Nomadic Radio: Speech & Audio Interaction for Contextual Messaging in Nomadic Environments

NITIN SAWHNEY and CHRIS SCHMANDT
MIT Media Lab

Mobile workers need seamless access to communication and information services while on the move. However current solutions overwhelm users with intrusive interfaces and ambiguous notifications. This paper discusses the interaction techniques developed for Nomadic Radio, a wearable computing platform for managing voice and text-based messages in a nomadic environment. Nomadic Radio employs an auditory user interface, which synchronizes speech recognition, speech synthesis, non-speech audio and spatial presentation of digital audio, for navigating among messages as well as asynchronous alerting of newly arrived messages. Emphasis is placed on an auditory modality as Nomadic Radio is designed to be used while performing other tasks in a user's everyday environment; a range of auditory cues provide peripheral awareness of incoming messages. Notification is adaptive and context sensitive; messages are presented as more or less obtrusive based on importance inferred from content filtering, whether the user is engaged in conversation and her recent responses to prior messages. Auditory notifications are dynamically scaled from ambient sound through recorded voice cues up to message summaries. Iterative design and a preliminary user evaluation suggest that audio is an appropriate medium for mobile messaging but care must be taken to minimally intrude on the wearer's social and physical environment.

Categories and Subject Descriptors: B.4.2 [**Input/Output and Data Communications**]: Input/Output Devices—*voice*; D.2.2 [**Software Engineering**]: Tools and Techniques—*modules and interfaces*; *user interfaces*; H.1.1 [**Models and Principles**]: Systems and Information Theory—*value of information*; H.1.2 [**Models and Principles**]: User/Machine Systems—*human factors*; *human information processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering*; *selection process*; H.4.3 [**Information Systems Applications**]: Communications Applications—*electronic mail*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*audio input/output*; *evaluation/methodology*; H.5.2. [**Information Interfaces and Presentation**]: User Interfaces—*evaluation/methodology*; *input devices and strategies*; *interaction styles*; *theory and methods*; H.5.3. [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*asynchronous interaction*

General Terms: Design, Human Factors

Additional Key Words and Phrases: Speech interaction, non-speech audio, spatial listening, passive awareness, contextual interfaces, wearable computing, adaptive interfaces, interruptions, notifications

Portions of this paper are adapted from work published in ISWC '98 and CHI '99 Conference Proceedings.

Authors' Addresses: N. Sawhney and C. Schmandt, Speech Interface Group, MIT Media Lab, E15-352, 20 Ames St., Cambridge, MA 02139; email: nitin@media.mit.edu, geek@media.mit.edu.

1. INTRODUCTION

Communication in the workplace is critical for workgroups; in an information rich environment people access a variety of content to stay abreast of developments. People rely heavily on voicemail and email for day-to-day interaction, along with news, traffic and weather updates while on the move. While many of these services are available on the desktop environment, users carry a variety of devices such as pagers, cell-phones and PDAs for managing personal information and timely communication. Within the workplace, where workers spend a significant portion of their time away from their desktop in meetings or social encounters, they wish to have urgent communication directed to them or simply have an awareness of important events. By observing usage patterns of mobile communication devices, one can infer four general dimensions for design of nomadic user interfaces:

1.1 *Temporal Nature of Communication.* Although synchronous voice communication is generally desirable, many users need asynchronous messaging such as voicemail and email when the other party is unavailable or if either is not currently near a phone. Recipients sometimes prefer to cache messages that they can browse at a later time. Alphanumeric and voice pagers as well as mobile phones and some wireless PDAs provide asynchronous messaging, however few offer an integrated solution.

1.2 *Communication Modality.* Users find voicemail more personal since it conveys inflection and urgency; it allows senders to compose messages easily and recipients to access it "anytime, anywhere" [Whittaker et al. 1998]. However, email is ideal for lengthy and descriptive messages; email is easily skimmed and searched as well as archived and shared with others. People also browse timely information such as news, traffic and weather in the form of text updates (on the web) or radio broadcasts. Hence in a mixed-media environment mobile devices must provide multi-modal interfaces with unified access.

1.3 *Overhead of Transactions.* While on the move, users are usually time-constrained and hence prefer terse interactions for handling communication. The overhead-incurred (setup and interaction time) should be proportional to the duration and value of the transaction. While users may be willing to flip open and launch an application on a laptop or PDA to edit or compose documents, they are less inclined to do so to simply read an email or check the status of an event. It can be argued that even the act of pulling out a pager from ones pocket to check if an incoming message is important, requires unnecessary and repetitive overhead if the information is not timely. Mobile devices should permit users to engage in both extended and rapid transactions in a fluid and unobtrusive manner.

1.4 *Active Usage vs. Peripheral Awareness.* Users take direct actions such as browsing information, making calls or composing messages on mobile devices. However, away from their desks they are usually engaged in other tasks that require their primary attention. In these situations, they wish to remain aware of significant events or be notified for urgent communication, while not having to be interrupted and shift their focus of attention. Although mobile phones and pagers provide alerting, they tend to disrupt both the user and others. Hence some form of peripheral awareness is necessary.

These dimensions suggest design of mobile audio devices that support multi-modal and asynchronous messaging using a unified interaction paradigm. Users must be able to casually browse information on the move as well as receive timely and peripheral awareness of events relevant to their situation. We suggest that the affordances of speech and audio in the mobile interface make it uniquely suited for spontaneous, peripheral and hands-free usage in nomadic environments. In this paper, we demonstrate Nomadic Radio, an audio-only wearable interface that unifies asynchronous communication and remote information services. Email, voicemail, hourly news broadcasts and calendar events are automatically downloaded to the wearable device throughout the day. To provide unobtrusive use in varying situations, we explored two key modes of interaction:

1) *Navigation*: users can actively browse messages via voice and tactile input, along with a synchronized combination of non-speech audio, synthetic speech feedback and spatial audio techniques. Although navigation was designed to be a hands and eyes-free activity, it requires greater attention from the nomadic user.
2) *Notification:* We recognize that users of mobile systems typically need to focus on their foreground task, rather than being disrupted by incoming messages. In Nomadic Radio, filtering and prioritization techniques determine the timely nature of information. Notification is dynamically scaled and adapted by inferring interruptability based on the user's recent activity and context of the environment.

In this paper, we first consider the attributes of speech and audio interaction that make it suitable for nomadic usage, as well as the unique challenges due to its limitations. We review novel audio interfaces developed for hand-held, portable and wearable systems, which have informed the design of Nomadic Radio. We then describe the audio interface, wearable design and system architecture of our working implementation. The paper demonstrates mechanisms for navigation i.e. casual browsing and rapid scanning of audio/text messages. The majority of the paper focuses on the notion of determining when and how to interrupt a listener when a message is received. We consider the characteristics of a scaleable notification model, responsive to the context of the user. We follow this with a preliminary evaluation of the effectiveness and usage of the auditory interaction and notification techniques. We close with a discussion of the design and research implications for Nomadic Radio as well as future mobile and wearable systems.

2. USING SPEECH AND AUDIO ON MOBILE DEVICES

We suggest a non-visual approach for interaction on mobile and wearable devices, both as an alternative to their existing visual displays as well as a secondary modality to complement their functionality. We recognize that an auditory modality may not serve as a general-purpose interface for all nomadic applications, however audio may be better suited within certain domains of information (particularly for content that is intrinsically voice/audio) and in specific usage contexts (when the user's hands or eyes are busy). Several characteristics of speech and audio make it appropriate for *navigation* and *notification* in nomadic systems, yet it is essential to design such systems with an awareness of its limitations.

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, mobile and wearable devices. Hence new I/O modalities must be

explored to provide natural and direct interaction. Speech and audio allow the physical interface to be scaled down in size, requiring only the use of lightweight and strategically placed speakers and microphones [Stifelman et al. 1993] rather than a large keyboard or touch-display. Listeners also wish to hear varying levels of information content in different situations. We will show techniques for scaleable auditory notification of text and audio messages.

2.2 *Unobtrusive Operation.* Hand-held and head-mounted displays demand a certain level of perceptual load 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 or mobile device, such as when walking or driving. A "hands and eyes-free" approach, using audio-based augmentation allows the user to simultaneously perform other tasks while listening or speaking [Martin 1989]. For nomadic access and control, speech provides a natural and convenient mechanism. Simple and reliable voice instructions can replace an awkward series of keyboard inputs or point and click actions using an impractical set of visual prompts. We will discuss effective use of speech I/O for navigation and browsing.

