Design of Spatialized Audio in Nomadic Environments

Nitin Sawhney and Chris Schmandt
Speech Interface Group
MIT Media Laboratory
20 Ames Street, Cambridge, MA 02139
{nitin, geek}@media.mit.edu

ABSTRACT

This paper describes an on-going research project at the MIT Media Lab, exploring the use of audio as a primary modality for nomadic computing applications. We are developing a framework for use on a wearable audio platform, *Nomadic Radio*, that presents dynamic information within a spatialized audio environment. The contextual state of the *nomadic listener* indicated by time of day, physical positioning, scheduled tasks, message content, and level of interruption is used to present relevant information in the user's listening space. In this paper we will consider issues related to auditory presentation and spatial techniques for awareness and browsing of audio messages on wearable computing.

INTRODUCTION

In an information rich environment, people access a multitude of content such as news, weather, stock reports, and data from a variety of information sources. Much of this information is now accessible from desktop computers. People also increasingly communicate via email, fax, and voice mail. It is desirable to provide seamless access to personal information and communication services to individuals situated away from their desktops (i.e. *nomadic listeners*) such as in meetings, classes or simply while walking. Such services should be made available on-demand in a passive and unobtrusive manner, based on the user's level of attention and interruptability. The MIT Media Lab's Nomadic Computing Environment [1] enables subscribers to manage personal information via fax, pagers and telephony access using digitized audio and synthesized speech. Pagers and cellular phones provide remote access to information, yet such devices offer extremely low-bandwidth for communication and the interface does not afford rich delivery of information content. Personal Digital Assistants (PDAs) offer personal applications in a smaller size, yet they generally utilize pen-based graphical user interfaces, which are not ideal when the user's hands and eyes are busy such as while driving. Hand-held [2, 3] and mobile [4, 5] audio devices with localized computing and richer interaction mechanisms certainly point towards audio interfaces and networked applications for a new personal information platform, *Wearable Audio Computing* (WAC) [6]. Auditory displays can be used to enhance an environment with timely information and provide a sense of peripheral awareness [7] of people and background events. A combination of wearable auditory and tactile interaction can provide users a means of unobtrusively augmenting a physical environment. Our goal is to provide access to messaging services such as email, voice mail and news on a wearable device, with audio as the primary interaction modality.

Wearable computers are generally always operational and can monitor the user and the physical environment on a continual basis. Longer-term interaction with such devices provides an opportunity to personalize information and the interface, as the system learns and adapts to the user's daily tasks and preferences. Yet most wearable computers today derive their interfaces from concepts in desktop computing such as keyboards, pointing devices, and graphical user interfaces. If wearable computers are expected to become as natural as clothing, we must consider the role of audio and tactile interaction as the primary interface for wearable computing. Speech input is a natural means of interaction, yet the user interface must be carefully devised to permit recognition and recording [2] on a wearable device. Speech and button input can be combined to provide users more flexibility for interaction in different situations. Button input allows operation of the WAC when background noise or social protocol constrains the use of speech recognition. Buttons are more suitable for navigation or tasks requiring fine control such as speed and volume adjustment. In contrast, speech input is preferable to buttons for complicated tasks such as "Send response to John" or when the user's hands and eyes are busy.

Audio output on wearables requires use of speakers worn as headphones or appropriately placed on the listener's body. Headphones are not entirely suitable in urban environments where users need to hear other sound sources such as traffic or in offices where their use is considered anti-social as people communicate frequently. In these situations speakers worn on the body could instead provide directional sound to the user (without covering the ear), yet they must be designed to be easily worn and least audible to others. Audio output is sequential and exists only temporally; the ear cannot browse around a set of recordings the way the eye can scan a screen of text and images. Hence techniques based on time-compressed speech, simultaneous audio and spatialization must be considered for browsing audio-based messages more efficiently.

NOMADIC RADIO: WEARABLE AUDIO MESSAGING

Nomadic Radio [8] is being developed as a unified messaging system that utilizes spatialized audio, speech synthesis and recognition, on a wearable audio platform. Messages such as hourly news broadcasts, voice mail, email and weather reports are downloaded to the device throughout the day. Selected messages are presented in the user's listening space, based on her context and desired level of awareness or interruptability. We are incorporating timely message filtering [9] to selectively present the appropriate messages, coupled with user location and environmental context.

