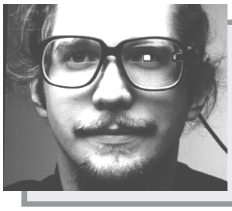


Wearable Computing

Editor: Thad E. Starner ■ Georgia Institute of Technology ■ thad@cc.gatech.edu

The Role of Speech Input in Wearable Computing

Thad E. Starner, Georgia Institute of Technology

Speech recognition seems like an attractive input mechanism for wearable computers, and as we saw in this magazine's first issue, several companies are promoting products that use limited speech interfaces for specific tasks. However, we must overcome several challenges to using speech recognition in more general contexts, and interface designers must be wary of applying the technology to situations where speech is inappropriate.

LOMBARD SPEECH

On a tram in Zürich the other day, my research group started to discuss a counter-terrorism scenario I'm writing for the US Defense Advanced Research Projects Agency (DARPA). Going up a hill, the tram became quite noisy, so naturally I adjusted my voice to compensate. The tram stopped, and I suddenly found myself yelling, in very clear English, "Unless, of course, you are an American and get kidnapped by al-Qaeda." All conversation on the tram ceased.

Although amusing, this story is no doubt familiar. Almost everyone has had a similar experience in a crowded restaurant or at a cocktail party. We speak differently in the presence of noise: increased amplitude, reduced word rate, and clearer articulation.[1] The effect is sometimes called *Lombard speech*, after the French physician who noted it in 1911. Unfortunately, Lombard speech represents a difficulty for using speech recognition on a wearable computer. Most commercial speech recognizers are trained for dictation in an office environment and fail miserably when used in a noisy mobile environment. Although noise-canceling microphones help considerably with reducing noise level, the speaker's change in voice due to ambient noise means that even speech systems trained specifically to understand the user could have problems.

MANAGING MOBILE SPEECH ERRORS

In a recent talk, Chalapathy Neti from IBM showed results that begin to address the first problem of mobile speech recognition, that of background noise.[2] IBM tested its speech recognition system in conditions where it kept adding louder "speech babble" background noise to the clean speech to be recognized. The clean speech had a 19.5-decibel signal-to-noise ratio (SNR), and the recognition engine performed at a 12 percent word error rate on this speech. When researchers added additional noise to achieve an SNR of 12 dB, the word error increased to 60 percent. By 0 dB, word error approached 100 percent! Training the system on speech that was degraded by the same level and source of noise helped considerably: under 30 percent error at 12 dB and 80 percent error at approximately 0 dB. The IBM researchers then added visual features from video of the speakers' mouths to train the speech recognition system. The combined audio and visual feature system performed at better than 50 percent word error rate in 0 dB conditions. This result is better than human performance when only audio is available and comes close to human performance when both audio and video are available. For smaller vocabulary and digit recognition tasks, the system shows even stronger improvement when you use video features and perform training with noise (see Figure 1).[3]

Does this mean that we'll solve our problems with mobile speech or at least approach human performance? Imagine a PDA with a built-in camera running such a system for subway travelers or, similarly, a pocket-sized wireless microphone and camera peripheral that tracks audio-visual features and transmits them to the user's wearable computer. Unfortunately, the issue still remains complicated. Although the IBM results are promising, a speech recognizer trained in one noise domain could have significant problems in another. For example, bursty street traffic noise and microphone noise due to wind could cause different effects than noise in an airplane. This problem will probably be more significant in large-vocabulary, continuous-speech systems, where spurious noise is more likely to interrupt utterances and there are more classes (words) to match against the noise.