2.3 *Expressive and Efficient Interaction.* Voice is 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 [Chalfonte et al. 1991]. The intonation in speech also provides many implicit hints about the message content. 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. 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 [Cohen 1994]. We will describe how 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" [Arons 1992], 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 [Wenzel 1992]. However experiments show that increasing the number of channels beyond three causes an increase in cognitive overload and hence a degradation in comprehension. Arons [1992] 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. We will demonstrate how such techniques can be used for browsing and rapidly *scanning* audio messages.

2.1 Problems with Speech and Audio in a Nomadic Environment

Several characteristics of noisy and social environments make the use of voice input and audio output on mobile devices and wearables, ineffective or simply awkward. These aspects must be carefully examined to consider appropriate solutions or alternatives.

2.1.1 *Speech can be Slow and Tedious.* Excessive speech interaction can be tedious especially if the user must repeat the same command in a noisy environment. Tasks requiring fine control or repetitive input such as "move forward", make speech input awkward ("faster, faster....") [Stifelman et al. 1993]. Such tasks are better accomplished using button input. In addition, hearing excessive recorded and especially synthetic speech can be a burden on a listener, due to its sequential and transient nature. Hence techniques such as skimming compressed speech [Arons 1997] and *scanning* audio streams are useful.

2.1.2 *Acquiring Application Vocabulary.* Unlike on-screen buttons, speech commands on a non-visual application must be recalled by the user. The vocabulary can be designed to provide intuitive commands, yet the user must be made familiar with their syntax and extended functionality over time. In many cases, the user may not recall the right command to speak, and either needs to inquire the application or use an alternative means for accomplishing the task easily.

2.1.3 *Effect of Environment on Speech I/O.* In noisy environments, the accuracy of recognition is seriously degraded and the interface can be less responsive. Directional microphones and noise cancellation[1] techniques ease the problem to some extent. In addition, if users are stressed or frustrated they will not articulate spoken commands clearly and this too will affect recognition accuracy. Hence in situations where speaking is less desirable or recognition is simply not feasible, tactile input provides a potential solution that can be coupled with speech to provide a responsive and unobtrusive hybrid interface. Similarly it is challenging to consider how speech and audio output on mobile devices can be presented and adjusted in noisy environments, without isolating the listener from events in their auditory environment (this is particularly important for the visually impaired).

2.1.4 *Managing Social Conventions.* Using speech recognition on a wearable device typically requires use of body-worn microphones. Most social and cultural conventions assume that it is awkward for people to be speaking to themselves, especially with no prior warning (unlike taking calls on a cell-phone). It can be confusing whether the user is addressing her wearable or the person next to her in an elevator or meeting room. An explicit push-to-talk button that produces an audio cue heard by others and a continuous visual indicator (flashing light on the wearable) can reduce some of the confusion. Users speaking during a meeting or lecture can be distracting to others in the room (unless they whisper). Over time, people adopt new social conventions or simply get accustomed to such technologies, as they have with people using cellular phones in social environments. However, during the early phase of adoption, many users will be less inclined to use speech on wearables in public places.

---

[1] Good directional microphone design provides some form of noise cancellation. Yet active noise cancellation requires pre-sampling the level of noise in the background and subtracting it from the user's speech (a much more difficult problem).

2.1.5 *Lack of Privacy, Security or Confidentiality.* In a social environment, speech I/O poses a number of additional problems. Users will not feel comfortable speaking names or passwords and hearing confidential information (financial or medical transactions) aloud while near their coworkers. The application must be designed to allow alternative means of input and discreet audio output.

We stress that design for nomadic computing requires attention to the affordances and constraints of speech and audio in the interface coupled with the characteristics of the user's physical environment. Designing a versatile interface plays an important role in determining how well the wearable or mobile system will be adopted in certain situations.

3. NOMADIC AUDIO INTERFACES

3.1 Paging and Telephony

Simple devices, such as pagers, provide a convenient form of alerting users to remote information. These devices are lightweight and offer low-bandwidth for communication; subsequently the interface does not afford rich delivery of information content. Recent 2-way pagers provide small displays and customizable alerts, however messaging is not seamless (users must shift their attention to scan text on the device) and richer audio sources are not integrated on the same device. Notification is a key problem on most devices, where the user is unable to determine whether the current message is relevant. On some pagers and mobile phones users can define ringing tones for preferred people or caller groups, however most users are less inclined to continuously maintain such static filtering rules. Recent telephony services such as WildFire and Portico offer speech interaction and rule-based filtering. Phoneshell [Schmandt 1994] offers subscribers integrated access to unified messaging using digitized audio and synthesized speech. Speech Acts [Yankelovich 1994] and MailCall [Marx and Schmandt 1996] provide conversational interaction and assistance if the user is not understood.

Phoneshell and MailCall incorporate *Clues*, a dynamic filtering system that prioritizes messages by inferring user interest based on correlation with their calendar, rolodex and recent email replies. Such telephone-based messaging systems generally require synchronous usage, i.e. users must dial-in to the messaging service each time they wish to browse their messages. Hence, there is less incentive to use the devices on an on-demand basis or for short transactions. Specialized phones from AT&T and Nokia allow asynchronous use but do not provide an unobtrusive hands-free interface or *contextual notification*. The *Profile* function on the Nokia 6100 phones enables users to adjust the ringing tones according to various situations and caller group identification. However, this requires users to pre-designate important callers and manually set active profiles continuously. In contrast to phone-based messaging, a wearable computer can cache all messages for browsing later, and by sensing the user, message and environmental context at all times, it can decide when it is most appropriate to interrupt the listener and scale its feedback accordingly.

3.2 Mobile Audio Devices

Nomadic users want continuous access to relevant information using natural and unobtrusive interfaces. Hand-held audio devices such as VoiceNotes [Stifelman et al. 1993] and NewsComm [Roy and Schmandt 1996] provide speech and button interfaces

for browsing user-authored notes or personalized news audio. VoiceNotes utilized a number of techniques for modeless navigation, scanning lists as well as speech and non-speech feedback. Nomadic Radio uses similar techniques for browsing spoken text and spatial audio, while focusing on notification. In NewsComm, audio servers extract structural descriptions of the news programs by locating speaker changes and pauses. The interface consists of button-based controls to allow the user to select recordings, indicate which ones were interesting and navigate within a recording using structural information.

Nomadic Radio does not utilize structured audio for navigation, but techniques such as *spatial scanning* and *foregrounding* allow simultaneous playback and scaleable presentation of audio and text messages. Instead of utilizing listener profiles, messages once loaded in Nomadic Radio are dynamically presented based on a contextual notification model. A recent system by Clarion, the AutoPC, provides email, phone and real-time information to drivers using a speech interface. Our key concern is how to minimize distraction to drivers and mobile users, by providing notifications when the user seems most likely to be interruptible.

3.3 Wearable Audio Computing

Several recent projects utilized speech and audio I/O on wearable devices to present information. A prototype augmented audio tour guide [Bederson 1995] played digital audio recordings indexed by the spatial location of visitors in a museum. *SpeechWear* [Rudnicky et al. 1996] enabled users to perform data entry and retrieval using speech recognition and synthesis. In a recent paper by Starner et al. [1997] suggests the use of sensors and user modeling to allow wearables to infer when users should be interrupted by incoming messages. They suggest waiting for a break in the conversation to post a message summary on the user's heads-up display. *Audio Aura* [Mynatt et al. 1998] explored the use of background auditory cues to provide serendipitous information coupled with people's physical location in the workplace. In Nomadic Radio, the user's inferred context rather than actual location is used to decide when and how to deliver scaleable audio notifications.

A good example of a wearable audio interface is embodied in the Wristwatch-type PHS telephone developed by NTT [Suzuki et al. 1998]. A combination of voice recognition and four operational buttons are used for all call-handling functions. In field trials with 40 users at the Nagano Olympic Games, most users preferred communication via the watch's built-in loudspeaker, rather than wear ear-microphone units. Over half the users wore the device as a wristwatch whereas a third wore it around the neck as a pendent. In addition, the same proportion of users utilized voice, button input or both, suggesting a pattern of interaction best suited to their environment. Hence users are willing to adopt new modes of usage and physical configurations for enhanced personal communication. A novel approach towards a hand-worn microphone and earpiece set provides a natural and compact interface for voice I/O [Fukumoto and Tonomura 1999]. In this paper we describe a primarily non-visual approach to provide timely information to nomadic listeners, based on a variety of contextual cues.