The current system operates primarily as a wearable audio-only interface, although a visual interface is used for development purposes. A combination of speech and button inputs are used to control the interface. Textual messages such as email, weather forecasts, and stock reports are delivered via synthesized speech. Users can select a category such as news or email and browse messages in it sequentially as well as save and delete them from the server. In one usage scenario, the listener begins hearing an ABC news summary at a certain time of day, and moments later a voice message arrives reminding her of a meeting later that day while the news broadcast fades down. As the system gains location awareness (discussed later in this paper), we envision a scenario where the listener's location context enables the system to provide relevant messages as needed. For example as the user moves close to a particular room, she may hear a voice message left by a colleague or more importantly she is reminded of a meeting if she is not in a desired location at a specific time. In *Nomadic Radio*, we utilize the metaphor of *radio* to present personalized information as *active broadcasts* delivered within the user's listening environment. Several such broadcasts can be presented simultaneously as spatialized audio streams, to enable the listener to better segregate and browse multiple information sources. We will now focus on issues related to techniques for spatial presentation of audio messages and consider radio as a metaphor for browsing audio.

SIMULTANEOUS AND SPATIAL LISTENING

People using wearable devices must primarily attend to events in their environment, yet need to be notified of background processes or messages. Speech and music in the background and peripheral auditory cues can provide an awareness of messages or signify events, without requiring one's full attention or disrupting their foreground activity. Audio easily fades into the background, but users are alerted when it changes [10]. It is possible for listeners to attend to multiple background processes via the audio channel as long as the sounds representing each process are distinguishable. This well known cognitive phenomenon, called the "Cocktail Party Effect" [11], provides the justification that humans can in fact monitor several audio streams simultaneously, selectively focusing on any one and placing the rest in the background. A good model of the head-related transfer functions (HRTF) permits effective localization and externalization of sound sources [12]. Yet the cognitive load of listening to simultaneous channels increases with the number of channels. Experiments show that increasing the number of channels beyond three causes a degradation in comprehension [13]. Bregman claims that stream segregation is better when frequency separation is greater between sound streams [14]. Arons suggests that the effect of spatialization can be improved by allowing listeners to easily switch between channels (providing perceptual handles on each channel) and pull an audio stream into focus as well as allowing sufficient time to fully fuse the audio streams [15].

A spatial sound system can provide a strong metaphor by placing individual voices in particular spatial locations. The effective use of spatial layout can be used to aid auditory memory. The *AudioStreamer* [16] detects the gesture of head movement towards spatialized audio-based news sources to increase the relative gain of the source, allowing simultaneous browsing and listening of several news articles. Kobayashi introduced a technique for browsing audio by allowing listeners to switch their attention between moving sound sources that play multiple portions of a single audio recording [17]. An audio landscape with directional sound sources and overlapping auditory streams (*audio-braiding*) can also provide a listening environment for browsing multiple audio sources easily [18]. On a wearable device, spatial audio requires the use of headphones or shoulder mounted directional speakers. In noisy environments there will be a greater cognitive load to effectively use spatial audio, yet it can help segregate simultaneous audio streams more easily. Here the exact location of the sound is less important, but can provide cues about aspects of the message such as its category, urgency and time of arrival.

Designing an effective spatial layout for a diverse set of messages requires a consideration of their content and scalability issues. In *Nomadic Radio*, messages such as email, voicemail, news and traffic reports must be presented to the listener throughout the day. Each message category has a different level of urgency and messages within that category may themselves be considered timely or higher priority. One approach is to map messages in auditory space based on content category. Hence news would be played in one direction and all voice messages in another. Yet this does not scale well as many new messages arrive and the spatial layout only indicates categorical information. The first few seconds of a message can allow users to distinguish their category quite easily. For example listeners can become quite adept at distinguishing a news summary from a voice message after hearing just the first second or so of the audio.

