Speaking and Listening on the Run: Design for Wearable Audio Computing

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

Abstract

The use of speech and auditory interaction on wearable computers can provide an awareness of events and personal messages, without requiring one's full attention or disrupting the foreground activity. A passive "hands-and-eyes-free" approach is appropriate when users need convenient and timely access to remote information and communication services. Nomadic Radio is a distributed computing platform for wearable access to unified messaging via an auditory interface. We demonstrate the use of auditory cues, spatialized audio, and speech I/O in the wearable interface for passive awareness, scaleable notification and navigation/control. The architecture is designed for wired audio wearables and has been extended for distributed wireless operation.

1: 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. People increasingly communicate via services such as email, fax, and telephony. Such a growth in information and communication options is fundamentally changing the workplace and "beginning to have a seismic effect on people's professional and personal lives" *(see the recent Pitney Bowes Study, April 8, 1997[1]).* A partial solution to information overload is to give people timely and filtered information, most relevant to the context of their current tasks. Seamless access to personal information and communication services should be made available to users in a passive and unobtrusive manner, based on their level of attention and interruptability.

Simple devices such as pagers provide a convenient form of alerting users to remote information. Such devices offer an extremely low-bandwidth for communication and the interface does not afford rich delivery of information content. Telephones are ubiquitous and cellular services offer mobility. Computer-based telephony services, such as *Phoneshell* [20], offer subscribers integrated access to information such as email, voice mail, news and scheduling, using digitized audio and synthesized speech. However telephones primarily operate on a synchronous model of communication, requiring availability of both parties, and high tolls for accessing services since a connection must be maintained while listening to the news or stock report. All processing must be done on the telephony servers, rather than the phone itself. The requirement of synchronous connection prevents the device from continuously sensing its user and environment.

Portable computing devices can download messages when not being actively used, allowing asynchronous interaction similar to email. Personal Digital Assistants (PDAs) offer the benefit of personal applications in a smaller size; however they generally utilize pen-based graphical user interfaces, which are not ideal when the user's hands and eyes are busy. Hand-held audio devices [24, 15] 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) [16]. Wearable auditory displays can be used to enhance an environment with timely information and provide a sense of peripheral awareness of people and background events.

Several wearable computing projects have considered the use of speech and audio in the interface. *Ubiquitous Talker* [14] is a camera-enabled system that provides information related to recognized physical objects using a display and synthesized voice, and accepts queries via speech input. A prototype augmented audio tour guide [3] presented digital audio recordings indexed by the spatial location of visitors in a museum. *SpeechWear* [17] enabled users to perform data entry and retrieval using speech recognition and synthesis. A speech-enabled web browser allowed users to access local and remote documents through a wireless link. *Audio Aura* [13] explored the use of background auditory cues to provide serendipitous information coupled with people's physical location in the workplace. In a recent paper [22], researchers suggest the use of sensors and user modeling to allow wearables to infer when users should be interrupted by incoming messages. They suggest an approach based on waiting for a break in the conversation to post a summary of an urgent message onto the user's heads-up display.

[1] *http://www.pitneybowes.com/pbi/whatsnew/releases/workers_overwhelmed.htm*

Proceedings of ISWC'98, International Symposium on Wearable Computing,
19-20 October 1998 in Pittsburgh, Pennsylvania. Copyright 1998 IEEE.

In this paper we consider a primarily non-visual approach to present timely information to listeners, both as an alternative to typical HMD-based systems as well as a secondary modality to complement current wearable systems. We recognize that an auditory modality may not serve as a general-purpose interface for all wearable applications, however it is better suited within certain domains of information (particularly for content that is intrinsically voice/audio) for specific user tasks and contexts (hands-eyes busy). We will consider the trade-offs regarding such issues in the next section.

Nomadic Radio has been developed as a wearable audio platform to allow active navigation of personal messages and timely information, as well as peripheral awareness via notifications. An integrated use of speech recognition, spatial audio, and synthetic speech along with ambient and auditory cues, provides rich forms of foreground and passive interaction. Continuous sensing of the user and environment as well as filtering and prioritization of incoming information allows the system to change its operating characteristics and dynamically scale the amount of information presented.

In this paper we discuss several research challenges for *wearable audio computing*: design of a hands-free and robust audio interface, development of a unified and extensible architecture for messaging and communication, as well as mechanisms to enable distributed information and audio services for wireless operation. We consider techniques for contextual notification i.e. knowing when and how to interrupt the listener. The goal of this work is to develop distributed infrastructure and interaction techniques for scaleable notification, messaging and communication services on a wearable device, with audio as the primary interaction modality.