4. NOMADIC RADIO

Nomadic Radio is a messaging application for a wearable platform [Sawhney and Schmandt 1998], which unifies remote information services such as email, voice mail,

hourly news broadcasts, and personal calendar events. 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 modeless 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 an audio-only wearable device (Figure 1), although a visual interface is used for setting user preferences and server-related functions. Textual messages such as email and calendar events are spoken via synthetic speech, whereas voicemail and broadcast news segments are presented as simultaneous spatial audio streams. Special emphasis has been placed on the design of appropriate auditory cues to indicate system activity, message notification, confirmations and breakdowns. Users can navigate messages and control the interface using voice commands, coupled with button input for situations where it is too noisy for recognition or socially intrusive to speak.

4.1 Wearable Audio Platform

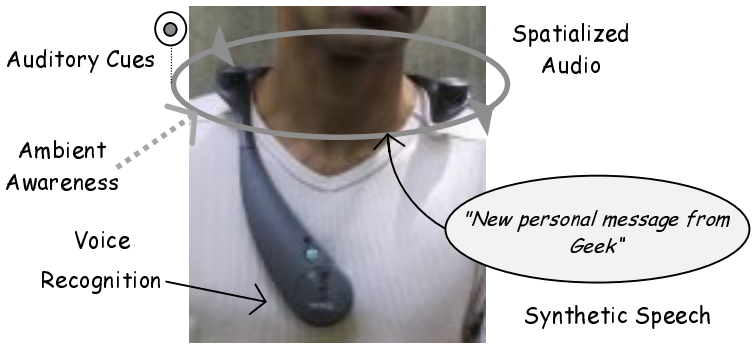

Fig. 1. The primary wearable audio device for inter-office use, the *Soundbeam Neckset,* with directional speakers and microphone.

On mobile systems, audio output is generally provided via monaural speakers; audio is sometimes difficult to hear and such speakers do not allow a rich delivery of multiple audio streams. In addition, this approach minimizes privacy and causes disruption (and sometimes embarrassment) when used in public environments. 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 provide directional sound to the user (without covering the ear), yet must be designed to be worn easily and least audible to others.

To provide a hands-free and unobtrusive interface to a nomadic user, our goal was to provide a wearable audio-only device. The *SoundBeam Neckset,* a research prototype patented by Nortel for use in hands-free telephony, was adapted as the primary wearable platform in Nomadic Radio. It consists of two directional speakers mounted on the user's shoulders, and a directional microphone placed on the chest (Figure 1). The volume on the Neckset can be set to a level that ensures that audio output is heard primarily by the user, while still being within conversational distance with others. In October 1997, we integrated these audio components into a new wearable and modular configuration called the *Radio Vest*[2], designed for a more rugged outdoor and mobile usage. Here the clip-on speakers and microphone modules can be easily detached when not needed.

[2] *http://www.media.mit.edu/~nitin/projects/NomadicRadio/WhatNR.htm#Design*

4.2 System Architecture

Timely messaging and remote information access requires nomadic computing infrastructure. In the Speech Interface group, we have developed an environment [Schmandt 1994] that allows subscribers at the MIT Media Lab to access desktop information using a variety of mobile interfaces such as 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.

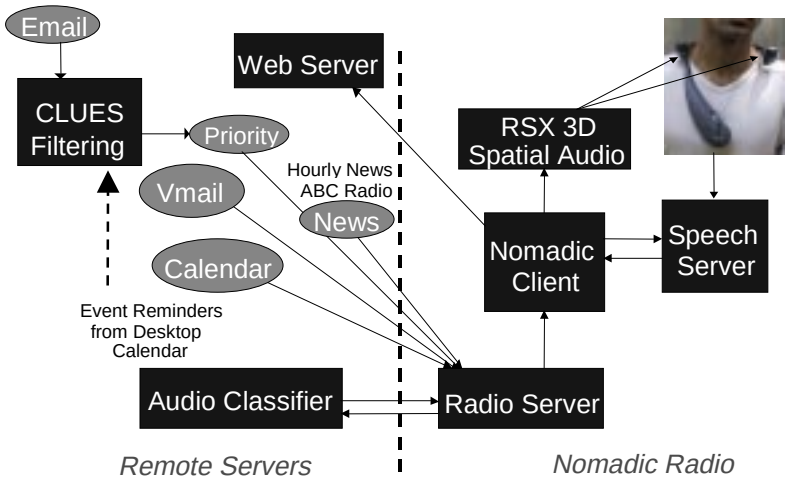

Fig. 2. Architecture of Nomadic Radio, showing communication between remote server processes and the *Nomadic Client.*

Nomadic Radio consists of Java-based clients on the wearable PC and remote server components communicating within a building on an 802.11 wireless LAN. The current architecture (Figure 2) relies on server processes, written in C and Perl running on Sun SPARCstations, that extract information from live sources including voice-mail, email, hourly updates of ABC News, personal calendar, weather, and traffic reports. Email messages are prioritized based on content filtering via *Clues* (described later). The clients, when notified, download the prioritized text and audio files from the web server. The audio classifier, running on a remote server, detects whether the user is speaking or if there is a conversation occurring nearby and dynamically adjusts the level of notification provided for incoming messages.

In Nomadic Radio, spatialized audio, speech synthesis/ recognition and audio monitoring are provided as local and distributed services to wearable and wireless platforms. The system currently runs on a Toshiba Libretto 100CT mini-portable PC, however it can be used with wireless audio I/O. 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. Multi-threaded design 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.

5. NAVIGATION

Speech I/O provides a rich means for communication and it can be effectively leveraged for interaction with devices in our environment (or wearables on our body). Voice-enabled applications for wearables and mobile systems must be designed to be unobtrusive and responsive, as the user expects to use them casually and instantaneously

in a variety of nomadic environments. We will now consider issues related to speech recognition, vocabulary design and voice-based navigation. We also discuss presentation and scanning via synthetic speech and spatial audio techniques.

5.1 Voice Navigation and Control

A networked speech recognition and synthesis module was developed for Nomadic Radio, based on AT&T's *Watson* Speech API[3]. The system, which runs in real-time on a Pentium-based wearable, allows speaker-independent recognition based on a custom-defined application grammar. The recognition rate is over 90% for short phrases spoken by users accustomed to the system, in a noise-free environment. Performance reduces to nearly 40-50% in noisy situations, where button-based interaction becomes necessary. Future work in active noise suppression devices [Akoi et al. 1998] and adaptive speech recognition is necessary to improve performance.

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 ToCHI paper draft".*

Nitin: *"move back"*

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

Nitin: *"Read this message"*

NR: <audio cue> *"Message Preview: Dimitri, 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"* <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."*

Fig. 3. A user wakes up Nomadic Radio and actively browses messages and calendar events.

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 | slow-down | speed-up} Audio"*. The user can say *"Help {command}"* for specific spoken instructions or ask, *"I am confused. What can I say?"* to hear an overview of all help commands. As mentioned earlier, Nomadic Radio utilizes a modeless interface for unified messaging such that all voice commands are always valid within each category (see scenario in Figure 3). Hence, commands like *"move back"*, *"play message preview"* or *"remove this message"* can apply to email, voice mail, news or calendar events. The vocabulary was redesigned after several iterations to select intuitive and consistent commands with minimal acoustic similarity. In Nomadic Radio, speech prompts are designed to be brief, yet convey sufficient information. Concise feedback permits faster interaction and requires the listener to retain less information in working memory. 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.

Speech input 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. A time-out setting

[3] *http://www.research.att.com/projects/watson/*

allows the user to press the listen button and start speaking. The system automatically detects end of the utterance and stops listening. Here, the system prompts the user with spoken feedback (*"Say that again?"*) if it does not recognize a phrase when the user presses the listen button. A *push-to-talk* mode is also necessary for Wearable PCs without full-duplex audio support. Here the speech module notifies the *Nomadic Client* to silence all audio playback when the system is listening, and reactivate audio once it has heard an utterance. In *continuous monitoring*, the user can explicitly place the system in *listen* or *sleep* mode using the trigger phrases *"Nomadic Wake-up"* and *"Nomadic Sleep"*. 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. While listening, the system tries to recognize any commands heard and will notify the user via spoken feedback or audio cues *only* when it has confidence in a recognized phrase. No feedback is provided for unrecognized commands to minimize annoying prompts. *Continuous monitoring* supports the ability for users to *barge-in* with spoken commands while the system is speaking or playing an audio stream.