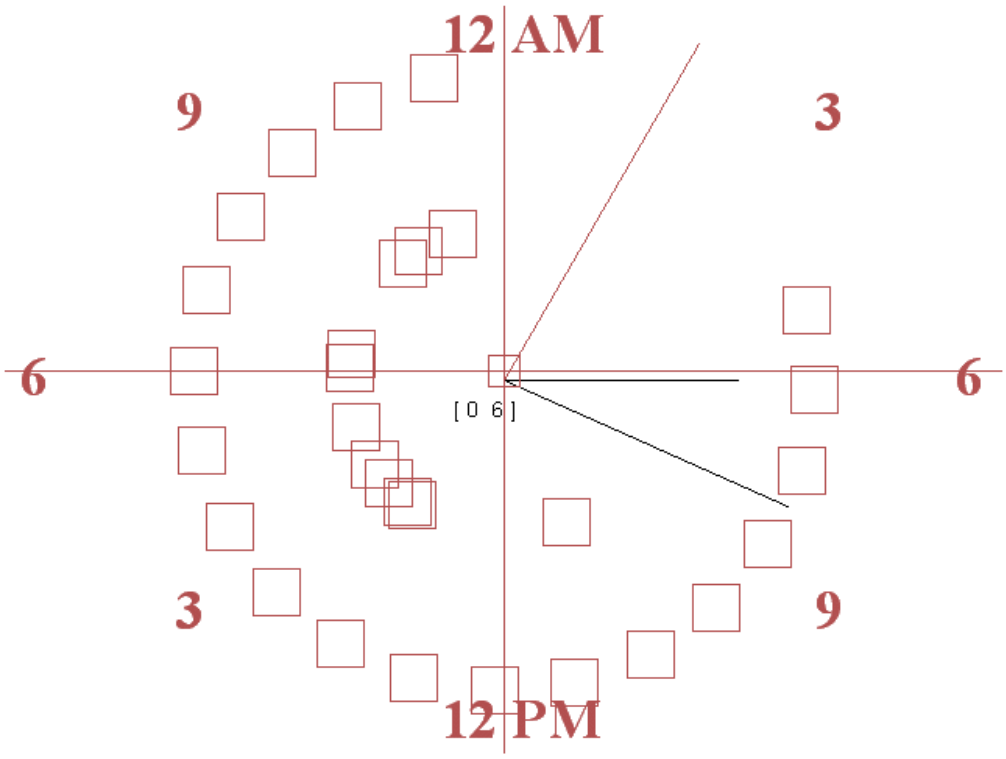

Figure 1: Time-based spatial layout of hourly news (outer orbit) and voice messages (inner orbit). The cross-hair represents the center of the user's listening space, acting as a *virtual ear* to browse audio in the soundspace.

Each message arrives at a different point in time, hence its date and time of arrival provide a unique parameter for spatial layout. Thus a suitable approach is to utilize arrival time to position messages in chronological order around a listener's head (Figure 1). A spatial clock can permit messages arriving at noon to be positioned in the front and those at 3:00 PM on the right and so on. A twelve hour clock does not scale well for messages arriving throughout the day. Here messages arriving after 12:00 PM will overlap with existing ones from the AM. Auditory cues, played at the start of the message can indicate AM or PM, yet a better solution is to use a twenty four hour clock, that can better represent messages for an entire day. Using such a metaphor all messages arriving during the day occupy a unique position in the listening space. The listener can discern the approximate time of arrival based on the general direction that the message is heard. The message category also determines the distance of the messages from the listener, indicating general importance of the category. For instance all voice messages are positioned closer and news messages placed further away.

In *Nomadic Radio*, spatial listening is utilized for message playback in three specific modes:

Broadcasting: When new messages arrive, they are *broadcast* to the listener from a specific spatial location. These messages are continuously heard in the *background* and fade away if the user does not pay attention to them (i.e. activate them via selecting a button). This is based on the metaphor of traditional radio broadcasting, where listeners passively listen to news stories and only pay attention when a relevant article is heard.

Browsing: This is an active form of listening where users can select a category and browse sequentially through all the messages, playing each one as needed (using forward/back buttons on a wireless mouse). When an important or desirable message is heard the user can stop and listen to the entire message in the *foreground*. This is similar to the metaphor of switching stations on a radio until a station playing desirable music is found.

