Wearable Audio Computing:
A Survey of Interaction Techniques

Deb Roy, Nitin Sawhney, Chris Schmandt and Alex Pentland
Perceptual Computing Group and Speech Interface Group
MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139
{dkroy, nitin, geek, sandy}@media.mit.edu

Abstract

We consider wearable computing applications which rely on audio as a primary medium of the interface. This paper surveys a range of interaction techniques which may be applied to the design of wearable audio computers (WACs). A summary of several speech and audio processing technologies which can be used in the interface of WACs are reviewed. We present several usage scenarios and focus on two specific systems which we are currently implementing. *Nomadic Radio* is a 3-D audio application which provides the user with a personalized and dynamic audio-only information environment. We are also developing an adaptive speech recognition system; an application based on this system enables hearing impaired users to visualize speech. Future research areas include adaptive interfaces, automatic *situational awareness* and *focus of attention* in wearable audio computing.

Keywords

Wearable computing, mobile computing, speech recognition, speech synthesis, audio interfaces, information retrieval, audio indexing

1. Introduction

Most wearable computers today derive their interfaces from concepts in desktop computing such as keyboards, pointing devices, and graphical user interfaces. If wearable computers are expected to become as natural as clothing, we must re-examine the interface with different standards of usability. In this paper we consider the role of audio in the interface of wearable computers.

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 will be 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 driving. A solution in this situation is to use audio to display information.

In this paper we consider wearable computing applications which rely on audio as a primary medium of the interface. We refer to a wearable computer running such applications as a *wearable audio computer* (WAC). Section 2 presents three usage scenarios to motivate the use of WACs. Section 3 surveys several interaction techniques relevant for WACs, and Section 4 reviews several underlying audio and speech interface technologies. Section 5 presents two WAC applications which we are now developing. Finally in Section 6 we comment on future research and conclude.

2. Scenarios for Wearable Audio Computing

2.1 Continuous Audio Capture and Retrieval

Consider a simple note taking application which the user wished to use while engaged in a conversation. Rather than use a slow text-entry device which might distract the user, audio can serve as the input modality allowing speech from the conversation to be recorded transparently[1]. Audio retrieval techniques (discussed in Section 3.2) could then be used to access useful information at a later time. Such an application would be useful, for example, while attending a conference where the flow of spoken information (lectures, meetings and hallway conversations etc.) is typically overwhelming. The system could also be useful for journalists, students and criminal investigations.

2.2 Communications Management

A WAC can be used to manage personal communications naturally. The functionality of cellular phones, pagers

[1]This application immediately raises privacy issues concerning the recording of other people's speech. We acknowledge that this important issue must be addressed for such systems to be used in practice.

and email can be intermingled through a single interface. Synthetic speech can be used to read email and pages to the user; speech recognition can be used to convert the user's responses (with constraints on vocabulary and grammar) into text for email responses. A combination of speech synthesis, recognition, and audio recording, can be used to act as a virtual secretary which manages telephone communications.

2.3 Disability Aids

There are a variety of disability aids which may be built using WACs. Using GPS and digital maps a WAC can provide a visually impaired user with navigation guidance using synthetic speech. A similar system could also act as a guide for sighted individuals visiting a museum. More immediate information about the environment such as obstacles in the user's path can also be sensed using optical sensors and converted into meaningful audio cues. In another application, hearing-impaired users could use a system which converts speech to a visual display to help communicate (See Section 5.3).

3. Interaction Techniques

In this section we present a range of audio interaction techniques which may be used in designing WACs. We divide the techniques into three broad classes: conversational interfaces which employ a spoken conversation metaphor to structure interactions, non-linear audio access techniques which make utilize the random access capabilities of digital audio, and navigable audio information spaces.

3.1 Conversational Interfaces

The design of non-visual interfaces which rely primarily on speech recognition and synthesis may be based on conversational interaction. A spoken conversation metaphor allows a system to create a shared context and protocol with a user to communicate its abilities, constraints, and level of understanding [Hayes83]. It allows a system to gracefully deal with speech recognition errors due to out-of vocabulary utterances and unpredictable phenomenon like background noise. An effective conversational approach [Schmandt94a] utilizes directive and time-out prompts, implicit, explicit and context-free confirmations, and modeling of dialogue states and transitions to better represent conversational flow. Higher levels of linguistic knowledge based on syntactic and semantic constraints as well as discourse structure can aid in conversational modeling for speech interfaces.

