the time it takes to switch from one to the other, which is proportional to angle.

AudioStreamer takes advantage of the "cocktail party effect"; surrounded by spatially disparate conversations, a listener can selectively attend to one of them. AudioStreamer further enhances the listener's ability to browse between the three audio programs by enhancing the signal that is the current focus of interest. It attempts to provide an invisible interface, wherein the very act of listening is the interface. We often turn our faces towards the source of a sound to better localize and attend to it; AudioStreamer uses this gesture to imply interest. Head motion is detected using non-contact sensors built into a chair. Moving the head towards the sound increases the gain of that source 10 dB; the louder source now dominates.

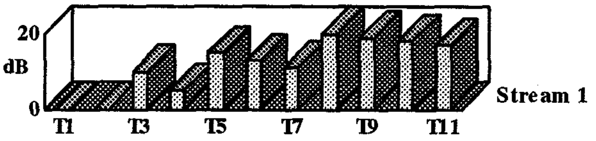

Figure 1. The interaction of time and gain. The listener selects the stream at T3, T5, and T8.

Interest is not static, but rather diminishes over time. AudioStreamer models listener interest by decaying the gain which is applied when the listener focuses on a particular source. When the gain starts to fade, the listener can turn towards the source again, and go to a higher state of interest (Figure 1). For each of the first three levels of interest, the selected sound is increasingly louder, and the decay time constant is proportionally longer. At the highest level of attention, the other sources are silenced (until a story boundary, see below).

CHI' Companion 95, Denver, Colorado, USA
© 1995 ACM 0-89791-755-3/95/0005...$3.50

GETTING THE LISTENER'S ATTENTION

While the listener is attending to one channel, salient events, i.e., interesting discourse, may occur on an another channel. AudioStreamer alerts the listener to these events because, as is discussed in the next section, very little information "leaks" through from non-attended channels.

But how can we know when a "salient event" occurs in audio? AudioStreamer uses techniques previously described as semi-structured audio: without knowing the actual words that were spoken, acoustic cues can give hints as to content. Related projects explored detection of segments the speaker meant to emphasize based on pauses and pitch range variation. Although these appear to work well for lectures and informal speech, they failed due to the strict timing structure and intonational patterns of broadcast news. Pause structure alone worked very well on 15 minute BBC newscasts, but less well on longer programs and failed on North American programs.

So, for the television news programs we used closed-caption information; closed-captioning is a rough text transcription of the spoken program. For the BBC radio source, we used a speaker differentiation algorithm in addition to pauses to find story boundaries.

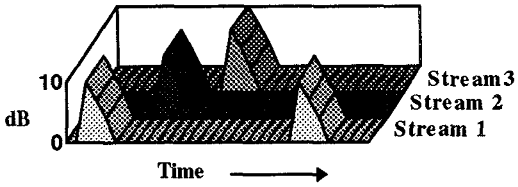

Figure 2. The system suggests possible points of interest by increasing the gain at story boundaries.

AudioStreamer uses story boundary information to attract the listener's attention. At a story boundary, the channel emits a 400 Hz, 100 msec tone, and the gain of the audio is increased 10 dB. This attention indication decays rapidly (5 seconds). The effect is that while listening to one story, a listener is mildly interrupted by awareness of a change to something potentially more interesting on another channel, and can easily switch attention back and forth (Figure 2). This can be accomplished quickly enough that little information is lost at the primary channel.

PERFORMANCE AND EVALUATION

AudioStreamer is one of a series of projects exploring various means of browsing and scanning audio, all with the goal of acquiring information from recorded speech in less time than real-time. SpeechSkimmer [1] uses time compression and frequent manual input to facilitate random access. AudioStreamer defines a complementary approach, wherein serial time compression is replaced by multi-channel presentation, and the manual user interface is replaced by possibly more spontaneous natural head gestures. Both place heavy cognitive demands on the listener, and are meant for active involvement.

AudioStreamer tests the viability of simultaneous listening for audio browsing. Enhancement to listener selective attention is very powerful, but how much does one hear on the secondary channels? Most of the experimental evidence [3] suggests that relatively little leaks past short term memory, and that listeners may not even be aware of what language is being spoken on secondary channels. AudioStreamer confronts this problem in a unique way, by eliciting attention shifts at potentially interesting times. This technique seems to be effective, but merits more formal evaluation.

How well does AudioStreamer work? Although we have not performed any formal user studies, informal evaluations are rather bimodal. While some listeners find the interface exciting and potentially valuable, especially those already used to dealing with large quantities of audio information, others find it confusing, tiring, and impenetrable. Certainly simultaneous listening will have to be carefully situated with respect to a particular task. As with listening to synthetic or time compressed natural speech, we expect an adaptation effect whereby listeners will gain facility at this style of listening with exposure; this has been true of the developers of this interface.

ACKNOWLEDGMENTS

Deb Roy wrote the speaker differentiation software, and Chris Horner wrote the closed-caption software. Tom Zimmerman designed the non-contact sensors. Crystal River provided the Beachtrons. This work was supported by the MIT News in the Future consortium and Sun Microsystems.

REFERENCES

1. Arons, B., SpeechSkimmer: Interactively Skimming Recorded Speech. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, ACM Press, Nov. 1993.

2. Bregman, A.S., *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA: MIT Press, 1990.

3. Norman, D.A., *Memory and Attention: An Introduction to Human Information Processing.*, New York, NY: Wiley, 1976.