Scanning: Sometimes listeners want to get a preview of all their messages quickly without manually selecting and playing each one. This is not unlike a *scan* feature on modern radio tuners that allows users to alternatively hear each station for a short duration. In *Nomadic Radio, message scanning* cycles through all messages by moving each one to the center of the listening space for a short duration of time, and fading it out as the next one starts to play. All messages are played sequentially in this manner, yet there is some overlap as one message fades away and the next one begins to play. This

simultaneity allows for more efficient browsing while presenting the important content of the messages easily (i.e. identity of caller or main news headlines). The scanning algorithm ensures that the messages are quickly brought into perceptual focus by pulling them to the listener rapidly, yet the messages are pushed back slowly to provide an easy fading effect as the next one is heard. As the message is pulled in, its direction is maintained allowing the user to retain a sense of message arrival time. This spatial continuity is important in discriminating and holding the auditory streams together [15].

Several levels of awareness are used to determine the form of message delivery. Since users can be actively engaged in conversations or meetings, they may not wish to be distracted by audio playing at certain times of the day. Instead auditory cues or message previews can be provided. Currently three levels of awareness can be set by the user. A unique auditory cue is always played before the message to distinguish the message. *Message Summary* provides the sender and subject of email messages (via speech synthesis) or plays the first 2-3 seconds of a voice message or audio news summary. *Message Preview* provides a short preview of an email message (first 100 characters) or plays a fifth of the audio message. Finally, *Message Body* plays the entire message in the background or foreground based on the urgency of the message. Scanning and message awareness modes can be activated by the user via speech commands. Eventually the system may learn to predict the appropriate awareness level based on the contextual state of the user.

During informal demonstrations of *Nomadic Radio* (on desktop PCs), listeners could typically segregate at least two to three audio streams of synthesized speech or digital audio content, and browse messages easily. We must continue to evaluate the usability of these interaction techniques on the wearable platform for effective presentation and browsing of audio messages.

IMPLEMENTATION

Nomadic Radio consists of client and remote server components that communicate over the wireless LAN. The *Nomadic Clients,* developed in Java, operate on Pentium-based wearable PCs (worn on the waist). The *Soundbeam Neckset* (Figure 2), worn around the neck, is being modified for audio I/O from the Wearable. The *Neckset* is a patented research prototype originally developed by *Nortel* for use in hands-free telephony. It consists of two directional speakers, mounted on the user's shoulders, and a directional microphone placed on the chest. The system uses speaker-independent speech recognition based on AT&T's Watson API. We are currently incorporating an adaptive speaker dependent speech recognizer, developed at the Media Lab [19]. It will allow creation and use of different contexts, each containing a unique lexicon of recognizable words. A button on the *Neckset* can activate speech recognition or deactivate it in noisy environments. Spatialized audio is rendered in real-time and delivered to the *Neckset,* using a Java interface to Intel's *RSX 3D* audio libraries. Tactile input is provided by a three-button wireless mouse. On the wearable device, IR-based receivers will provide positioning data using the *Locust Swarm*, a distributed IR location system at the Media Lab.

Figure 2: The wearable audio device, the *Soundbeam Neckset,* with directional speakers and microphone

The current architecture relies on server processes (written in C and Perl) running on Sun SPARCstations that utilize the telephony infrastructure in the Media Lab's Speech Interface group [1]. The servers extract information from live sources including voice-mail, email, hourly updates of ABC News, weather and traffic reports. The clients, when notified, download

the appropriate text/audio files stored on the web server. Text files are converted to synthesized speech (using the AT&T *Watson* Speech API) and rendered as digitized audio in the spatial environment of the listener.

CONCLUSIONS AND FUTURE WORK

Messaging on a wearable device can be enhanced via simultaneous and spatial auditory presentation. Using the metaphor of radio, several techniques for broadcasting and browsing audio-based messages have been developed. *Nomadic Radio* can be considered an active information agent that adaptively manages the user's listening space, based on their location, context of activity, prior listening patterns and desired level of interruption. Current work-in-progress includes integration of location awareness and refinement of the audio interaction based on field evaluations of the wearable audio device. Future work includes using timely message filtering and situational awareness on the wearable system via a classification of sounds in the environment. This would allow the system to use the environmental sound as a contextual cue for delivery of appropriate information, such as a train schedule if it heard the sound of a train approaching or to turn down audio sources if the listener was engaged in a conversation. The design of effective audio environments that are sensitive to nomadic listeners, presents a challenge for mobile and wearable computing devices today. We stress that attention to the affordances and constraints of speech and audio in the interface coupled with the physical design of the wearable audio device play an important role in determining how it will be adopted in social environments over prolonged usage.