2: Using Speech and Audio on Wearables

We now consider several techniques and issues related to the use of speech and audio in wearable interfaces.

2.1: Scalability

Traditional input/output modalities such as keyboards and screens lose their utility as devices get smaller. The functionality and ease of use of GUIs does not scale well on small, wearable devices. Hence new I/O modalities must be explored to provide natural and direct interaction with wearable computing. Speech and audio allow the physical interface to be scaled down in size, requiring only the use of strategically placed speakers and microphones [24] and perhaps tactile input for privacy and noisy environments. Yet scalability is also an issue with an audio modality where listeners may wish to hear varying levels of information content in different situations. We will show techniques for scaleable auditory notification and summarization of text and audio messages.

2.2: Unobtrusive Operation

Consider the head mounted display, which is used as the primary output device of most wearable computers. One criticism is that such a display is far too noticeable and therefore socially unacceptable to be considered a serious solution (although commercial systems are quickly driving down the size of such devices). A more serious objection regards the type of perceptual load it places on the user. There are situations in which the user’s eyes are busy although she is otherwise able to attend to information from her wearable computer, such as when walking or driving. An “eyes-free” approach, using audio-based augmentation allows the user to simultaneously perform other tasks while listening or speaking [11]. Such an interface can be unobtrusive and designed to be discreet. We will discuss the use of speech recognition in the interface and consider design solutions for two different wearable configurations.

2.3: Expressive and Efficient Interaction

Voice is considered more expressive and efficient than text, as it places less cognitive demands on the speaker and permits more attention to be devoted to the content of the message [5]. The intonation in human voice also provides many implicit hints about the message content. *VoiceCues* (short audio signatures) played as notifications in *Nomadic Radio,* imply the sender of email messages in a quick and unobtrusive manner. Speech provides a natural and direct means for capturing user’s input. On a wearable, interactions can be structured as small and infrequent *transactions* such as receiving notifications, listening and browsing messages, or communicating with people. Here speech input can be utilized more effectively, where a few phrases allow sufficient control to complete the transaction, and allow the user to focus on the task at hand.

2.4: Peripheral Awareness

People using wearable devices must primarily attend to events in their environment yet need to be notified of background processes or messages. People have a good sense of awareness of background events and can easily shift their focal attention to significant events. 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 [7]. In *Nomadic Radio*, ambient auditory cues are used to convey events and changes in background activity.

2.5: Simultaneous Listening

It is possible for listeners to attend to multiple background processes via the auditory channel as long as the sounds representing each process are distinguishable. This well known cognitive phenomenon, called the "Cocktail Party Effect" [1], 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 [26]. 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 [25]. Bregman [4] claims that stream segregation is better when frequency separation is greater between sound streams. Arons [1] suggests that the effect of spatialization can be improved by allowing listeners to easily switch between channels and pull an audio stream into focus, as well as by allowing sufficient time to fully fuse the audio streams. Such techniques are used in *Nomadic Radio* for *scanning* personal messages.

2.6: Limitations of Speech and Audio

Speech input is a natural means of interaction, yet the user interface must be carefully devised to permit recognition and recording on a wearable device. Issues related to privacy and level of noise in the environment constrain speech input on wearables; in these situations tactile interaction can be used to control the interface. Speech is fast for the author but slow and tedious for the listener. Speech 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 such as interactive skimming [2], non-linear access and indexing [15, 24] and audio spatialization [10, 21] must be considered for browsing audio.

Many of these audio techniques can be used to augment existing visual wearable interfaces. We stress that design for wearable audio computing requires attention to the affordances and constraints of speech and audio in the interface coupled with the physical form of the wearable itself. The physical design and social affordances of the audio interface play an important role in determining how well the wearable will be adopted in certain situations.

3: *Nomadic Radio:* Wearable Audio Messaging

Nomadic Radio [18] provides a unified audio interface to a number of remote information services. Messages such as email, voice mail, hourly news broadcasts, and personal calendar events are automatically downloaded to the device throughout the day. Messages are dynamically structured within categories by filtering and creating views based on attributes such as message type, unread status, priority or time of arrival. A unified messaging interface permits all operations to be performed on any category of messages. Users can select a category such as email or voice mail, browse messages sequentially, and save or delete them on the server.