SpeechActs [Yankelovich95] uses a spoken conversation metaphor and a speech-only interface to provide users with access to a variety of data, including email, calendar, weather, and information on publicly-traded stocks. The *Conversational Desktop* [Schmandt85] facilitated integrated office telecommunication and message handling using a telephony and desktop speech interface. It allowed users to place calls, take voice messages, record reminders, schedule meetings, and access traffic information. A key feature of the system was to engage users in a sub-dialogue to obtain missing information or correct recognition errors. *NewsTalk* [Herman95] is a speech-only conversational interface that enables interactive retrieval of personalized news over the telephone, using speech recognition, synthesis and digitized news broadcasts.

The MIT Media Lab's Nomadic Computing Environment [Schmandt94b] enables subscribers to manage personal information via fax, pagers and telephony access to digitized and synthesized audio. These services are integrated via *Phoneshell*, which provides remote telephony access to voice mail, email, calendar, rolodex, and a variety of news, weather and traffic information. It displays information via speech synthesis and uses DTMF (touch tones) for input, structured using the conversational metaphor.

3.2 Non-linear Audio Access

The temporal nature of audio makes it a difficult medium to browse in the way we can skim the contents of a page of text. For example it is usually a frustrating experience to find a specific passage lost somewhere on a long tape recording. Digital storage of audio eliminates the serial access constraint of conventional tape media. This leads to an active area of research in interaction techniques for non-linear audio access. We note that in order to have non-linear access the interface needs some structural information about the audio which may be derived automatically (Section 4.3) or manually [Degen92].

Audio content represented as hyper-linked nodes can be browsed using spatial and temporal techniques. *Hyperspeech* [Arons91] is a speech-only hypermedia application that uses speech recognition to maneuver in a database of digitally recorded hyper-linked speech segments without a visual display. *Espace 2* [Sawhney96] is an audio-only environment that consists of a hierarchy of hyper-linked "ambient" spaces containing voice conversations. Temporal audio cues indicate the presence of hyper-links, during the playback of audio content.

VoiceNotes [Stifelman93] was designed as an application for a voice-controlled hand-held computer which allows the creation, management and retrieval of user-authored voice notes. The interface combines speech recognition and button input, allowing the user to combine input modalities depending on what is convenient.

Several systems use speaker segmentation information to provide non-linear audio access. *NewsComm* [Roy96] delivers personalized news and audio programs to mobile listeners through a hand-held playback device. Structural descriptions of the news programs are generated by locating speaker changes and analyzing pause structure. A simple button interfaces lets the user navigate through audio news based on the structural information. Kimber *et.al.* have built a system which automatically divides a multi-speaker recording into speaker segments and displays the information in a graphical browsing tool [Kimber95]. Roy and Malamud have made audio proceedings of the US House of Representatives publicly available on the Internet by making the recordings (comprising several hundred hours of captured speech) searchable by speaker identity [Roy97a].

Some recent mobile applications use spatial memory in handwritten notes to access temporal indices in audio recordings. The *Audio Notebook* [Stifelman96] is a paper-based augmented audio system that allows a user to capture and access an audio recording synchronized with handwritten notes and page turns. *Filochat* [Whittaker94] also co-indexes speech recording to pen-strokes of handwritten notes taken with a digital notebook. Users can access a portion of audio data by gesturing at an associated note. *Dynomite* [Wilcox96] is a portable electronic notebook for the capture, search and organization of handwritten and digital audio notes. *Dynomite* utilizes user-defined ink properties and keywords to dynamically generate new structured views of the recorded information. Synchronized audio is selectively captured via active user highlighting for minimal storage on mobile devices. A similar methodology can be utilized in wearable audio system to index captured audio with contextual information such as user location, time, and current activity for later retrieval as structured views into the user's sonic memories.