ACKNOWLEDGMENTS

We would like to thank Alex Pentland, Deb Roy, Lisa Stifelman, Michael Jacknis, Brian Clarkson and Travell Perkins at the MIT Media Lab as well as Lisa Fast and Andre Van Schyndel at Nortel for their support of this project.

REFERENCES

1. Schmandt, Chris. "Multimedia Nomadic Services on Today's Hardware". IEEE Network, September/October 1994, pp12-21.
2. Stifelman, Lisa J., Barry Arons, Chris Schmandt, Eric A. Hulteen. "VoiceNotes: A Speech Interface for Hand Held Voice Notetaker". *Proceedings of INTERCHI'93*, April 1993.
3. Roy, Deb K. and Chris Schmandt. "NewsComm: A Hand-Held Interface for Interactive Access to Structured Audio". *Proceedings of CHI '96*, April 1996, pp. 173-180.
4. Stifelman, Lisa J. "Augmenting Real-World Objects: A Paper-Based Audio Notebook". *Proceedings of CHI '96*, April 1996.
5. Wilcox, Lynn D., Bill N. Schilit, Nitin Sawhney. "Dynomite: A Dynamically Organized Ink and Audio Notebook". *Proceedings of CHI '97*, March 1997, pp. 186-193.
6. Roy, Deb K., Nitin Sawhney, Chris Schmandt, Alex Pentland. "Wearable Audio Computing: A Survey of Interaction Techniques" Technical Report, MIT Media Lab, April 1997.
7. Mynatt, E.D., Back, M., Want, R. and Frederick, R. "Audio Aura: Light-Weight Audio Augmented Reality". *Proceedings of UIST '97 User Interface Software and Technology Symposium*, Banff, Canada, Oct 15-17, 1997.
8. Sawhney, Nitin and Chris Schmandt. "Nomadic Radio: A Spatialized Audio Environment for Wearable Computing." *Proceedings of the International Symposium on Wearable Computing*, IEEE, Cambridge, October 1997.
9. Marx, Matthew and Chris Schmandt. "CLUES: Dynamic Personalized Message Filtering". *Proceedings of CSCW '96*, November 1996, pp. 113-121.
10. Cohen, J. Monitoring background activities. *Auditory Display: Sonification, Audification, and Auditory Interfaces.* Reading MA: Addison-Wesley, 1994.
11. Handel, S. *Listening: An Introduction to the Perception of Auditory Events.* MIT Press, 1989.
12. Wenzel, E.M.. Localization in virtual acoustic displays, Presence, 1, 80, 1992.
13. Stifelman, Lisa J. "The Cocktail Party Effect in Auditory Interfaces: A Study of Simultaneous Presentation". Technical Report, MIT Media Lab, September 1994.
14. Bregman, Albert S. *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, 1990.
15. Arons, Barry. "A Review of the Cocktail Party Effect". *Journal of the American Voice I/O Society*, Vol. 12, July 1992.
16. Schmandt, Chris and Atty Mullins. "AudioStreamer: Exploiting Simultaneity for Listening". *Proceedings of CHI '95*, May 1995, pp. 218-219.
17. Kobayashi, Minoru and Chris Schmandt. "Dynamic Soundscape: Mapping Time to Space for Audio Browsing". *Proceedings of CHI '97*, March 1997.
18. Maher, Brenden. "Navigating a Spatialized Speech Environment through Simultaneous Listening and Tangible Interactions". M.S. Thesis, Media Arts and Sciences, MIT Media Lab, Fall 1997.
19. Roy, Deb K. and Alex Pentland. "Adaptive Multimodal Interfaces." *Proceedings of the Workshop on Perceptual User Interfaces*, Banff, Canada, October 1997.