To provide a hands-free and unobtrusive interface to a nomadic user, the system primarily operates as a wearable audio-only device (see figures 2 and 3), although a visual interface is used for setting user preferences and server-related functions. Information, notifications and feedback is provided to the user through a combination of auditory cues, spatial audio playback and synthetic speech. Textual messages such as email and calendar events are spoken in a concise manner. Special emphasis has been placed on the design of appropriate auditory cues to indicate system activity, message notification, confirmations and break-downs. Auditory cues become especially important when synthetic speech feedback must be scaled back if the user is inferred to be busy.

In *Nomadic Radio* we utilize the metaphor of radio to present audio sources such as voice messages and hourly news 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. Users can navigate messages and control the interface using voice commands, coupled with tactile input in noisy environments or social situations.

3.1: A Usage Scenario

A demonstration of how *Nomadic Radio* (NR) is used during a single session is shown in Figure 1. If the system detects no user activity for some time, it turns off speech synthesis and recognition and goes into sleep mode, yet it can monitor the user and be activated when spoken to.

Wednesday, 1:15 PM

Nitin says: *"Nomadic Wake Up!"*

NR speaks: *"Ok, I'm Listening."*

Nitin: *"Go to my messages"*

NR: <audio cue for command understood> *"Nitin, you have 17 unread messages out of 40 total messages and 3 scheduled events today."* <waits momentarily> <audio cue for a most important message followed by the related VoiceCue> *"Last most important message 40 from Geek about reply to ISWC paper draft".*

Nitin: *"move back"*

NR: <audio cue + VoiceCue> *"Unread very important short message 30 from Tony Jebara about Lets hit the gym?"*

Nitin: *"Read this message"*

NR: <audio cue> *"Message Preview: Tony Jebara, says I'm heading for the gym in 15 minutes, where are you?"*

Nitin: *"Go to my calendar"*

NR: <audio cue> *"Nitin, you have 3 scheduled events today."* <pauses momentarily> *"Special Event 3: Meeting with AT&T at 2:00 PM for 30 minutes".*

Nitin: *"Go to my voice mail"*

NR: <audio cue> *"Nitin, you have 7 voice messages."* <pauses momentarily> *"Last Unread short message 7: Voice message from 225-6406"* <a 3 second audio preview of the message is heard before it slowly fades away> <audio cue>

Nitin: *"Nomadic Sleep!"*

NR: <audio cue> *"Ok, I'll stop listening now."*

Figure 1: A user wakes up *Nomadic Radio* and actively browses messages and calendar events.

This scenario demonstrates the user actively browsing messages using voice commands. Messages are spoken via text-to-speech synthesis and different auditory cues indicate the urgency of messages as well as whether the voice commands were actually recognized. Voice messages, hourly news broadcasts and *VoiceCues* for email are played in a specific spatial direction around the listener's head, based on the time of message arrival (this technique is described in the section on spatial listening).

3.2: Design of the Wearable Audio Platform

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. Earphones are discreet, yet do not allow effective delivery of spatial and simultaneous audio. In these situations speakers worn on the body can 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.

The *SoundBeam Neckset* is the primary audio I/O device used in *Nomadic Radio*. The *Neckset*, a research prototype embodying the first implementation of the *SoundBeam* technology (patented by Nortel), was originally developed for use in hands-free telephony and desktop communications. It consists of two directional speakers mounted on the user's shoulders, and a directional microphone placed on the chest (see figure 2). Nortel allowed us to modify the *Neckset* for spatial audio and evaluate it in the context of wearable audio computing.

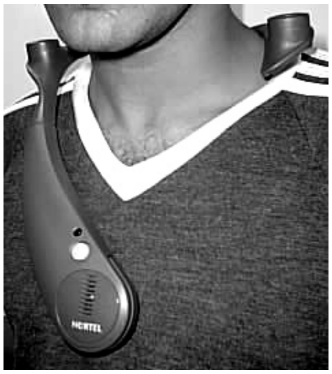

Figure 2: The primary wearable audio device for inter-office use, the *Soundbeam Neckset,* with directional speakers and microphone.

The original *Neckset* was adapted for use in *Nomadic Radio* by patching the amplifier circuit to allow two independent audio channels for stereo output. Direct audio I/O from the wearable PC was added to deliver real-time spatialized audio on the directional speakers and provide voice recognition on the *Neckset*. A button on the *Neckset* could be used to activate speech recognition or deactivate it in noisy environments. Currently this functionality is provided via wireless tactile input and voice commands to put the recognition in “listen” or "sleep" mode.