5.2 Spatial and Simultaneous Listening

Spatialized audio is a technique by which the characteristics of sound sources are perceptually modeled such that a listener hears them at distinct locations around the head. In Nomadic Radio, spatial audio is rendered using the RSX 3D audio API developed by Intel, which is based on a model of head-related transfer function (HRTF) measurements by Gardner and Martin [1995].

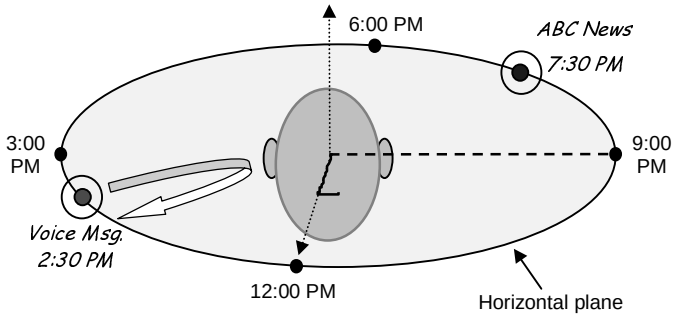

Fig. 4. Audio messages are positioned around the listener's head based on their time of arrival, providing a scaleable presentation for browsing. Here, while the user is listening to an ABC News broadcast in the background, an incoming voice message begins to play, gradually fading in and out of the listener's foreground. Spatialized audio allows the listener to segregate both messages, while focusing attention on the primary audio stream.

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* [Schmandt and Mullins 1996] 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 [Kobayshi and Schmandt 1997] 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 helps segregate simultaneous audio streams more easily. Here the exact location of the sound is less important, but provides cues about the message such as its category, urgency, and time of arrival.

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 (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. Segregated audio also enables the listener to focus on an audio stream such as voicemail in the foreground while hearing a news broadcast in the background.

5.3 Message Scanning

Sometimes listeners want to get a preview of all their messages quickly without manually selecting and playing each one. A recent study of strategies for managing voicemail [Whittaker et al. 1998] suggested the need for techniques that allow users to scan their messages, under time constraints. In Nomadic Radio, users can automatically hear all messages by *scanning* i.e. saying, *"scan messages"*. This presents each message for a short duration, beginning with the current one and moving backwards until all messages are played. The duration is based on the user's selected presentation level, i.e. audio cue, summary, preview or full message. The scan function works identically on both text and audio messages in any view. For text messages, spoken via synthetic speech, there is a 1-second delay (period of silence) between messages during scanning. The user has a temporal target window of 2 seconds to select the last message played after hearing it, similar to VoiceNotes. Once the user hears a message of interest or wants to stop scanning, issuing any command will deactivate scanning.

For audio messages, spatial foregrounding techniques allow listeners to hear several messages within a short duration. Spatial audio 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 (Figure 4). 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.

6. SCALEABLE AND CONTEXTUAL NOTIFICATION

There are several problems with notifications on existing mobile devices, described here.

6.1 *Lack of Differentiation in Notification Cues.* Every device provides some unique form of notification. In many cases, these are distinct auditory cues. Yet, most cues are generally *binary* in nature, i.e. they convey only the occurrence of a notification and not its urgency or dynamic state. This prevents users from making timely decisions about received messages without having to shift focus of attention (from the primary task) to interact with the device and access the relevant information.

6.2 *Minimal Awareness of the User and Environment.* Such notifications occur without any regard to the user's engagement in her current activity or her focus of attention. This interrupts a conversation or causes an annoying disruption in the user's task and flow of thoughts. To prevent undue embarrassment in social environments, users typically turn off cell-phones and pagers in meetings or lectures. This prevents the user from getting notification of timely messages and frustrates people trying to get in touch.

6.3 *No Learning from Prior Interactions with User.* Such systems typically have no mechanism to adapt their behavior based on the positive or negative actions of the user. Pagers continue to buzz and cell-phones do not stop ringing despite the fact that the user may be in a conversation and ignoring the device for some time.

6.4 *Lack of Coordinated Notifications.* All devices compete for a user's undivided attention without any coordination and synchronization of their notifications. If two or more notifications occur within a short time of each other, the user gets confused or frustrated. As people start carrying around many such portable devices, frequent and uncoordinated interruptions inhibit their daily tasks and interactions in social environments.

Given these problems, most devices fail to serve their intended purpose of notification or communication, and thus do not operate in an efficient manner for a majority of their life cycle. New users choose not to adopt such technologies, having observed the obvious problems encountered with their usage. In addition, current users tend to turn off the devices in many situations, inhibiting the optimal operation of such personal devices. A recent observational study [O'Conaill and Frohlich 1995] evaluated the effect of interruptions on the activity of mobile professionals in their workplace. An interruption, defined as an asynchronous and unscheduled interaction, not initiated by the user, results in the recipient discontinuing the current activity. The results revealed several key issues. On average, subjects were interrupted over 4 times per hour, for an average duration slightly over 2 minutes. Hence, nearly 10 minutes per hour was spent on interruptions. Although a majority of the interruptions occurred in a face-to-face setting, 20% were due to telephone calls (no email or pager activity was analyzed in this study). In 64% of the interruptions, the recipient received some benefit from the interaction. This suggests that a blanket approach to prevent interruptions, such as holding all calls at certain times of the day, would prevent beneficial interactions from occurring. However, in 41% of the interruptions, the recipients did not resume the work they were doing prior to it. But active use of new communication technologies makes users easily vulnerable to undesirable interruptions.

These interruptions constitute a significant problem for mobile professionals using tools such as pagers, cell-phones and PDAs, by disrupting their time-critical activities. Improved synchronous access using these tools, benefits initiators but leaves recipients with little control over the interactions. The study suggests development of improved filtering techniques that are especially light-weight, i.e. do not require more attention from the user and are less disruptive than the interruption itself. By moving interruptions to asynchronous media, messages can be stored for retrieval and delivery at more appropriate times. Personal messaging and communication, demonstrated in Nomadic Radio, provides a simple and constrained problem domain in which to develop and evaluate a contextual notification model. Messaging requires development of a model that dynamically selects a suitable *notification strategy* based on message priority, usage level, and environmental context. Such a system must infer the user's attention by monitoring her current activities such as interactions with the device and conversations in the room. The user's prior responses to notifications must also be taken into consideration to adapt the notifications over time. We now discuss techniques for *scaleable auditory presentation* and an appropriate parameterized approach towards *contextual notification.*

6.1 Scaleable Auditory Presentation

A scaleable presentation is necessary for delivering sufficient information while minimizing interruption to the listener. In Nomadic Radio, the user is notified of message arrival using a variety of auditory cues, synthetic speech and scaled audio, based on the inferred priority of the message and user context. Messages are scaled dynamically to unfold as seven increasing levels of notification (Figure 5).

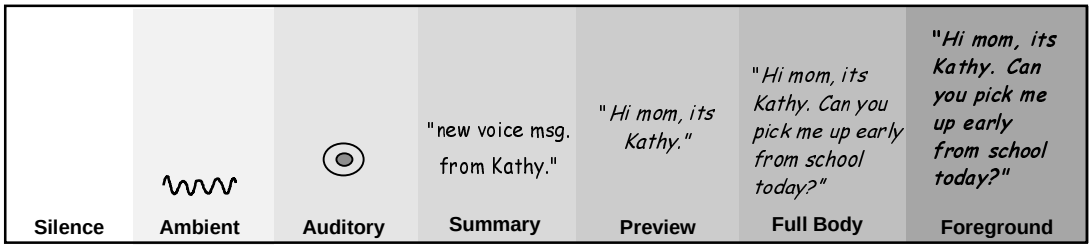

Fig. 5. Dynamic scaling of an incoming voice message during its life cycle based on the interruptability of the listener. The message is presented at varying levels: from a subtle auditory cue to foreground presentation.

6.1.1 *Silence for Least Interruption and Conservation.* In this mode all auditory cues and speech feedback are turned-off. Messages can be scaled down to silence when the message priority is inferred to be too low for the message to be relevant for playback or awareness to a user, based on her recent usage of the device and the conversation level. This mode also conserves processing, power and memory resources on a portable device or wearable computer.