A prototype audio augmented reality-based tour guide [Bederson96] presented digital audio recordings indexed by the spatial location of visitors in a museum. This is a early implementation of a wearable audio system which provides only fixed information and does not consider the listener's usage history or model their preferences. *SpeechWear* [Rudnicky96] is a mobile speech system developed at CMU which enables users to perform data entry and retrieval using an interface based on speech recognition and synthesis. A speech-enabled web browser allows users to access local and remote documents through a wireless link. *Ubiquitous Talker* [Nagao95] is a camera-enabled system developed at Sony Research Labs, that provides the user information related to a recognized physical object via display and synthesized voice. It accepts queries through speech input. We believe that a personalized information services and local computation can be provided on wearable devices with audio serving as the primary interaction modality.

3.3 Audio Information Spaces

This section reviews several efforts to display complex information spaces using only audio output.

In contrast to most screen-readers designed for visually impaired users which attempt to convey the entire contents of the visual display via speech synthesis, *Emacspeak* [Raman96] integrates spoken feedback into the application itself. It does so by having the user interface components of the application communicate directly with the speech subsystem. This enables *Emacspeak* to provide rich, context-sensitive spoken feedback. to *Emacs* functions such as document editing, calendar, and WWW access. This approach reduces cognitive load on the user and optimizes the nature of information conveyed.

Several attempts have been made to represent GUIs (Graphic User Interfaces) by mapping speech and auditory cues to visual interface artifacts, applications and window events (the first was *Sonic Finder* [Gaver89]). The *Mercator* [Edwards94] system provided visually-impaired users access to GUI applications on the X-Windows platform. The interface transparently maps the graphical and textual output of existing X-applications to audio and synthesized speech output within the *Mercator* environment. In such environments, the user must continually seek and manipulate acoustic representations of visual artifacts in the GUI. Much unnecessary visual information in the GUIs is often also encoded into audio cues providing no real value to the user. Thus simply adding auditory cues and speech synthesis to existing GUIs is not guaranteed to be an adequate means for providing non-visual access to information.

Auditory icons are short non-speech sounds which are generated to convey the nature of a digital artifact or event and its dynamically changing attributes [Gaver89]. In some situations, objects and events are better represented with continuous patterns of sound rather than discreet auditory icons. Data changing over time or the presence of persistent objects in the environment require the use of continuous audio. Ambient sound textures can be specially designed or algorithmically generated (as earcons) to deliver a perception of object persistence or subtle changes in object attributes. Continuous audio can indicate the presence of background activity [Cohen93] or the sense of location within an audio environment.

Speech and audio interfaces are considered to be sequential, while visual interfaces are simultaneous. The “cocktail-party effect” 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 [Bregman90]. A good model of the head-related transfer functions (HRTF) permits effective localization and externalization of sound sources [Wenzel92]. A spatial sound system can provide a strong metaphor by placing individual voices in particular spatial locations or allowing the user to create a personalized map of the audio recordings in space. The *AudioStreamer* [Schmandt95] 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. Speaker differentiation and pauses were used to find story boundaries (indicated to the listener by changes in tone). Such applications provide access to information through new listening experiences. We must consider how such audio presentation techniques can now be coupled with tactile input to provide non-visual interaction with a wearable audio computer.

4. Supporting Technologies

In this section we review some of the speech and audio technologies which are used in audio interfaces.

4.1 Automatic Speech Recognition (ASR)

Speech recognition is the process of converting a spoken utterance into a text transcript. Speech recognizers may be trained on a specific user’s voice (speaker-dependent) or it may be speaker-independent. Recognizers can either process isolated-word speech requiring the user to pause after each word or can deal with natural continuous speech (see [Rabiner93] for a thorough treatment of speech recognition fundamentals). K*eyword spotting* is the process of detecting specific words or phrases in a stream of speech (for an overview see [Rose96]).

Speech recognition technologies can play an important role in the interface of WACs. It is useful to review some practical issues in using ASR with WACs. Commercial speech recognizers usually contain acoustic models trained on speech from either headset microphones, desk-top microphones, or telephone speech. Alternate setups (which are likely to occur in wearable applications) will degrade the accuracy of such systems. Ideally, an adaptive recognizer which can change its acoustic models should be employed.