In October 1997, we incorporated elements of Nortel's *Neckset* into a new solution, based on collaboration with Zoey Zebedee, a fashion designer from the Parsons School of Design, New York. The speaker enclosures were molded on a 3D printer to provide better directional audio. The audio components were integrated into a new wearable and modular configuration called the *Radio Vest*, designed for a more rugged and mobile usage (see figure 3). Here the clip-on speakers and microphone modules can be easily detached when not needed. This configuration delivers sufficient quality for spatialized audio and the sound enclosures ensure private listening for the user.

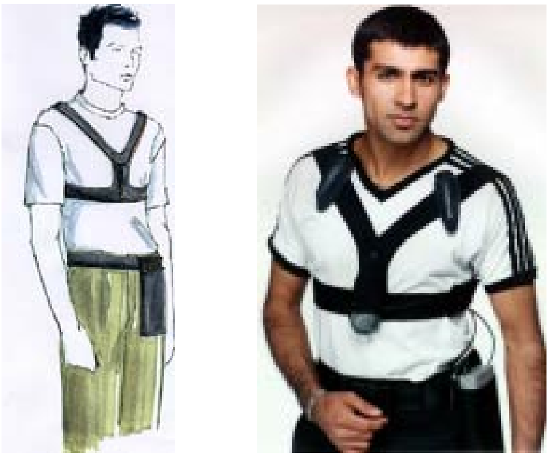

Figure 3: The *Radio Vest*, conceptual design and final implementation of an outdoor wearable configuration with clip-on directional speakers and powered microphone.

The *Radio Vest* is an experimental prototype for outdoor wearable usage, whereas the *Neckset* remains the primary device for inter-office use. We must evaluate the ergonomics and social affordances of such designs as they are worn more readily and consider appropriate refinements. For instance, what social conventions must be used to address the wearable in public spaces? How can others be made aware of the user listening to (and hence distracted by) an incoming message especially when it is inaudible to them, e.g. a visual indicator on the wearable?

4: Auditory Interaction Techniques

A variety of auditory techniques must be supported in a non-visual wearable interface. Spatial listening can be used for browsing and scanning messages easily. Synthetic speech enables explicit feedback in conjunction with speech recognition that provides hands-free navigation and control. In addition to speech, auditory cues provide effective awareness and notifications. These auditory techniques must be designed to function synchronously without overwhelming or confusing the listener.

4.1: Simultaneous and Spatial Listening

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* [21] 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 [10] 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. 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.

In *Nomadic Radio*, audio files are rendered in the spatial environment of the listener using a Java interface to the *RSX 3D* audio API developed by Intel. The perceptual audio models used in *RSX 3D*, are based on a set of head-related transfer function (HRTF) measurements of a KEMAR (electronic mannequin) done by Bill Gardner at the Media Lab [8]. The measurements consist of the left and right ear impulse responses from a loudspeaker mounted 1.4 meters from the KEMAR. The HRTF model allows real-time rendering of several monophonic sound sources, positioned arbitrarily around the head and permits control of their elevation, azimuth, and distance cues.

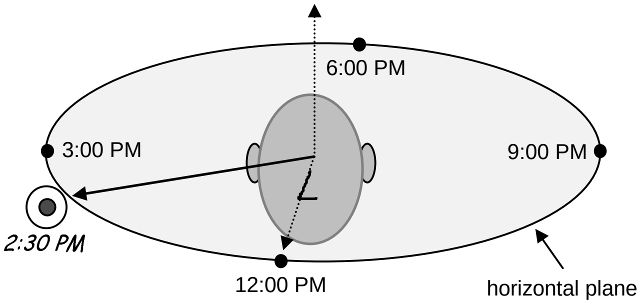

Figure 4: Localization of an incoming message on a chronological 12-hour spatial audio display. The mapping is based on the time of arrival of email and voice messages.