6.1.2 *Ambient Cues for Peripheral Awareness.* In Nomadic Radio, 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 (Figure 6). The sound of flowing water provides an unobtrusive form of ambient awareness that indicates the system is active (silence indicates sleep mode). Such a sound tends to fade into the perceptual background after a short time, so it does not distract the listener. The pitch is increased during file downloads, momentarily foregrounding the ambient sound. A short e-mail message sounds like a splash while a two-minute audio news summary is heard as faster flowing water while being downloaded. This implicitly indicates message

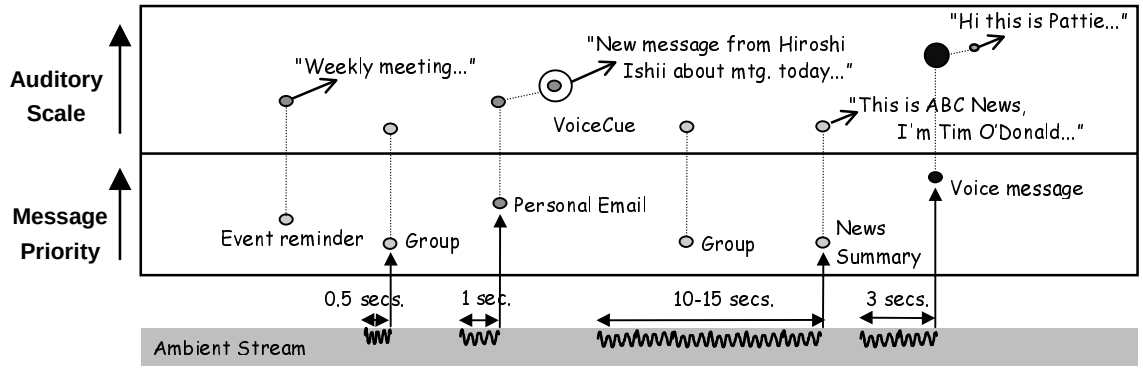

Fig. 6. Users wearing Nomadic Radio are provided a continuous awareness of incoming messages through a well-paced progression of auditory alerts interlaced with ambient sound. Initially, the ambient background sound (of flowing water) is heard speeded-up while messages are being downloaded (pre-cueing the listener to an incoming message). Scaleable auditory cues notify the listener regarding message priority, while *VoicesCues* identify the sender.

size without the need for additional audio cues and prepares the listener to hear (or deactivate) the message before it becomes available. Such peripheral awareness minimizes cognitive overhead of monitoring incoming messages relative to notifications played as distinct auditory cues, which incur a somewhat higher cost of attention on part of the listener.

In *ARKola* [Gaver et al. 1991], an audio/visual simulation of a bottling factory, repetitive streams of sounds allowed people to keep track of activity, rate, and functioning of running machines. Without sounds people often overlooked problems; with auditory cues, problems were indicated by the machine's sound ceasing (often ineffective) or via distinct alert sounds. The various auditory cues (as many as 12 sounds play simultaneously) merged as an auditory texture, allowed people to hear the plant as a complex integrated process. Background sounds were also explored in *ShareMon* [Cohen 1994], a prototype application that notified users of file sharing activity. Cohen found that pink noise used to indicate %CPU time was considered "obnoxious", even though users understood the pitch correlation. However, preliminary reactions to wave sounds were considered positive and even soothing. In *Audio Aura* [Mynatt et al. 1998], alarm sounds were eliminated and a number of "harmonically coherent sonic ecologies" were explored, mapping events to auditory, musical or voice-based feedback. Such techniques were used to passively convey the number of email messages received, identity of senders, and abstract representations of group activity.

6.1.3 *Auditory Cues for Notification and Identification.* In Nomadic Radio, auditory cues are a crucial means for conveying awareness, notification and providing necessary assurances in its non-visual interface. Different auditory techniques provide distinct feedback, awareness and message information.

6.1.3.1 *Feedback Cues.* Several types of audio cues indicate feedback for a number of operational events in Nomadic Radio:

1) Task completion and confirmations - button pressed, speech understood, connected to servers, finished playing or loaded/deleted messages.
2) Mode transitions - switching categories, going to non-speech or ambient mode.
3) Exceptional conditions - message not found, disconnected with servers, and errors.

6.1.3.2 *Priority Cues for Notification.* In a related project, "email glances" [Hudson and Smith 1996] were formulated as a stream of short sounds indicating category, sender and content flags (from keywords in the message). In Nomadic Radio, message priority inferred from email content filtering provides distinct auditory cues (assigned by the user) for group, personal, timely, and important messages. In addition, auditory cues such as telephone ringing indicate voice mail, whereas an extracted sound of a station identifier indicates a news summary.

6.1.3.3 *VoiceCues for Identification.* A novel approach for easy identification of the sender of an email is based on creating a unique auditory signature of the person. *VoiceCues* are created by manually extracting a 1-2 second audio sample 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. The authors have found *VoiceCue*s to be a remarkably effective method for quickly conveying

the sender of the message in a very short duration. This technique reduces the need for synthetic speech feedback, which can often be distracting.

6.1.4 *Message Summary Generation.* A spoken description of an incoming message can present relevant information in a concise manner. Such a description typically utilizes header information in email messages to convey the name of the sender and the subject of the message. In Nomadic Radio, message summaries are generated for all messages, including voice-mail, news and calendar events. The summaries are augmented by additional attributes of the message indicating category, order, priority, and duration. A study on voicemail usage [Whittaker et al. 1998] indicated that users generally listen to the first few seconds of a message (for intonation in caller's voice) to determine if the message requires immediate action, rather than listening to voicemail headers. Hence for audio sources, like voice messages and news broadcasts, the system plays the first 2.5 seconds of the audio. This identifies the caller and the urgency of the call (from intonation) or provides a station identifier for news summaries.

6.1.5 *Message Previews.* Messages are scaled to allow listeners to quickly preview the contents of an email or voice message. In Nomadic Radio, a preview for text messages extracts the first 100 characters of the message. This heuristic generally provides sufficient context for the listener to anticipate the overall message theme and urgency (however text summarization techniques, based on tools such as ProSum[4], would allow a scaleable summary of arbitrarily large text). For email messages, redundant headers and previous replies are eliminated from the preview

It's 1:15 PM and Jane is wearing Nomadic Radio. She has a meeting in a conference room in 15 minutes. She receives an early notification via an auditory cues and synthetic speech.
NR: <auditory cue for early event reminder> ***"Jane, you have a scheduled event at 1:30 PM today."*** <pause> ***"Meeting with Motorola sponsors in the NIF room for 30 minutes."***
Jane scans her email messages to hear one about the meeting and check who else is coming. A new group message arrives.
NR: <ambient sound speedup> ***"New group message from Geek about where is lunch?"***
Jane ignores the message and heads over to the conference room. At this point, since Jane has been inactive for some time and the conversation level in the room is higher, the system scales down notifications for all incoming messages. Moments later a timely message arrives (related to an email Jane sent earlier) and the conversation level is lower. The system first plays an auditory cue and gradually speeds up the background sound of water to indicate to Jane that she will hear a summary soon.
NR: <auditory cue for timely message> + <faster ambient sound>
Jane is now engrossed in the meeting so she prevents the system from playing a summary of the message, by pressing a button on Nomadic Radio (she does not speak to avoid interrupting the meeting). The sound of water slows down and message playback is aborted. The system recognizes Jane is busy and turns down notification for subsequent messages.
Its 1:55 PM and the meeting is nearly over. The system is currently in *sleep* mode. A very important voice message from Jane's daughter arrives. It recognizes the priority of the message and despite the high conversation level and low usage, it plays auditory cues to notify Jane. The ambient sound is speeded-up briefly before playing a preview of the message.
NR: <audio cue for voice message "telephone ringing" sound> + <VoiceCue of Kathy>
Jane hears her daughter's voice and immediately presses a button to play the message. The system starts playing the full voice message in the foreground (instead of just a preview), two seconds earlier than its computed latency time.
NR: <human voice> ***"Hi mom, its Kathy. Can you pick me up early from school today?"*** <audio cue for end of message>
Jane excuses herself from the meeting and browses her email on Nomadic Radio while walking back to her car.

Fig. 7. A scenario showing Jane using Nomadic Radio to listen to notifications while engaging in other tasks.

[4] *http://transend.labs.bt.com/prosum/on_line/*

for succinct playback to the listener (however, we have not considered means for incorporating replies that may constitute an intrinsic part of the message). A preview for an audio source such as a voice message or news broadcast presents a fifth of the message at a gradually increasing playback rate of up to 1.3 times faster than normal. There are a range of techniques for time-compressing speech without modifying the pitch [Arons 1997], however twice the playback rate usually makes the audio incomprehensible. A better representation requires a structural description of the audio, based on pauses in speech, speaker and topic changes [Roy and Schmandt 1996].