We note that most current research in speech recognition is focused on the task of speaker-independent continuous speech. This is largely a result of application areas such as telephone network services and public kiosks which must deal reliably with first-time users. Speech input on WACs will not require some features such as speaker independence. On the other hand it the WAC is to be a natural extension of the user’s clothes, the recognizer must adapt to the vocabulary and pronunciation habits of each user. Take for example the "large vocabulary problem" described by [Furnas87]. Furnas created a simple application with 25 commands and provided subjects with a list of names to be assigned to each command. Furnas found the likelihood that two people chose the same term meant the same command by it was only 15%. One can imagine a speech recognition system for a wearable in which the mapping of words to commands is adapted to each user rather than preprogrammed by the system designer.

Since speech recognizers will likely never reach 100% performance in practical situations, the designer must be careful not to depend on ASR for critical interface functions. In addition it may be socially unacceptable to speak at certain times[1] so a secondary method of input (for example using a keypad or some other tactile input device) should also be considered.

4.2 Structuring Audio

Non-linear access to audio (Section 3.2) requires structural information about the audio recording. This structure provides “handles” into the recording which can help the user efficiently access its contents. Rather than provide indices at arbitrary locations in the recording (for example at equal time intervals in CD players) access may be improved by finding events of interest in the audio recording. This section describes a variety of speech and audio processing technologies which extract structural information from audio.

Short-time energy can be used to find pauses in recordings [Rabiner93]. A variety of methods based on auto-correlation based peak picking may be used to estimate the fundamental frequency of speech. Arons reports that indexing into speech recordings based on pause and pitch structure increased access efficiency [Arons93].

Speaker identification is the process of identifying which member of a previously known set of speakers is most likely to have generated a speech sample [Furui96].

[1] Although we should not be too worried about this in the long run; people once thought that telephones would not succeed since no one would feel comfortable talking into a machine.

Speaker indexing is the process of finding and indexing (assigning unique indices) to every unique speaker in a recording [Gish91, Wilcox94, Roy97b]. Both speaker identification and indexing can be used to index into recordings at speaker change locations for efficient non-linear access.

Sound classification techniques can be used to automatically segment audio. For example speech and music can be separated by characterizing statistical patterns inherent in each class of acoustics. A variety of methods can be used to classify sound patterns including clustering techniques [Wold96] or neural networks [Feiten94].

Similar to speech recognition, the performance of all the technologies mentioned in this section are dependent how clean the recordings are made.

4.3 Digital Audio Recording and Playback

Simple digital audio recording and playback may be used in a variety of ways in the audio interface. As noted earlier, recording may be used to capture the user's acoustic environment as a form of augmented memory. Applications with high storage demands can use compression algorithms to reduce storage without significantly degrading quality (see [Schroeder85] for a popular method to compress speech).

Pre-recorded digital speech may be used to create output prompts for an audio interface. Although the result sounds natural, the main disadvantage is that prompts must be planned in advance; the application designer must predict every necessary prompt and pre-record them. For complex applications this may require impractical amounts of memory.

4.4 Speech Synthesis

A common alternative to recorded prompts is text-to-speech synthesis. Speech synthesis is flexible since arbitrary prompts may be generated at run time by an application. The principal weakness of synthesized speech is that the resulting prosody lacks naturalness. Speech may be synthesized by concatenating sub-word segments of prerecorded speech or by controlling parameters of a vocal tract model [Klatt87]. The results of the two methods are roughly comparable although is it easier to create new voices using concatenation.

5. Applications Under Development

In this section we present two WAC applications which are in the early stages of development. Currently we have completed the construction of two WAC hardware prototypes, and implemented each of the applications on work stations. We are now porting the applications to the WAC platform and also continuing to develop the applications.

5.1 The Prototype Hardware

The current applications will primarily run on the Lizzy wearable computing platform [Starner97] pictured in figure 1. The system currently consists of a 486 100 MHz CPU connected to a variety of I/O devices and memory. The audio related I/O components include an adaptive speech recognizer designed for small vocabulary user trained tasks, full duplex digital audio I/O, and the AT&T FlexTalk text-to-speech synthesizer.