Designing an effective spatial layout for a diverse set of audio messages requires a consideration of content, priority and scalability issues. One approach is to map messages in specific quadrants of the listening space, based on category or urgency. However, this does not scale well as new messages arrive. In *Nomadic Radio*, voicemail and news arrive at different times throughout the day, hence their time of arrival provides a unique parameter for spatial layout. Messages are positioned in chronological order around a listener's head (see figure 4). The listener can discern the approximate time of arrival based on the general direction of the audio source and retain a spatial memory of the message space. Based on this framework, three different forms of spatial listening are utilized:

Broadcasting

When new messages arrive, they are *broadcast* to the listener from a specific spatial location. These messages are heard in the *background* and fade away if the user does not explicitly activate message playback. This mode 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 and Foregrounding

Browsing is an active form of listening where users can select a category and browse sequentially through all messages, playing each one as needed. This mode is similar to the metaphor of switching stations on a radio until a station playing desirable music is found. When a desirable message is heard, the user can listen to the entire message in the *foreground* or request a preview. The *foregrounding* 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 spatial direction is maintained allowing the listener to retain a sense of message arrival time. This spatial continuity is important for discriminating and holding the auditory streams together [1].

Spatial Scanning

Sometimes listeners want to get a preview of all their messages quickly without manually selecting and playing each one. This is similar to the *scan* feature on modern radio tuners. In *Nomadic Radio*, message scanning cycles through all messages by moving each one to the center of the listening space for a short duration and fading it out as the next one starts to play. All messages are played sequentially in this manner, with some graceful overlap as one message fades away and the next one begins to play. The *scanning algorithm* interlaces audio streams in parallel by running each one in its own thread. This simultaneity allows listeners to hear an overall preview of the message space in an efficient manner with minimum interaction.

4.2: Synthetic Speech and Voice Navigation

Synthetic speech allows flexible feedback and playback of text-based messages. Concise speech prompts permit faster interaction and require the listener to retain less information in working memory. In *Nomadic Radio*, speech prompts are designed to be brief, yet convey sufficient information (see figure 1). A conversational model for feedback and navigation allows a natural means of interaction. Implicit and explicit confirmations allow the user to know the effect of her commands and learn the vocabulary over time. Novice users are provided explicit feedback for all voice commands, such as "Going to your voice messages" or "Say that again?". As the user becomes more proficient with the interface, the user can reduce speech feedback for shorter phrases and auditory cues.

Voice navigation is provided in two different modes: *push-to-talk* and *continuous monitoring*. In noisy environments, a push-to-talk strategy allows users to explicitly direct commands to the system or deactivate recognition completely. In normal usage, continuous monitoring allows users to place the system in *listen* or *sleep* mode directly via speech commands. In listen mode the system tries to recognize any commands heard and notifies the user (using speech or audio cues) when it has confidence in a recognized phrase. No feedback is provided for unrecognized commands to minimize annoying prompts. We implemented a networked interface module for speech recognition and synthesis based on AT&T's *Watson* Speech API. The system allows speaker-independent recognition based on a custom-defined application grammar. In *Nomadic Radio*, the vocabulary is structured into 12 meta-commands each of which support a unique set of modifiers, such as *"Go to my* {email | news | calendar | voice-mail}", "*Move* {forward | back}" or "{play | stop | pause | slow-down | speed-up} *Audio*". The user can say "*Help* {command}" for spoken instructions or ask "I am confused. What can I say?" to hear a help overview. *Nomadic Radio* utilizes a *modeless* interface for *unified messaging* such that all voice commands are always valid within each category. Hence commands like "move back" or "play message preview" can apply to email, voice mail, news or calendar events. The vocabulary was redesigned after several usage iterations to select intuitive and consistent commands with minimal acoustic similarity.

4.3: Auditory Cues for Awareness and Feedback

Excessive speech feedback is tedious and can slow down interactions on a wearable. Synthetic speech can be distracting to listeners while they are performing other tasks or in conversation. On the other hand, non-speech audio in the form of *auditory icons* [9] or cues can provide such feedback via short everyday sounds. Auditory cues are a crucial means for conveying awareness, notification and providing necessary assurances in a non-visual interface. Four different sets of auditory cues were designed for handling distinct aspects of the *Nomadic Radio* interface:

(1) Feedback cues indicate events such as *task completion and confirmations* (button pressed, speech understood, connected to servers, finished playing message, loaded or deleted messages), *mode transitions* (switching categories, going to non-speech, ambient, or sleep mode) and *exceptional conditions* (message not found, lost connection with servers or operational errors).

(2) Message priority inferred from email content filtering is used to provide distinct auditory cues for group, personal, timely, and important messages.