Another difficulty relates to the Lombard effect mentioned earlier. We continually change our voices depending on our emotion, stress level, social situation, environment, and many other factors. Imagine a soldier under enemy fire not being able to communicate to his wearable owing to his yelling commands (although smaller vocabularies and utterances might help). Sharon Oviatt's ex-

1536-1268/02/$17.00 © 2002 IEEE

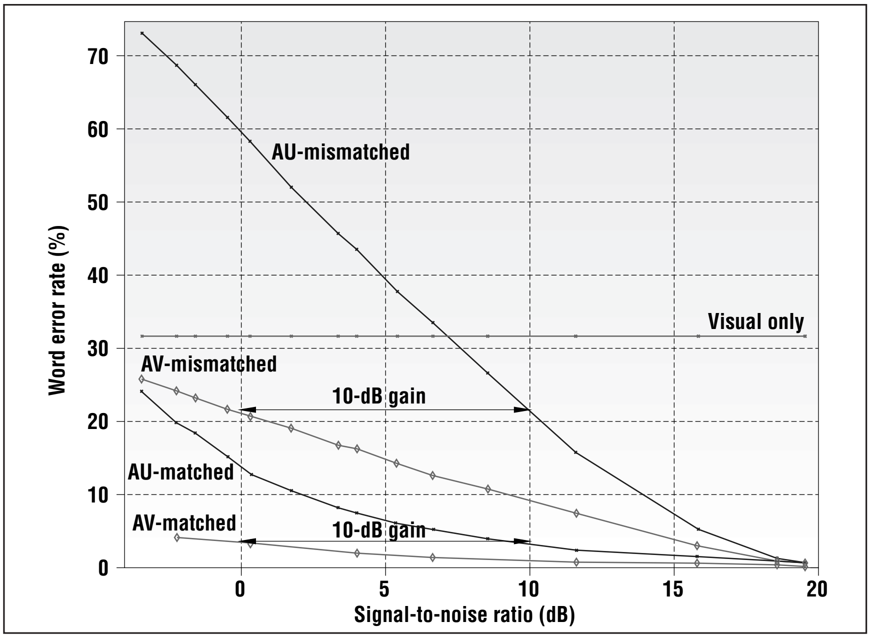


Figure 1. IBM's results for recognizing spoken digits in the presence of increasing levels of added noise. AU indicates audio-only recognition; AV indicates both audio and visual data. Matched means both training and test speech included the same noise level. Mismatched means the training speech was not corrupted with additional noise. Adding visual features improved the results by 10 db.[3]

periments at the Oregon Graduate Institute offer possibilities for less extreme circumstances. Oviatt's research suggests that speakers naturally adapt their voices to better match the speech of a computerized agent with whom they converse.[4,5] Oviatt mentions converging adaptation by the user to the agent's amplitude, pitch, pauses, and latency in response. Could wearable agents use their own "voices" to encourage more easily recognizable speech from their users in mobile settings?

SPEECH INTERFACES ARE NOT THE HOLY GRAIL FOR WEARABLES

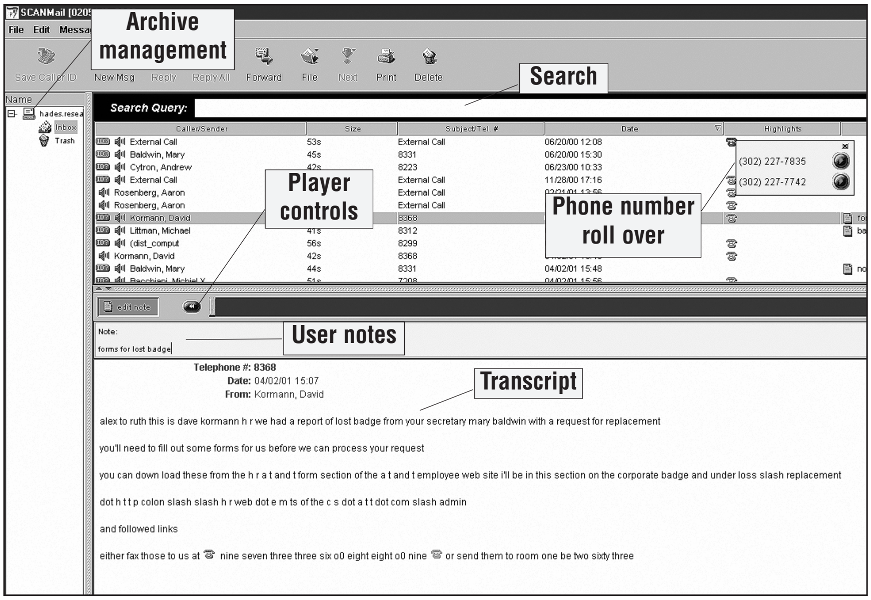
Many commercial wearable computer companies have promoted the combination of speech recognition and a head-up display as an ideal hands-free interface. Reflection Technology, maker of a portable head-up display sold in 1989, suggested such a combination in its advertising. In 1993, I bought a commercial wearable called a CompCap designed to accommodate both Reflection Technology's Private Eye and a limited speech recognizer. The company that made the CompCap, Park Engineering, suggested its wearable for use by telephone linemen. The idea was that linemen, who usually hang from telephone poles, could simply speak the telephone number they wanted to test. Such constrained speech tasks in which digits are spoken or a choice is made from a small list of options are certainly possible in a mobile environment. However, is it possible to create wearable systems with agents you can talk to like a fellow traveler? Maybe the agent could even listen in on the user's conversations and proactively provide assistance through an earphone. Such a conversational interface would be the ultimate in wearable computing, wouldn't it?