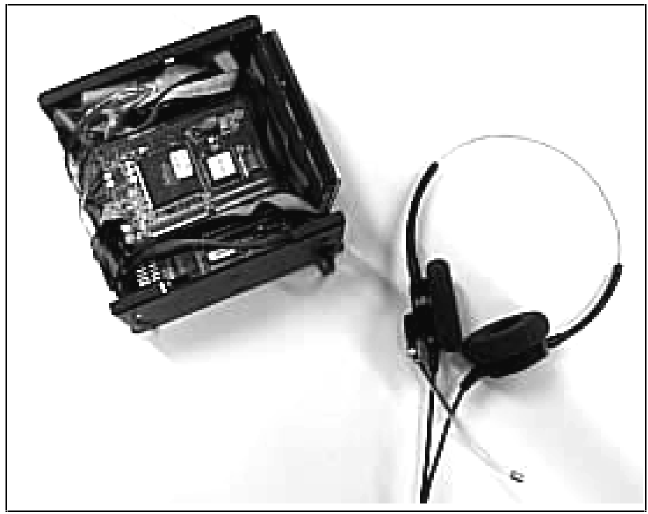

Figure 1: The main CPU and headset microphone

The system includes an optional head mounted display for visual output and a chorded keyboard (the "Twiddler") for tactile input. We are experimenting with several different headphone and microphone configurations and new tactile input devices.

5.2 Nomadic Radio

Wearable computing can be utilized to provide users with highly personalized and timely information, based on the context of their tasks. *Nomadic Radio* is an attempt towards a personalized and dynamic audio-only information environment. It uses the rich metaphor of *Radio* to structure simultaneous and spatialized audio streams as *radio broadcasts* of timely information [Marx96]. The wearable audio system will allow a mobile user (a nomadic listener) to access broadcasts such as voice-mail, news, appointments, weather information, traffic reports, and music. The information is delivered via digitized audio streams previously downloaded by the system from an audio server. The audio streams are presented simultaneously in specific spatial locations in the user's listening space.

During informal demonstrations of *Nomadic Radio* (on desktop PCs), listeners could typically segregate at least three different audio streams of synthesized speech or digital audio content, permitting more effective retrieval and browsing of audio. Some streams are initially presented at the center of the user's listening space, and progressively fade in a desired direction as other audio streams gain priority or the user chooses to shift focus. The location, speed and direction of movement of the audio streams signify the content type, level of urgency and associative characteristics of the information presented. We are currently experimenting with several techniques for the design of dynamic audio streams to maximize audible delivery and related information cues as well as to minimize the potential perceptual load on the listener.

Such asynchronous audio broadcasts allows an effective means of both transmitting and communicating information to users at their own listening pace. In a synthesized email application, *MailCall* [Marx96], *timely messages* were filtered based on the priority and state of the information being presented as well as the context of the user's tasks and related information. For example, if the user has received recent email from John or has a meeting scheduled with him that afternoon, new voice mail from John would be considered timely messages and would have a higher priority as they were dynamically presented in his/her listening space. Users can also send short voice messages to a specific person and asynchronously reply to messages at their own pace. This form of communication is not unlike email or textual MUDs (multi-user domains) with the added convenience and intonational properties of voice.

The current implementation of *Nomadic Radio* is being developed in Java and utilizes Intel's RSX 3D audio libraries. It runs primarily on the Microsoft Windows 95/NT platforms. The wearable versions of the system will communicate with a remote Java client using Wireless LAN to upload recently stored messages from the audio server such as voice-mail, traffic reports or news broadcasts.

Nomadic Radio can be considered an active information agent that will adaptively manage the user's listening space, based on their location, context of activity, prior listening patterns and desired level of interruption. GPS and IR-based positioning will be used for determining a user's contextual location. Reinforcement learning of user's evolving listening patterns will determine the spatial and temporal playback of individual audio streams. We will experiment with live audio streaming over the network and automated generation of structured WWW-based transcripts of recorded audio.

5.3 Robust and Adaptive Speech Input

We are also developing a speech recognition system designed for WACs. Figure 2 shows the main components of the system. Currently only the spectral front-end processor and the recurrent neural network (RNN) have been implemented. After giving an overview of how the system is expected to function, we present an implemented application which uses the spectral analysis and RNN components of the system.