6.1.6 *Playing Complete Message Content.* This mode plays the entire audio file or reads the full text of the message at the original playback rate, in the background.

6.1.7 *Foreground Rendering via Spatial Proximity.* An important message is played in the foreground of the listening space. The audio source of the message is rapidly moved closer to the listener, allowing it to be heard louder, and played there for $4/5^{th}$ of its duration. The message gradually begins to fade away, moving back to its original position and amplitude for the remaining $1/5^{th}$ of the duration. The *foregrounding* algorithm ensures that the messages are quickly brought into perceptual focus by pulling them to the listener rapidly. Messages are pushed back slowly to provide an easy fading effect as the next one is heard (Figure 4). Spatial direction is maintained so listeners can retain focus on an audio source even if another begins to play. This spatial continuity is important for discriminating and holding the auditory streams together [Arons 1992].

The example in Figure 6 shows show these techniques are combined to provide a fluid auditory presentation[5]. Here the user is made aware of events and incoming messages via subtle auditory cues and ambient sound. E.g. a new email message from Hiroshi is heard as a *VoiceCue* followed by a synthetic speech summary. An ABC news summary is heard in the background, while a voicemail from Pattie is heard as a ring followed by 3-4 seconds of her voice in the foreground. Hence a range of techniques provide scaleable forms of background awareness, auditory notification, spoken feedback and foreground rendering of incoming messages (see an example scenario incorporating these techniques in Figure 7). Along with scaleable presentation, the *pace of disclosure* for incoming messages is carefully controlled to ensure that users are easily pre-cued to attend to a forthcoming message summary. The gradually unfolding presentation (Figure 6) permits sufficient time to take an action (deactivate or scale playback) before the message is fully presented. The pace of audio interaction is also maintained during browsing and scanning modes, based on related techniques for time synchronization utilized in VoiceNotes.

6.2 Contextual Notification

In Nomadic Radio, context dynamically scales the notifications for incoming messages. The primary contextual cues used include: *message priority* from email filtering, *usage level* based on time since last user action, and the *likelihood of conversation* estimated from real-time analysis of the auditory scene. In our experience these parameters provide sufficient context to scale notifications, however data from motion or location sensors can also be integrated in such a model. We utilize a linear and scaleable notification model, based on a notion of estimating costs of interruption and the value of information

[5] On-line example of scaleable audio – http://www.media.mit.edu/~nitin/NomadicRadio/nr_media/alert.wav

to be delivered to the user (Figure 8). This approach is related to recent work [Horvitz and Jed 1997] on using perceptual costs and a focus of attention model for scaleable graphics rendering.

6.2.1 *Message Priority.* The priority of incoming messages is explicitly determined via content-based email filtering using *Clues* [Marx and Schmandt 1996], a filtering and prioritization system. *Clues* relates incoming messages with items in the user's calendar, rolodex, to-do list, as well as a record of outgoing messages and phone calls. Email messages are also prioritized if the user is traveling and meeting others in the same geographic area (via area codes in the rolodex). Static rules created by the user prioritize specific people or message subjects. When a new email message arrives, keywords from its sender and subject header information are correlated with static and generated filtering rules to assign a priority to the message. The current priorities include group, personal, very important, most important, and timely. These priorities are parameterized by logarithmically scaling them within a range from zero to one.

6.2.2 *Usage Level.* A user's last interaction with the device determines her usage level. If users are engaged in voice commands to the system or browsing recently, they are probably more inclined to hear new notifications and speech feedback. When a new message arrives, its time of arrival is compared with that of the last action taken by the user, scaled based on the system sleep time (default at 15 minutes). High usage is indicated by values closer to one and any message arriving after the sleep time are assigned a zero usage level. Logarithmic scaling ensures that there is less variance in usage values for recent actions relative to that computed during less activity. Actions such as stopping audio playback or deactivating speech are excluded from computing the usage, to avoid creating a recursive usage pattern.

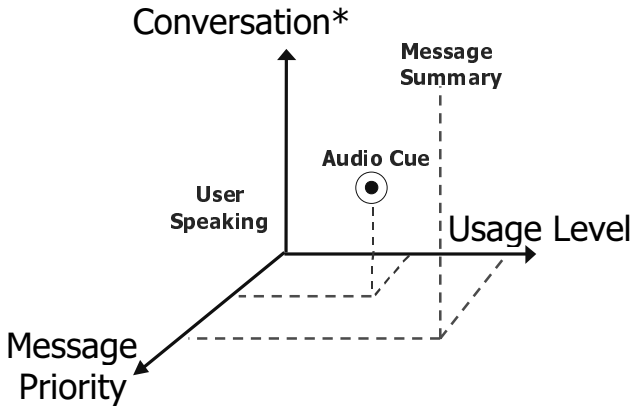

Fig. 8. The contextual notification model; here only an auditory cue is presented for an incoming group message if the user is speaking and not actively using the device. However, a summary is spoken for urgent messages when the conversation level is lower. * The conversation scale is shown inverted for simplicity.

6.2.3 *Likelihood of Conversation.*

Conversation in the environment can be used to gauge whether the user is in a social context where an interruption is less appropriate. If the system detects the user speaking or the occurrence of several speakers over a period of time, that is considered an indication of a conversational situation. Auditory events are first detected by adaptively thresholding total energy and incorporating constraints on event length and surrounding pauses. The system uses mel-scaled filter-bank coefficients (MFCs) and pitch estimates to discriminate, reasonably well, a variety of speech and non-speech sounds. HMMs (Hidden Markov Models) capture both the temporal characteristics and spectral content of sound events. Once the system is trained on a number of auditory events, each is designated as either interruptible (room noise, outdoors, background speech) or uninterruptible (user's voice, foreground speech). When an incoming message is received

the likelihood of such events in the auditory environment allows the system to scale its feedback, minimizing disruption when necessary. The classifier runs in real-time and detects known classes with over 90% accuracy. The techniques for feature extraction and classification of the auditory scene using HMMs are described by Clarkson et al. [1998].

6.2.1 *Notification Level.* A weighted average for all three contextual cues provides an overall notification level. The conversation level has an inversely proportional relationship with notification i.e. a lower notification must be provided during high conversation.

$$Notify_{Level} = ((Priority \times P_{wt}) + (Usage \times U_{wt}) + ((1 - Speech) \times S_{wt})) / 3$$

Here P_{wt}, U_{wt} and S_{wt} are weights for priority, usage and conversation levels. This numerical notification level must be translated to a discrete state, by comparing it to the thresholds for each of the 7 presentation scales, to play the message appropriately.

6.2.3 *Presentation Latency.* Latency represents the period of time to wait before playing a message to the listener, after a notification cue is delivered. Latency is computed as a function of the notification level and the maximum window of time that a lowest priority message can be delayed for playback. The default maximum latency is set to 20 seconds, but can be modified by the user. A higher notification level will cause a shorter latency in message playback and vice versa. For example, an important message will play as a "preview" within 3-4 seconds of arrival (heard as auditory cues), whereas group messages play in "summary" after 10-12 seconds. The use of latency primarily allows a user sufficient time to interrupt and deactivate an undesirable message before it is presented.

6.3 Dynamic Adaptation of Notifications

The user can initially set the weights for the notification model to high, medium, or low (interruption). These weight settings were selected by experimenting with notifications over time using an interactive visualization of message parameters. This allowed us to observe the model, modify weights and infer the effect on notification based on different weighting strategies. Pre-defined weights provide an approximate behavior for the model and help bootstrap the system for novice users. The system allows users to dynamically adjust these weights by their implicit actions while playing or ignoring messages.

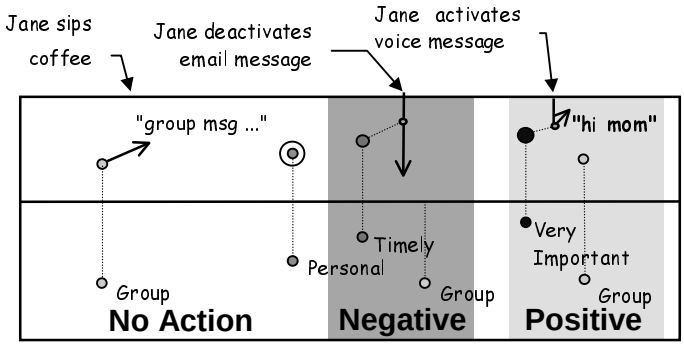

Fig. 9. Adaptation of notification weights based on Jane's actions while listening to messages.