(3) *VoiceCues* are used to provide a unique auditory signature for easy identification of the sender of an email message. *VoiceCues* are created by extracting 1-2 second audio samples from the voice messages of callers and associating them with their respective email login. When a new email message arrives, the system queries its database for a related *VoiceCue* for that person before playing it to the user as a notification, along with the priority cues.

(4) Ambient auditory cues are continuously played in the background to provide an awareness of the operational state of the system and ongoing status of messages being downloaded. The sound of flowing water provides an unobtrusive form of ambient awareness. The pitch is increased during file downloads; hence a short email message sounds like a splash while a two minute audio news summary is indicated by faster flowing water, while it is being downloaded.

4.4: Dynamic Operational Modes

Nomadic Radio runs under several dynamically changing modes of operation. The system continuously degrades its auditory interface services based on time elapsed since the last user action. A dynamic state model allows the system to transition to states with minimum auditory interruption, and consequently lower memory usage and optimal CPU performance. In *active speech mode*, speech feedback is scaled down over time, from full message playback to *preview* (1/5th of message) and finally to *summary* (only message header spoken). After a few minutes of inactivity, the system switches to *non-speech mode* by turning off speech synthesis and recognition, although it continues monitoring audio for a "wake up" utterance from the user. In this state the system relies on auditory cues to convey feedback and notifications. Eventually the system switches to an *ambient mode* where most auditory cues are turned off, yet awareness is provided via continuous ambient sounds. After 15 minutes of inactivity the system enters *sleep mode* (1% CPU usage) which conserves power and provides the least interruption if the user has not been paying attention. At any state, the system switches to higher (or lower) modes when a new message arrives (based on its inferred priority level) or if a valid speech command is detected.

5: Unified Messaging Architecture

Timely messaging and remote information access requires a nomadic computing infrastructure. In the Speech Interface group, we have developed an environment [20] that allows subscribers at the MIT Media Lab to access information such as email, voice messages, weather, hourly news and calendar events using a variety of interfaces such as desktops, telephones, pagers, fax, and more recently on wearable platforms. For wearable access, such services have been unified in a manner that is scaleable and easy to navigate using an audio-only modality. In *Nomadic Radio*, spatialized audio, speech synthesis/recognition and audio monitoring are provided as local and distributed services to wearable and wireless platforms. The architecture is modular and extensible such that users can create and subscribe to new services or continue using the system reliably even if one becomes unavailable. This approach allows robust access, coordination and distribution of information/interface services in nomadic environments.

Nomadic Radio consists of client and remote server components that communicate over the wireless LAN. The current architecture (shown in Figure 5) 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. The servers extract information from live sources including voice-mail, email, hourly updates of ABC News, personal calendar, weather, and traffic reports.

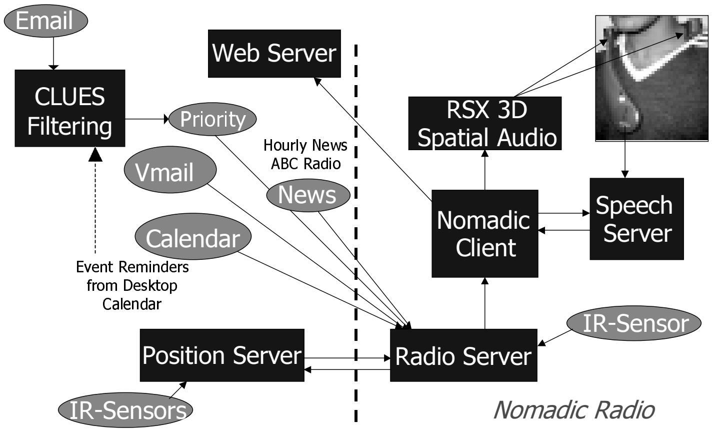

Figure 5: Architecture of *Nomadic Radio*, showing communication between remote server processes and the *Nomadic Client*

Content-based email filtering using CLUES [12], a filtering and prioritization system, has been integrated in *Nomadic Radio*. CLUES determines the timely nature of messages by finding correlation between a user's calendar, rolodex, to-do list, as well as a record of outgoing messages and phone calls. The clients, when notified, download the prioritized text and audio files from the web server.