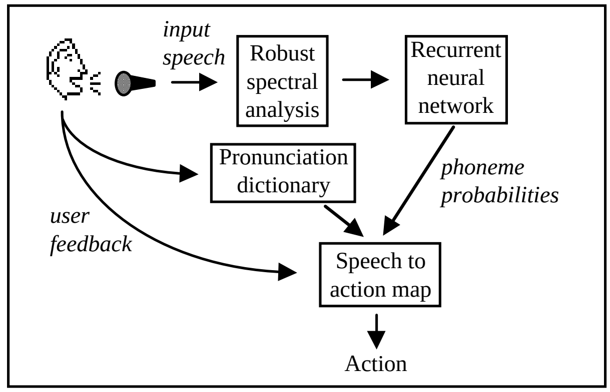

Figure 2: Speech input system for WACs

Overview of the System: The user's speech is analyzed by the RASTA method which has proven to be superior in speech recognition experiments where there are mismatches in the background noise of the training and testing environments [Hermansky94]. This is a good choice for wearable computing applications since the background noise is. The output of the RASTA analyzer is fed into a recurrent neural network (RNN) which produces phoneme probabilities 100 times per seconds.

A dictionary will contain keywords which are used by the interface of the active application. The action-map will spot the keywords and respond using a speech-to-action association map. Feedback from the user regarding the appropriateness of the system's action in response to speech input is used to adapt the dictionary and action map. Over time the system is expected to adapt to an individual user's specific word usage patterns.

The RNN has been trained to perform speaker independent phoneme recognition. The purpose of the RNN is transform the input speech into a speaker independent representation so that dictionaries and action maps may be shared between different users. The speaker independence allows the interface designer to create a generic interface which has some predefined speech input capabilities. The interface can then adapt further to meet the needs of each individual user.

An Application for the Hearing-Impaired: We have developed a hearing aid application which uses the

RASTA processor and RNN to convert speech into a visual display. The intended user is a person who lost hearing later in life. The goal of this application is to map speech input to a human-readable visual representation and make it accessible on a wearable computing platform. As opposed to ASR systems which try to recognize what is spoken, the speech display displays probability levels for each phoneme. The task of recognizing the speech (which requires knowledge of lexical, grammatical, and contextual constraints) is left to the user.

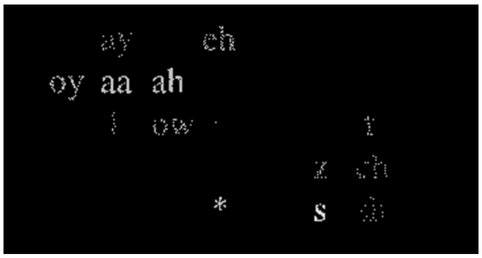

Figure 3: Snap-shot of the speech display midway through the word "such". The phonemes /s/ and /ah/ are highlighted along with other acoustically similar phonemes found by the RNN.

The user could use this system as an aid during face-to-face communication, augmenting the view of the speaker's face with graphical information which may help lip readers or possibly provide sufficient information for stand-alone speech understanding. The system is implemented on an SGI workstation requiring approximately 25% of full load on an R4400 workstation. We are now porting the system to our 486 CPU based. We are also experimenting with alternate designs for the visual design based on usability factors.

6. Conclusions and Future Directions

We have presented a survey of several interaction techniques and interface technologies which are useful for the design of WACs. Desktop computers come equipped with large monitors providing high bandwidth output and essentially obviating the need for audio output. Similarly the keyboard reduces the need for speech input. Wearable computing changes the picture dramatically since the traditional interface components are not incorporated as easily into the situation. We suggest that this is an ideal area to integrate audio as a primary interface medium.

Most of the work we have reviewed in this paper are areas of ongoing research. To conclude we would like to highlight two areas which are particularly challenging but important to consider. Given the personal nature of wearable computers, adaptive interfaces will be especially important. If the computer is to be as natural as clothing, it must be as malleable to the particular whims and idiosyncrasies of each user. Thus researchers need to develop natural interaction methods for interfaces which adapt to the habits and styles of each user.