Georgia Tech students Fleming Seay, Kuleen Mehta, and Tracy Westeyn prototyped such a system using a continuously open, two-way cellular phone connection to a team of undergraduates who played the part of a conversational wearable agent named Jane. The undergraduates monitored a cell phone line from 10 am to 10 pm; Jane would answer direct queries as well as volunteer information when it felt it was appropriate (for example, "I hear the television. Would you like to know what's on?"). Due to the slowness of the simulated agent's responses to queries (the response to "What is the current score of the basketball game?" might take several minutes of searching online), the emulation of a computerized agent was not as high fidelity as hoped. However, the system did show some distinct limitations to the audio-only approach. First, the agent could not recover enough of the user's context from the audio channel to be proactive except in rare circumstances (such as in the television example). Secondly, using voice to present information proved not to be socially graceful in many situations. For example, the user could not attend to both the agent's voice and the conversation around her at the same time. This conflict caused the user to appear distracted and at times say, "Hold on—Jane is trying to tell me something." Thus, even with humans playing the part of the agent and "hearing" the user's context, the interface proved awkward.

For many everyday applications, social gracefulness could be a major feature. Although cell phones and earbuds let users communicate by seemingly talking into thin air, voice control of a direct manipulation interface (for example, "Move down. Again. Over one. Sell that!") would seem even more out of place if overheard by others than one side of a human conversation would. For such direct manipulation applications, a multimodal interface combining a knob or a keyboard with voice input might be more usable and socially effective.[6–8] In addition, speech is socially interruptive and hard to ignore—bystanders eavesdrop, even if they don't intend to. Correspondingly, the colleagues of a wearable speech interface user might restrain their speech if they think the user's microphone is recording their conversation.

Although the issues of social gracefulness and privacy are particularly highlighted by a wearable's mobility, human computer interface researchers have emphasized some of the pitfalls of using speech recognition as an interface in desktops as well. Ben Shneiderman summarizes these points:

> *Human–human relationships are rarely a good model for designing effective user interfaces. Spoken language is effective for human–human interaction but often has severe limitations when applied to human–computer interaction. Speech is slow for presenting information, is transient and therefore difficult to review or edit, and interferes significantly with other cognitive tasks.*[9]



Caller/Sender	Size	Subject/Tel. #	Date	Highlights
External Call	53s	External Call	06/20/00 12:08	☎
Baldwin, Mary	45s	8331	06/20/00 15:30	
Cytron, Andrew	42s	8223	06/23/00 10:33	
External Call		External Call	11/28/00 17:16	☎
Rosenberg, Aaron		External Call	02/21/01 13:56	☎
Rosenberg, Aaron		External Call		☎
Kormann, David		8368		☎ fo
Littman, Michael	41s	8312		ba
(dist_comput	56s	8299		☎
Kormann, David	42s	8368		☎
Baldwin, Mary	44s	8331	04/02/01 15:48	no
Becchiani, Michiel Y	51s	7208	04/02/01 15:56	☎