5.1: Wired Wearable vs. Distributed Wireless

The *Nomadic Clients* have been developed in Java and use a sockets protocol for communication with remote information services, and for integration with the speech synthesis, recognition and audio monitoring modules. These modules are distributed on networked machines and run remotely, based on the needs of the specific wearable configuration. A "wired" configuration requires that the *Nomadic Clients* and the speech and audio modules to operate on the wearable itself. We have evaluated the operation of these modules on Pentium-based wearable PCs, such as the *VIA Wearable* from Flexipc and the Toshiba *Libretto 50* mini-notebook PC. In our experience, the memory and CPU requirements of spatial audio, speech recognition and the Java-based clients, make their reliable operation a challenging task on current wearable platforms. In addition, running multiple interface services requires independent audio channels (even full-duplex audio-boards only support 2 channels). A software-based audio mixer is needed to manage audio from each service, each of which must share a common protocol for audio I/O. However, independent API's for speech recognition and spatial audio do not allow audio control via a mixer.

An alternative configuration allows distributed wireless operation where the clients run on standalone desktop PCs or wearables, but the speech and audio components run remotely on networked PCs. This minimizes the computing and memory requirements on the wearable, and allows independent audio channels to be assigned for speech recognition, audio monitoring, speech synthesis and 3D audio, such that two or more systems can be simultaneously active. In our current architecture we utilize wireless microphones and wireless stereo transmitters for delivering full-duplex audio from two PCs each dedicated to different audio interface services.

The *Nomadic Clients* automatically detect the presence of local or remote speech/audio modules and switch their functionality accordingly. If local modules are detected, the system *interlaces* (activates one while deactivating others) the spatial audio playback with speech recognition and synthesis; here recognition must be activated by a button press. For remote modules, the system operates to provide continuous speech recognition and audio playback. This allows the user to *barge-in* or interrupt speech/audio playback with spoken commands. When a new message arrives and the user activity is determined to be low, audio monitoring is switched on to detect conversation level (and speech recognition deactivated if operating on the same PC). This architecture provides a flexible and modular mechanism to scale the *quality of services*. Scaling is based on the local and distributed processing available and the number of independent audio channels supported in a range of wired and wireless wearable audio platforms.

A multi-threaded design of the *Nomadic Clients* ensures fluid interaction for users, by performing all asynchronous operations as background parallel processes. This is necessary for real-time operation in an audio-only interface. Threads synchronize the timing of spatial audio streams with synthetic speech and auditory cues to provide a coherent and well-paced presentation.

6: Future Work

Deciding when and how to interrupt the listener allows a system to provide effective and unobtrusive notification. Contextual notification is based on four factors: *message priority level* from email filtering, *activity level* based on time since last user action, relevance of user's current *location*, and the *conversation level* estimated from real-time analysis of environmental audio. We are currently developing a system to infer environmental context by audio classification [6]. Our initial efforts have focused on discriminating speech from ambient noise in the environment. In the future, a rich classification of background sounds will help establish user context [19] in offices, classrooms, and the outdoors. We are integrating IR-based receivers on the wearable to provide positioning data to a *Position Server* (that tracks users and messages) via the Locust Swarm [23], a distributed IR location system at the Media Lab. Integration with *Nomadic Radio* will allow contextual notification based on location and auditory characteristics of the environment. As messages arrive, an appropriate notification level will be computed based on these factors, and weighted appropriately for high or low interruption. We are evaluating the effectiveness of such contextual notification for email and voice messaging.

7: Conclusions

We have developed an extensible and distributed architecture for *Nomadic Radio* that provides wearable access to remote information via integrated audio services. We demonstrated a working wearable audio interface that utilizes a variety of interaction techniques for notification, feedback, navigation and control. The system and interface have been refined over several design iterations based on continuous usage by the authors. However, we wish to gain further insight from user experiences in more formal usability settings. Design of a robust and flexible speech/audio interface for wearables provides an effective platform for *unified* messaging and communication. The goal is to provide scaleable and contextually relevant information to nomadic users, based on *situational awareness* and their inferred *focus of attention*.

Acknowledgments

We would like to acknowledge the contributions of Brian Clarkson for ongoing work on the environmental audio classifier, John Holmes and Natalia Marmasse for the *Position Server*, and Bin C. Truong for hardware integration on the *Radio Vest*. We also thank Lisa Fast, Andre Van Schyndel and Richard Levesque at Nortel.

References

[1] Barry Arons. "A Review of the Cocktail Party Effect". *Journal of American Voice I/O Society*, Vol. 12, July 1992.

[2] Barry Arons. "SpeechSkimmer: Interactively Skimming Recorded Speech". *Proceedings of UIST' 93*, November 1993.