Finally, in the future, wearable computers must not only have audio input and output, but must know when to talk and when to listen. Audio output must not be generated at inappropriate times. For example a WAC should not start reading email to its owner while she is engaged in conversation. Thus the WAC must have some situational understanding, knowing when it is OK to talk and when it should keep quiet. Similarly the WAC must have some notion of focus of attention. The speech recognizer should not be triggered by car traffic. The note taker should not record the background television (or should it?). The WAC must constantly listen but actively decide when to act upon auditory events and when to ignore them.

Acknowledgments

We would like to thank Thad Starner and Lisa Stifelman for insightful comments on the paper. Thanks to Travell Perkins for setting up speech recognition and synthesis and assembling one of our WAC hardware platforms.

References

[Arons91] Barry Arons. "Hyperspeech: Navigating in Speech-only Hypermedia". *Proceedings of Hypertext '91*, December 1991.

[Arons93] Barry Arons. "SpeechSkimmer: Interactively Skimming Recorded Speech". *Proceedings of UIST' 93*, ACM Press, Nov. 1993.

[Bederson96] Bederson, Benjamin B. "Audio Augmented Reality: A Prototype Automated Tour Guide". *Proceedings of CHI '95*, May 1996, pp. 210-211.

[Bregman90] Bregman, Albert S. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, 1990.

[Cohen93] Cohen, J. Monitoring background activities. Auditory Display: Sonification, Audification, and Auditory Interfaces. Reading MA: Addison-Wesley, 1994.

[Degen92] Degen, L., Mander, R. and Salomon, G. (1992) "Working with Audio: Integrating Personal Tape Recorders and Desktop Computers". *Proceedings of CHI '92*, pp. 413-418, ACM.

[Edwards94] W. Keith Edwards, Elizabeth D. Mynatt, and Kathryn Stockton. "Providing Access to Graphical User Interfaces - Not Graphical Screens". ACM *Proceedings on ASSETS '94*, November 1994.

[Feiten94] Feiten, B. and S. Gunzel, "Automatic Indexing of a Sound Database using Self-Organizing Neural Nets". Computer Music Journal, 18:3, pp. 53-65, Fall 1994.

[Furnas87] Furnas, G., Landauer, T., Gomez, L. and Dumais, S. (1987). “The vocabulary problem in human-system communications”. Communications of the ACM, 30: 964-971.

[Furui96] Furui, S. (1996). An Overview of Speaker Recognition Technology. In: Automatic Speech and Speaker Recognition. Edited by Lee, C., Soong, F., Paliwal, K. Kluwer Academic Press.

[Gaver89] William W. Gaver. The Sonic Finder: An interface that uses auditory icons. Human Computer Interaction, 4:67-94, 1989.

[Gish91] Gish, H., Siu, M., Rohlicek, R. (1991). “Segregation of Speakers for Speech Recognition and Speaker Identification”. *Proc. Int. Conf. Acoustics, Speech and Signal Processing*. Vol. 2 (pp. 873-876).

[Hayes83] Hayes, P. And D. Reddy. “Steps towards graceful interaction in spoken and written man-machine communication.”, International Journal of Man Machine Studies, 19, pp. 231-284, 1983.

[Herman95] Herman, Jeffery Allen. “NewsTalk: A Speech Interface to a Personalized Information Agent”. M.S. Thesis, MIT Media Lab, June 1995.

[Hernamsky94] Hermansky, H. and Morgan, N. (1993) “RASTA Processing of Speech”. IEEE Trans. Speech and Audio Proc. 1(1) pp. 39-49, Jan., 1994.

[Kimber95] Kimber, D., Wilcox, L., Chen, F., Moran, T. "Speaker Segmentation for Browsing Recorded Audio". *Proceedings. of CHI ‘95*, pp. 212-213, ACM, 1995.

[Klatt87] Klatt, D.H. (1987), "Review of text-to-speech conversion for English,". Journal of the Acoustic Society of America, 82(3): 737-793.

[Marx96] Marx, Matthew and Chris Schmandt. “CLUES: Dynamic Personalized Message Filtering”. *Proceedings of CSCW ‘96*, pp. 113-121, November 1996.