The system permits a form of *localized* positive and negative reinforcement of the weights by monitoring the actions of the user during notifications. As a message arrives, the system plays an auditory cue if its computed notification level is above the necessary threshold for auditory cues. It then uses the computed latency interval to wait before playing the appropriate summary or preview of the message. During that time, the user can request the message be played earlier or abort any further notification for the message via speech or button commands. If aborted, all weights are reduced by a fixed percentage (default is 5%). If the user activates the message within 60 seconds after the notification,

the playback scale selected by the user is used to increase all weights. If the message is ignored, no change is made to the weights, but the message remains active for 60 seconds during which the user's actions can continue to influence the weights.

Figure 9 shows a zoomed view of the extended scenario introduced earlier, focusing on Jane's actions that reinforce the model. Jane received several messages and ignored most of the group messages and a recent personal message (the weights remain unchanged). While in the meeting, Jane interrupted a timely message to abort its playback. This reduced the weights for future messages, and Jane was not notified of lower priority messages. The voice message from Kathy, her daughter, prompted Jane to reinforce the message by playing it. In this case, the weights were increased. Jane was notified of a group message shortly after the voice message, since the system detected higher usage activity. Hence, the system correctly scaled down notifications when Jane did not want to be bothered whereas notifications were scaled up when Jane started to browse messages.

7. PRELIMINARY EVALUATION

The system has been in use by one of the authors for over a year during its development; this aided us in making iterative refinements to the user interface. Once the system implementation was completed on a wearable PC (Nov. '98), it could be reliably used by others on a daily basis. However any evaluation is complicated by the fact that the system must be used in a nomadic environment, and hence social constraints as well as system power, portability, network connectivity and robustness become critical. For the evaluation, we wished to focus primarily on notification (rather than navigation); the nature of this task requires usage over long duration with actual incoming messages (asynchronous and unpredictable), hence close observation was not always possible and we had to rely on the user's own subjective experiences. The preliminary evaluation was conducted with two novice users, wearing the system for 2-3 hours over a period of 3 days; both were active users of mobile phones and 2-way pagers. They were asked to replace their pagers with Nomadic Radio, and use the device both in their offices and while in meetings throughout the evaluation. We sat-in on meetings and observed the users when possible; post-evaluation interviews were also conducted. We recognize that significant variation in feedback may result from a wider study, however we believe that even a limited study allows us to obtain meaningful insights for improving the system.

7.1 Audio UI and Navigation

One user had some difficulty understanding the conceptual model for browsing messages non-visually, until it was described to him later. The user also had trouble recognizing where he was in the message space. One solution is to allow users to ask, "where am I", similar to VoiceNotes or indicate "place" via continuous background sounds. This is a recurring theme in audio interfaces and suggests an area for future work. Users initially utilized explicit spoken confirmations for all voice commands, however over time, as they trusted recognition, they requested responsive auditory feedback only. One user didn't use spoken commands very often; he found it took longer and was socially intrusive especially in meetings. However he did comment that spoken commands would be necessary for navigation while driving or walking. The other user spoke frequently, yet felt uncomfortable speaking in large groups. As expected, users also sometimes forgot commands and had to be reminded. This suggests a simplified vocabulary or an adaptive

mechanism that implicitly learns associations between desired actions and the user's natural utterances over time.

Both users commented that 2-button interaction would be helpful. Although users got accustomed to use "push-to-talk" to initiate a spoken command, they preferred speaking spontaneously. Continuous recognition was not offered in the wearable PC due to its half-duplex hardware and rate of false detection in noisy environments. A potential solution is *contextual recognition*; i.e. let the system briefly listen for voice commands only when it most expects user interaction. Subsequent to the evaluation, we implemented this mechanism to improve responsiveness to incoming messages. When a message is received, the user hears an auditory alert and recognition is automatically activated for a brief window of time until a few seconds after the message stops playing. Informal tests revealed that this novel approach provided a form of natural interaction that users expected, while apparently minimizing user confusion, response time, and recognition errors. Formal study is required to confirm the benefits and tradeoffs involved in *contextual recognition*.

7.2 Notification

Both users were able to listen to notifications while attending to tasks in parallel such as reading or typing. One user managed to have casual discussions with others while hearing notifications; however he preferred turning off all audio during an important meeting with his advisor. However both users sometimes lost concentration if a message was *spoken* while they were listening to others. One user commented that he was able to store part of the message in his temporary memory and could wait till the end of the conversation to retrieve it from memory. People nearby found the spoken feedback distracting if heard louder, however that also cued them to wait before interrupting the user. Users frequently lowered the volume on the device to minimize any disruption to others and maintain the privacy of messages. However they also later increased the volume to hear incoming messages if the surrounding sounds were too loud. Hence both users requested an automatic volume gain that adapted to the environmental noise level.

In contrast to speech-only feedback, the users were more willing to listen to ambient and auditory cues while engaged in other tasks, and these cues seemed to allow them to gradually switch their attention to incoming messages. Familiarization with the auditory cues was necessary. One user preferred longer and gradual notifications rather than an abrupt onset of auditory tones. The priority cues were the least useful indicator whereas VoiceCues provided obvious benefit to both users. Knowing the actual priority of a message was less important than simply having it presented in the right manner. One user suggested weaving message priority into the ambient audio (as increased pitch). He found the overall auditory scheme somewhat complex, preferring instead a simple notification consisting of ambient awareness, VoiceCues and spoken text. The other user desired greater familiarity with the auditory scheme to appreciate its effectiveness and gradually learn to actively listen for the pattern of sequential cues.

Both users stressed that the ambient audio provided the most benefit while requiring least cognitive effort. They wished to hear ambient audio *at all times* to remain reassured that the system was still operational. Even if the system went into *sleep mode*, users requested ambient feedback, rather then silence in the interface. An unintended effect discovered was that a "pulsating" audio stream indicated low battery power on the

wearable device. One user requested a "pause" button, to hold all messages while participating in a conversation, along with subtle but periodic auditory alerts for unread messages waiting in queue. However any such “message hold” facility must automatically deactivate after a short time as the user may no longer be attending to the device, and the system would then be ineffective in monitoring and alerting the user to timely messages. Both users found notifications helpful even at their desktop, since it allowed them to focus on other tasks, while the system “screened” incoming messages. They only read messages on-screen after inferring urgency from auditory cues. The users felt the system provided appropriate awareness and expressive qualities that justified its use over a pager.

7.3 Physical and Social Usage

Users commented that they wanted only one device to handle all their messaging and communication needs, and that its modality should be adapted to the complexity of the information presented (a complementary visual display would be helpful). They requested a wearable PC that was smaller in size, with a battery life of 8-10 hours. Both users found the *Soundbeam Neckset* to be comfortable to wear (and frequently forgot they had it on); however users preferred that it not be tethered to the PC. One user requested discreet notification, while the other didn’t mind if people heard him listen to a message (but not its content). In one situation, the user laughed after hearing a message; his office-mate realized why, as he overheard synthetic speech playing on the Neckset. In terms of physical affordances, one user didn’t mind wearing the Neckset in the workplace (but not outdoors) whereas the other preferred a discreet solution. However they both considered earphones less comfortable and just as noticeable. When asked about discreet use relative to phones, pagers and PDAs, one user remarked that the key difference was that wearables were publicly noticed by default, and could not be easily hidden from view. Would continuous usage over time resolve this social apprehension to wearable devices? This may change as a larger community of users actively begins wearing them. In addition, close attention must be paid to the physical design and social affordances of such a device for everyday usage.

Although one user clearly wished to actively browse messages on the wearable device, the other primarily used it for passive notification. A range of behaviors indicates the need to support both aspects in the interface. However we suspect that in a wider study, most users will utilize notification aspects to a greater extent. Since the system was evaluated for such a short duration, users did not get fully accustomed to it to rely on it for all their messaging tasks; however gradual adoption over time may yield a new range of interaction and social behaviors. Hence a long-term trial with several nomadic users is necessary to further validate these preliminary results.

8. FUTURE WORK

Based on user evaluations, a few key aspects could be further developed in the wearable audio interface. To make the system usable in a variety of situations, tactile input coupled with speech is necessary. Although a numeric keypad was sometimes provided for tactile input, users indicated the need for a simple 2-button interaction mechanism. Users also found it frustrating, not being able to reply to urgent messages that required their timely feedback. However audio transcription of natural speech cannot be reliably used to

compose replies. Hence mechanisms for voice capture and default email or voice replies can be provided in the interface. Finally, we must consider techniques for automatically adjusting the volume of the Neckset relative to environmental sound, to provide the listener a consistent auditory experience. Compact noise-suppressing devices [Aoki et al. 1998] that adjust audio output to the ambient noise level present a possible solution.