[3] Bederson, Benjamin B. "Audio Augmented Reality: A Prototype Automated Tour Guide". *Proceedings of CHI '95*, May 1996, pp. 210-211.

[4] Bregman, Albert S. *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, 1990.

[5] Chalfonte, B.L., Fish, R.S. and Kraut, R.E. "Expressive richness: A comparison of speech and text as media for revision". *Proceedings of CHI'92*, pp. 21-26. ACM, 1991.

[6] Clarkson, Brian and Alex Pentland. "Extracting Context from Environmental Audio". *Proceedings of the International Symposium on Wearable Computing*, IEEE, October 1998.

[7] Cohen, J. Monitoring background activities. In Auditory Display: Sonification, Audification, and Auditory Interfaces. Reading MA: Addison-Wesley, 1994.

[8] Gardner, W. G., and Martin, K. D. HRTF measurements of a KEMAR. *Journal of the Acoustical. Society of America*, 97 (6), 1995, pp. 3907-3908.

[9] Gaver, W. The Sonic Finder: An interface that uses auditory icons. Human Computer Interaction, 4:67-94, 1989.

[10] Kobayashi, Minoru and Chris Schmandt. "Dynamic Soundscape: Mapping Time to Space for Audio Browsing". *Proceedings of CHI '97*, March 1997.

[11] Martin, G.L. The utility of speech input in user interfaces. *International Journal of Man Machine Studies*, 30:355-375, 1989.

[12] Marx, Matthew and Chris Schmandt. "CLUES: Dynamic Personalized Message Filtering". *Proceedings of CSCW '96*, pp. 113-121, November 1996.

[13] Mynatt, E.D., Back, M., Want, R. and Frederick, R. "Audio Aura: Light-Weight Audio Augmented Reality". *Proceedings of UIST '97*, Banff, Canada, Oct 15-17, 1997.

[14] Rekimoto, Jun and Katashi Nagao. "The World through the Computer: Computer Augmented Interaction with Real World Environments". *Proceedings of UIST '95*, November 14-17, 1995, pp. 29-38.

[15] 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.

[16] Roy, Deb K., Nitin Sawhney, Chris Schmandt, Alex Pentland. "Wearable Audio Computing: A Survey of Interaction Techniques" MIT Media Lab Technical Report #434, April 1997. *http://www.media.mit.edu/~nitin/NomadicRadio/AudioWearables.html*

[17] Rudnicky, Alexander, Reed, S. and Thayer, E. "SpeechWear: A mobile speech system". *Proceedings of ICSLP '96*, 1996.

[18] Sawhney, Nitin. "Contextual Awareness, Messaging and Communication in Nomadic Audio Environments", M.S. Thesis, Media Arts and Sciences, MIT Media Lab, May 1998.

[19] Sawhney, Nitin. "Situational Awareness from Environmental Sounds", MIT Media Lab Technical Report, June 1997. *http://www.media.mit.edu/~nitin/papers/Env_Snds/EnvSnds.html*

[20] Schmandt, Chris. "Multimedia Nomadic Services on Today's Hardware". IEEE Network, September/October 1994, pp12-21.

[21] Schmandt, Chris and Atty Mullins. "AudioStreamer: Exploiting Simultaneity for Listening". *Proceedings of CHI 95*, pp. 218-219, May 1995.

[22] Starner, Thad, Steve Mann, Bradley Rhodes, Jeffery Levine, Jennifer Healey, Dana Kirsch, Rosalind Picard, and Alex Pentland, "Augmented Reality through Wearable Computing". *Presence*, Vol. 6, No. 4, August 1997, pp. 386-398.

[23] Starner, Thad and Dana Kirsch. "The locust swarm: An environmentally-powered, networkless location and messaging system." *Proceedings of the International Symposium on Wearable Computing*, IEEE, October 1997.

[24] Stifelman, Lisa J., Barry Arons, Chris Schmandt, Eric A. Hulteen. "VoiceNotes: A Speech Interface for a Hand-Held Voice Notetaker". *Proceedings of INTERCHI '93*, April 1993.

[25] Stifelman, Lisa J. "The Cocktail Party Effect in Auditory Interfaces: A Study of Simultaneous Presentation". MIT Media Lab Technical Report, September 1994.

[26] Wenzel, E.M. Localization in virtual acoustic displays, Presence, 1, 80, 1992.