(302) 227-7835
(302) 227-7742

Note:
forms for lost badge

Telephone #: 8368
Date: 04/02/01 15:07
From: Kormann, David

alex to ruth this is dave kormann h r we had a report of lost badge from your secretary mary baldwin with a request for replacement

you'll need to fill out some forms for us before we can process your request

you can down load these from the h r a t and t form section of the a t and t employee web site i'll be in this section on the corporate badge and under loss slash replacement

dot h t t p colon slash slash h r web dot e m ts of the c s dot a t t dot com slash admin

and followed links

either fax those to us at ☎ nine seven three three six o0 eight eight o0 nine ☎ or send them to room one be two sixty three

Figure 2. SCANMail uses voice recognition to roughly transcribe voicemail so that the user can visually scan his or her messages. Speech that is recognized as a phone number is marked so the user can find and play that audio segment quickly.[10]

USING VISUAL DISPLAYS TO SEARCH SPEECH SIGNALS

How might we overcome some of the limitations of using speech for input on a wearable computer? Recorded speech, by its nature, is difficult to search for a given piece of information. However, finding a given word or phrase displayed on a high-resolution screen can be very fast. Perhaps we could combine a display with a speech system to allow both fast searching and ease of use. In addition, perhaps pens or keyboards could augment speech interfaces for situations where uttering a sequence of commands to select something on the screen would take longer than a gesture. At the 2002 Conference on Computer–Human Interaction, Steve Whittaker demonstrated a desktop system for voicemail called SCANMail that demonstrates the possibilities.[10] In a previous study using voicemail messages,[11] Whittaker noted that when trying to jot down two facts in a traditional voicemail message, users usually play the segments containing those two facts a combined average of 7.9 times. In his survey on voicemail, 72 percent of the users reported almost always taking notes when playing a message. Whittaker observed that phone numbers and proper names are often the most important pieces of information that a person wants to retain from a voicemail, so he designed SCANMail to support this need.

When a voice message is left on the SCANMail system, the system records the message's context (time of day, caller ID, and so on) and attempts to recognize the speech. SCANMail creates a transcript of the message as best it can and aligns it with the audio's waveform. On the visual interface, the user sees a list of messages, the context information, and a summary of each (see Figure 2). The summary contains recognized snippets that SCANMail believes to be important, such as digits that could indicate telephone numbers. The user can click on these digits to hear the original audio. Although the transcript might be error prone, it is sufficient to cue the user's own knowledge of the information. As Whittaker said during his talk, "If you are dealing with your own voicemail, you know a lot of the content already." On the other hand, when the transcript is not sufficient to cue the user's memory or if the message contains new information, the transcript provides a fast index into the audio waveform. In tests of SCANMail, Whittaker found that users exploited the transcripts often. Moreover, SCANMail users began to view their voicemail as a more permanent, working archive and saved 98 percent of their messages instead of deleting them. SCANMail reduced the need for users to take notes because they could easily retrieve crucial information from the archive.

Imagine such a system applied to a wearable computer, archiving and indexing a user's everyday conversations. Chris Schmandt's students have explored ways of indexing audio using other modalities for quite some time. For example, Lisa Stifelman and her colleagues created an audio notebook that associates audio recorded during a meeting with the pen strokes made on the user's paper notebook.[12] One of my students, Ben Wong, experimented with a similar concept informally in his personal life over a period of six months. To help alleviate concerns of privacy for the people with whom he conversed, Wong used a noise-canceling microphone. Unless someone spoke very loudly, the microphone only picked up Wong's voice. In addition, Wong's speech was immediately converted to text, and only the text was stored. This virtually guaranteed that only one side of a conversation was stored, and Wong reported that most people were comfortable with the situation.

INTERFACES WITH SOCIAL IMPLICATIONS

Soon after Wong began using this system, I started to get compliments on how well-spoken my students were. When pressed, the person would often cite Wong in particular. Curious, I began to pay attention to Wong's speech patterns.

We all know that a good conversational habit is to repeat crucial information to a speaker. This form of "echoing" confirms to both parties that the information was communicated properly and understood to be important. In our weekly meetings, I noticed that Wong routinely exploited this technique. For example, if I mentioned that SCANMail was an interesting paper, Wong would finish the conversation with, "OK, I'll look up that SCANMail paper by Whittaker from CHI for next week." Of course, what Wong was doing was repeating the information so that his voice recognizer would capture enough to remind him later. After the meeting, Wong would heavily edit the transcript, which at that time contained mostly garbage except for the summary phrases he had carefully enunciated. Wong's speech was serving dual purposes: confirming information with an interlocutor and driving his wearable interface.