In the current system, auditory context is established via classification of the user's voice and that of conversational speech nearby. However the auditory environment provides a rich array of sounds that provide appropriate context to the listener. Sounds of phones ringing, cars honking, trains in the subway, and indoor/outdoor ambience, provide contextual cues that could influence the type of messages presented to the user or operational characteristics of the device. Early experiments indicate that such environmental audio classification is feasible [Sawhney 1997; Clarkson et al. 1998]. The system currently adapts its notification weights based on user actions on subsequent messages, however it does not learn optimal notification from user behavior. To allow the system to generalize an appropriate long-term notification policy for individual users, a variety of statistical and reward-based machine-learning strategies [Kaelbling and Littman 1996] must be explored.

Although the techniques described in this paper have focused on interaction for asynchronous messaging, they are useful for synchronous voice communication. Frequent interaction within a workgroup using an auditory channel creates a social media space, which has its own unique characteristics and affordances [Ackerman et al. 1997]. A mobile user in such audio-only spaces requires a means for establishing awareness of others as well as protocols for initiating contact, turn-taking and interruption. One must consider appropriate forms for filtering, awareness cues, contextual notification and simultaneous listening, that can augment synchronous communication in social spaces.

9. CONCLUSION

This paper began by considering the key affordances and limitations of speech and audio techniques for messaging in nomadic environments. We have demonstrated how a synchronized combination of synthetic speech, auditory cues and spatial audio, coupled with voice and tactile input can be used for navigation and notification on wearables. A key focus was determining user interruptability based on a simple set of contextual cues, and dynamically scaling message presentation in an appropriate manner. Preliminary evaluations indicate that users find auditory awareness and contextual notifications beneficial rather than extensive speech-based navigation and browsing functionality. Users need nomadic audio systems that adapt to background noise and their general usage patterns. Our efforts have focused on wearable audio platforms, however these techniques can be readily utilized in consumer devices such as pagers, PDAs and mobile phones to minimize disruptions while providing timely information to users on the move. It is important to design interaction techniques, keeping in mind the characteristics of the changing physical and social environment within which such devices are actively used.

ACKNOWLEDGEMENTS

Sandy Pentland, Pattie Maes, Hiroshi Ishii, Deb Roy and Tony Jebara provided valuable feedback and critical insight throughout the development of this work. Thanks to Brian Clarkson for ongoing work on the audio classifier. Rehmi Post, Yael Maguire and Bin C. Truong provided assistance with *Soundbeam* modifications. Stefan Marti and Keith Emnett aided us with informal user evaluations. We thank Lisa Fast and Andre Van Schyndel at Nortel for their support of the project. Finally, we wish to acknowledge the ToCHI editors and anonymous reviewers for their insightful comments on this article.

REFERENCES

ACKERMAN, M. S., DEBBY, H., MAINWARING, S. D. AND STARR, B. 1997. Hanging on the 'Wire: A Field Study of an Audio-Only Media Space. ACM Transactions on Computer-Human Interaction, Vol.4, No.1, 39–66.

AKOI, S., MITSUHASHI, K., NISHINO, Y., MURAKAMI, K. 1998. Noise-Suppressing Compact Microphone/ Receiver Unit. NTT Review, Vol.10, No.6, 102–108.

ARONS, B. 1992. A Review of the Cocktail Party Effect". *Journal of American Voice I/O Society*, Vol. 12.

ARONS, B. 1997. Speech Skimmer: A System for Interactively Skimming Recorded Speech. ACM Transactions on Computer-Human Interaction, Vol.4, No.1, 3–38.

BEDERSON, B. B. 1995. Audio Augmented Reality: A Prototype Automated Tour Guide. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'95).* ACM, New York, 210–211.

CHALFONTE, B.L., FISH, R.S. AND KRAUT, R.E. 1991. Expressive richness: A comparison of speech and text as media for revision. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'91)*. ACM, New York, 21–26.

CLARKSON, B., SAWHNEY, N. AND PENTLAND, A. 1998. Auditory Context Awareness via Wearable Computing, *Workshop on Perceptual User Interfaces*, 37–42.

COHEN, J. 1994. Monitoring Background Activities. Auditory Display: Sonification, Audification, and Auditory Interfaces. Reading MA: Addison-Wesley.

FUKUMOTO, M. AND TONOMURA, Y. 1999. Whisper: A Wristwatch Style Wearable Handset. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'99)*. ACM, New York, 112–119.

GARDNER, W. G., AND MARTIN, K. D. 1995. HRTF measurements of a KEMAR. *Journal of the Acoustical. Society of America*, 97 (6), 3907–3908.

GAVER, W.W., R. B. SMITH, T. O'SHEA. 1991. Effective Sounds in Complex Systems: The ARKola Simulation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'91)*. ACM, New York. 85–90.

HORVITZ, E. AND JED L. 1997. Perception, Attention, and Resources: A Decision-Theoretic Approach to Graphics Rendering. In *Proceedings of Uncertainty in Artificial Intelligence*, 238–249.

HUDSON, S. E. AND SMITH, I. 1996. Electronic Mail Previews Using Non-Speech Audio. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'96)*. ACM, New York, 237–238.

KAELBLING, L.P. AND LITTMAN, M.L. 1996. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, Vol.4, 237–285.

MARTIN, G.L. 1989. The utility of speech input in user interfaces. *International Journal of Man Machine Studies*, 30:355–375.

MARX, M. AND SCHMANDT, C. 1996. CLUES: Dynamic Personalized Message Filtering In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CSCW'96)*. ACM, New York, 113–121.

MYNATT, E.D., BACK, M., WANT, R., BAER, M., ELLIS, J.B. 1998. Designing Audio Aura. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'98)*. ACM, New York, 566–573.

O'CONAILL, B. AND FROHLICH, D. 1995. Timespace in the Workplace: Dealing with Interruptions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'95)*. ACM, New York, 262–263.

ROY, D. K. AND SCHMANDT, C. 1996. NewsComm: A Hand-Held Interface for Interactive Access to Structured Audio. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'96)*. ACM, New York, 173–180.

RUDNICKY, A., REED, S., THAYER, E. 1996. SpeechWear: A mobile speech system. In *Proceedings of ICSLP '96*. IEEE.

SAWHNEY, N. 1997. Situational Awareness from Environmental Sounds, MIT Media Lab Technical Report, June 1997. *http://www.media.mit.edu/~nitin/papers/Env_Snds/EnvSnds.html*

SAWHNEY, N., AND SCHMANDT, C. 1998. Speaking and Listening on the Run: Design for Wearable Audio Computing. In *Proceedings of the International Symposium on Wearable Computing*, IEEE, 108–115.

SAWHNEY, N., AND SCHMANDT, C. 1999. Nomadic Radio: Scaleable and Contextual Notification for Wearable Audio Messaging. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'99)*. ACM, New York, 96–103.

SCHMANDT, C. 1994. Multimedia Nomadic Services on Today's Hardware". IEEE Network, September/October, 12–21.

SCHMANDT, C., AND MULLINS, A. 1995. AudioStreamer: Exploiting Simultaneity for Listening. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'95)*. ACM, New York, 218–219.

STARNER, T., MANN, S., RHODES, B., LEVINE, J., HEALEY, J., KIRSCH, D., PICARD, R., AND PENTLAND, A. 1997. Augmented Reality through Wearable Computing. Presence, Vol. 6, No. 4, 386–398.

STIFELMAN, L. J., ARONS, B., SCHMANDT, C., HULTEEN, E.A. 1993. VoiceNotes: A Speech Interface for a Hand-Held Voice Notetaker. In *Proceedings of the ACM INTERCHI '93 Conference on Human Factors in Computing Systems*. ACM, New York, 179–186.

SUZUKI, Y., NAKADAI, Y., SHIMAMURA, Y., NISHINO, Y. 1998. Development of an Integrated Wristwatch-type PHS Telephone. NTT Review, Vol.10, No.6, 93–101.

WENZEL, E.M. 1992. Localization in virtual acoustic displays, Presence, 1, 80.

WHITTAKER, S., HIRSCHBERG, J., NAKATANI, C.H. 1998. All Talk and All Action: Strategies for Managing Voicemail Messages. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'98)*. ACM, New York, 249–250.

YANKELOVICH, N. 1994. Talking vs. Taking: Speech Access to Remote Computers. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'94)*. ACM, New York, 275–276.