[Nagao95] Nagao, Katashi and Jun Rekimoto. “Ubiquitous Talker: Spoken Language Interaction with real world objects”, SONY Computer Science Lab, 1995.

[Rabiner93] Rabiner, L.R. and Juang, B.H. (1993). Fundamentals of speech recognition. Prentice Hall.

[Raman96] Raman, T. V. “Emacspeak --A Speech Interface”. *Proceedings of CHI ’96*, April 1996.

[Rose96] Rose, R.C. (1996) "Word spotting - Extracting partial information from continuous utterances". In: Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer Academic Publishers: 303-330.

[Roy96] 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.

[Roy97a] Roy, Deb and Carl Malamud. (1997) "Speaker identification based test to audio alignment for an audio retrieval system". 1997. To appear in *the Proc. of the Int. Conf. Acoustics, Speech and Signal Processing*, Munich, Vol. 2, pp. 1099-1103.

[Roy97b] Roy, Deb. "Speaker indexing using neural network clustering of vowel spectra". International Journal of Speech Technology, Vol. 1, No. 2:143-149. 1997.

[Rudnicky96] Rudnicky, Alexander, Stephen Reed and Eric Thayer. “SpeechWear: A mobile speech system”, CMU, 1996.

[Sawhney96] Sawhney, Nitin. and Arthur Murphy. “ESPACE 2: An Experimental HyperAudio Environment”, *Proceedings of CHI ’96*, April 1996.

[Schmandt85] Schmandt, C., B. Arons, C. Simmons. “Voice Interaction in an Integrated Office and Telecommunications Environment.” *Proceedings of the 1985 American Voice I/O Society Conference*, pp. 51-57. September 1985.

[Schmandt94a] Schmandt, Chris. Voice Communication with Computers. Van Nostrand Reinhold, 1994.

[Schmandt94b] Schmandt, Chris. “Multimedia Nomadic Services on Today’s Hardware”. IEEE Network, September/October 1994, pp12-21.

[Schmandt95] Schmandt, Chris and Atty Mullins. “AudioStreamer: Exploiting Simultaneity for Listening”. *Proceedings of CHI 95*, pp. 218-219, May 1995.

[Schroeder85] Schroeder, M.R. and Atal, B.S. (1985). "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 937-940.

[Starner95] Starner, T. and Pentland, A. (1995) Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. MIT Technical Report 375.

[Starner97] Starner, T. (1997). Lizzy Wearable Computer Assembly Instructions. MIT Media Laboratory. *http://wearables.www.media.mit.edu/projects/wearables/*

[Stifelman93] Stifelman, Lisa J., Barry Arons, Chris Schmandt, Eric A. Hulteen. “VoiceNotes: A Speech Interface for Hand Held Voice Notetaker”. *Proceedings of INTERCHI ’93*, New York, April 1993.

[Stifelman96] Stifelman, Lisa J. “Augmenting Real-World Objects: A Paper-Based Audio Notebook”. *Proceedings of CHI ‘96*, April 1996.

[Wenzel92] Wenzel, E.M. Localization in virtual acoustic displays, Presence, 1, 80, 1992.

[Whittaker94] Whittaker, S., Hyland, P., Wiley M. "Filochat: Handwritten Notes Provide Access to Recorded Conversations". *Proceedings of CHI '94*, 1994.

[Wilcox94] Wilcox, L., Kimber, D., Chen, F. (1994). "Audio Indexing using Speaker Identification". Xerox PARC ISTL Technical Report No. ISTL-QCA-1994-05-04.

[Wilcox97] Wilcox, Lynn D., Bill N. Schilit, Nitin Sawhney. "Dynomite: A Dynamically Organized Ink and Audio Notebook". *Proceedings of CHI '97*, March 1997, pp. 186-193.

[Wold96] Wold, E., T. Blum, D. Keislar, and J. Wheaton. "Content-based Classification Search and Retrieval of Audio". IEEE Multimedia Magazine, Fall 1996.

[Yankelovich95] Yankelovich, N., G. Levow, and M. Marx. "Designing SpeechActs: Issues in Speech User Interfaces." *Proceedings of CHI '95*. ACM, May 1995.