LOW-RISK INTERFACES

Wong's system also demonstrates another key concept for successful conversational interfaces for wearables: When accuracy is low, the penalty to the user for errors should also be low. In this case, the speech transcription only needed to be complete enough to trigger Wong's natural memory so that he could complete the transcription by hand (in a to-do list, for example).

These ideas of using a wearable display for rapid feedback and using speech for low-penalty interfaces is embodied in the Calendar Navigation Agent (CNA) described last issue.[13] If the system recognizes the key information ("Perhaps I can meet you Friday next week"), the action taken (changing the calendar position to Friday next week) saves the user some time. However, if the system fails, which the user can quickly see in the display, the penalty is that the user must navigate the calendar via the slower, more traditional interface. In other words, the user loses very little from using the system and could gain some convenience.

So, how can we make speech interfaces such as the Calendar Navigation Agent more effective? One simple idea is exploiting a concept similar to "push to talk."

CONSTRAINING THE PROBLEM AND DUAL-USE SPEECH

So, how can we make speech interfaces such as the CNA more effective? One simple idea is exploiting a concept similar to "push to talk." Speech researchers discovered that speakers think out their sentences and articulate more clearly if they have to press a button before they speak to the computer. The CNA requires the user to press a button on the one-handed keyboard at the beginning and end of the phrase to be parsed. In addition, the CNA severely limits what the user can say to certain phrases. These constraints limit the system's "perplexity" for the recognizer. Moreover, we try to design our phrases so as to make them longer because longer constrained phrases are easier for recognizers to handle. Although all these constraints might seem overly burdensome, the phrases can be constructed so as to be socially appropriate and might even be good conversational practice. For example, instead of cueing the CNA with "Friday at 2 p.m.," the user might say, "Yes, I believe I can meet you on Friday at 2 p.m. Please give me a minute to be sure." Fortunately, even when the CNA fails completely, the user can still repair the conversation socially by saying, "Give me one more second. I need to confirm that on my computer's calendar," while he navigates the interface with keystrokes.

NEXT TIME: ATTENTION

Personally, I'm fascinated by the possible dual use of speech for social conversation and simultaneous control of a wearable interface. Why doesn't a wearable user simultaneously navigate his or her calendar via the mouse and continue a conversation at the same time? This thought raises Shneiderman's last point in the quote from earlier—that of cognitive interference. Cognitive interference and human attention are probably the most important issues in the use of wearable computers. They also represent the biggest opportunity. In the next issue, I'll attempt to provide an introduction to these thorny topics. P

REFERENCES

1. J. Junqua, "The Lombard Reflex and Its Role on Human Listeners and Automatic Speech Recognizers," *J. Acoustical Soc. Am.*, vol. 93, no. 1, Jan. 1993, pp. 51–524.
2. G. Potamianos et al., *Large-Vocabulary Audio-Visual Speech Recognition by Machines and Humans*, Eurospeech, Aalborg, Denmark, 2001.
3. G. Gravier, G. Potamianos, and C. Neti, *Asynchrony Modeling for Audio-Visual Speech Recognition*, Human Language Technology, San Diego, Calif., 2002.
4. C. Darves, S. Oviatt, and R. Coulston, "Adaptation of Users' Spoken Dialogue Patterns in a Conversational Interface," to be published in *Proc. Int'l Conf. Spoken Language Processing* (ICSLP 2002), 2002.
5. R. Coulston, S. Oviatt, C. Darves, "Amplitude Convergence in Children's Conversational Speech with

Animated Personas," to be published in *Proc. Int'l Conf. Spoken Language Processing* (ICSLP 2002), 2002.

6. A. Smailagic et al., "Very Rapid Prototyping of Wearable Computers: A Case Study of VuMan 3 Custom versus Off-the-Shelf Design Methodologies," *Design Automation for Embedded Systems*, vol. 3, no. 3, Mar. 1998, pp. 217–230.
7. G. Martin, "The Utility of Speech Input in User-Computer Interfaces," *Int'l J. Man/Machine Studies*, vol. 30, no. 4, 1989, pp. 355–375.
8. C. Schmandt, M. Ackerman, and D. Hindus, "Augmenting a Window System with Speech Input," *Computer*, vol. 23, no. 8, Aug. 1990, pp. 50–56.
9. B. Shneiderman, "The Limits of Speech Recognition," *Comm. ACM*, vol. 43, no. 9, Sept. 2000, pp. 63–65.
10. S. Whittaker et al., "SCANMail: A Voicemail Interface that Makes Speech Browsable, Readable, and Searchable," *Proc. CHI*, ACM Press, New York, 2002, pp. 275–282.
11. S. Whittaker, J. Hirschberg, and C. Nakatani, "Play It Again: A Study of the Factors Underlying Speech Browsing Behavior," *Proc. CHI*, ACM Press, New York, 1998, pp. 247–248.
12. L. Stifelman, B. Arons, and C. Schmandt, "The Audio Notebook: Paper and Pen Interaction with Structured Speech," *Proc. CHI*, ACM Press, New York, 2001, pp. 182–189.
13. B. Wong, T. Starner, and R.M. McGuire, *Towards Conversational Speech Recognition for a Wearable Computer Based Appointment Scheduling Agent*, tech. report 02-17, Georgia Tech., Dept. Graphics, Visualization, and Usability, Atlanta, July 2002.

Thad E. Starner is an assistant professor of computing at the Georgia Institute of Technology, where he directs the Contextual Computing Group in the Institute's College of Computing. Contact him at thad@cc.gatech.edu.

in brief...

Scientists Get Help from Handhelds

By Rebecca Deuel

Field scientists might eventually see an end to pages of handwritten notes and time-consuming data entry, thanks to personal digital assistant software being implemented from Africa to the Arctic.

Researchers in both regions are using CyberTracker, data collection software for the Palm OS platform that runs on Palm handhelds or Handspring's Visor. A CyberTracker field computer lets users gather large amounts of data at a detailed level not previously possible, according to the CyberTracker Web site (www.cybertracker.org).

"One of the most important innovations CyberTracker brings to field research is the use of tracking as a scientific method," CyberTracker World reports, adding that it lets researchers collect data about animals that haven't been disturbed by humans. CyberTracker World is a North American field data collection and environmental training program that combines conservation education with CyberTracker technology. It can also be used with a global positioning system; when the tracker saves the data, the GPS records the location of the observations. With mapping software, researchers can use the GPS information to create detailed maps of animal locations.

Scientists Louis Liebenberg and Lindsay Steventon developed CyberTracker while working with the San bushmen in South Africa. In their paper, "Rhino Tracking with the CyberTracker Field Computer," Liebenberg and Steventon write that expert trackers such as the San can greatly benefit animal research and conservation. Through tracks and signs, they can interpret animal behavior and provide information that might remain unknown to researchers using more conventional methods such as electronic radio tagging. It has been difficult for traditional trackers to document their data however, as most of them are illiterate. CyberTracker has changed that with an icon-based interface that lets the tracker record animal sightings, track observations, species, sex, individual animals, and activities such as feeding, drinking, running, sleeping, fighting, or mating.

In 1996, two trackers began testing CyberTracker at Karoo National Park in South Africa. Liebenberg and Steventon report that despite being unable to read or write, the trackers quickly learned to use the field computer and upload the data themselves.

Researchers are also using CyberTracker software in several parks and game reserves in Africa. One project, at Odzala National Park in the Congo, tracked gorillas and elephants and reportedly recorded 35,000 observations in its first 18 months. The CyberTracker Web site reports that researchers have implemented projects on almost every continent.

Boris von Luhovoy, German magazine *Palm-top Pro*'s editor, told Palm Infocenter that, "Icon approach is the only practical and simple solution given the environment. CyberTracker software enables anybody, no matter which nationality, literate or illiterate, to enter and manage even the most complex data."