An Evaluation of Audio - Centric CMU Wearable Computers

Asim Smailagic
Institute for Complex Engineered Systems
Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

Carnegie Mellon's Wearable Computers project is defining the future for not only computing technologies but also for the use of computers in daily activities. Fifteen generations of CMU's wearable computers are evolutionary steps in the quest for new ways to improve and augment the integration of information in the mobile environment. The complexity of their architectures has increased by a factor of over 200, and the complexity of the applications has also increased significantly. In this paper, we provide a taxonomy of their capabilities and evaluate the performance of audio-centric CMU wearable computers.

1. Introduction

Carnegie Mellon's Wearable Computers project is defining the future for not only computing technologies but also for the use of computers in daily activities. The goal of this project is to develop a new class of computing systems with a small footprint that can be carried or worn by a human and be able to interact with computer-augmented environments. By rapid prototyping of new artifacts and concepts, CMU has established a new paradigm of wearable computers [1]. The fifteen generations of wearable computers have been designed and built over the last five and a half years, and most have been field-tested. The interdisciplinary, user centered, rapid prototyping methodology [2] has lead to a factor of over 200 increase in the complexity of the artifacts while essentially holding design effort constant. The complexity of wearable computer applications has also increased significantly. The application domains range from inspection, maintenance, manufacture, and navigation to on-the-move collaboration, position sensing, global communication, real-time speech recognition and language translation. Since wearable computers represent a new paradigm in computing, there is no consensus on the mechanical/software human computer interface or the capabilities of the electronics. Thus iterative design and user evaluation made possible by our rapid design/ prototyping methodology is essential for quick definition of this new class of computers.

The use of speech and auditory interaction on our wearable computers can provide hand-free input for applications, and enhances the user's attention and awareness of events and personal messages, without the distraction of stopping current actions. It also minimizes the number of user actions required to perform given tasks. The speech and wearable computer paradigms came together in the form of several wearable computers built by CMU, such as: Integrated Speech Activated Application Control (ISAAC), Tactical Information Assistant (TIA-P and TIA-0), Smart Modules, Adtranz and Mobile Communication and Computing Architecture (MoCCA) [3],[4].

Some research addresses wearable auditory displays with a limited scope, such as using them to enhance one environment with timely information [5], and providing a sense of peripheral awareness [6] of people and background events. Nomadic radio has been developed as a messaging system on a wearable audio platform [7], allowing messages such as hourly news broadcast or voicemail to be downloaded to the device.

In this paper, we evaluate the performance of CMU's recently built audio-centric wearable computers and provide a metric for their comparison.

2. Approach

We define two classes of wearable computer applications employing audio processing: Speech Recognition and Collaboration. Figure 1 presents a taxonomy of CMU's wearable computers used for speech recognition, taking into account the type of processing (Local / Remote), and whether they are general or special purpose devices. Most of the wearable computers are general purpose devices, with both remote and local processing capabilities. Speech recognition and/or voice transmission have

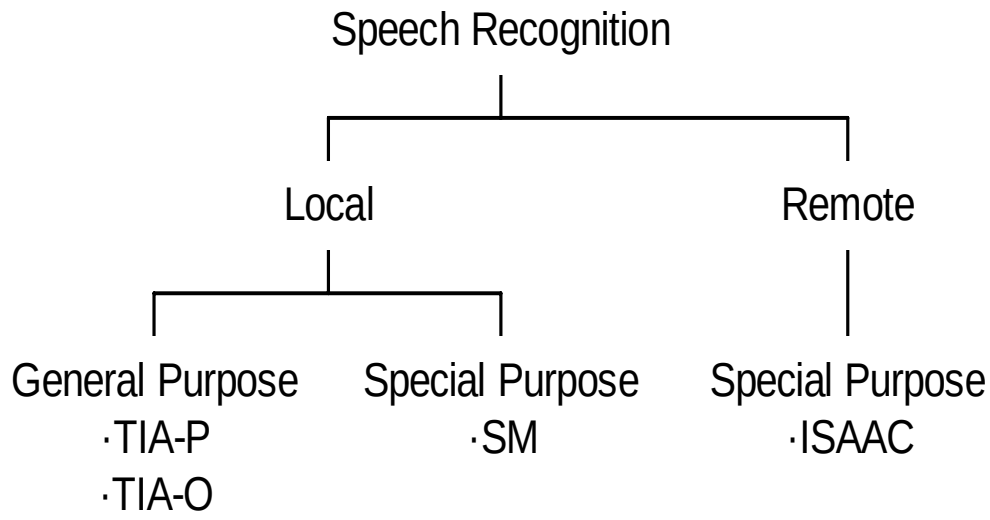

Figure 1: Classification of Wearable Computers in Respect to Speech Recognition

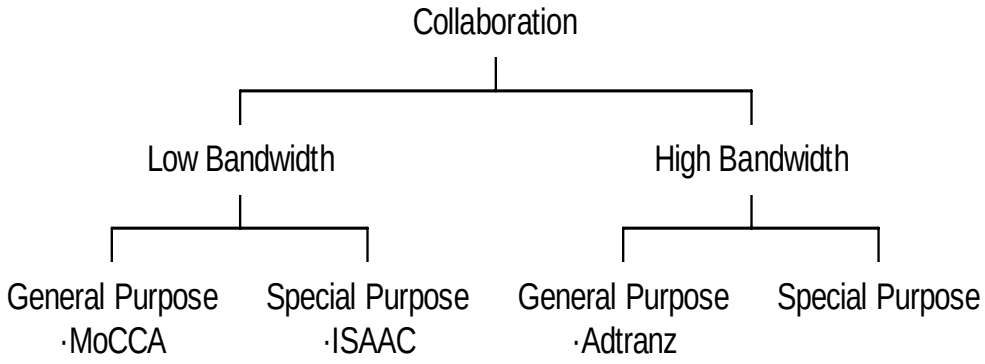

Figure 2: Classification of Wearable Computers in Respect to Collaboration

been used as basic components of collaboration systems. Figure 2 illustrates a taxonomy of CMU's wearable computers for Collaboration, making a distinction between low and high bandwidth systems as well as partitioning them into general and special purpose wearable computers. Their speech applications include audio control, audio transmission, speech recognition and translation. In the next section, we will describe the main characteristics of speech recognition wearable computers developed at CMU.

3. Speech Recognition

Four speech recognition wearable computers will be briefly described in this section.

3.1 ISAAC: Integrated Speech Activated Application Control

ISAAC enables users to control software applications on a base computer (a workstation) using a speech-only user interface and a wearable computer. A wireless microphone transmits analog speech to a speech recognition system on a workstation in the room. The workstation controls devices through infrared repeaters. The workstation also communicates with the user via an infrared wireless headset.

ISAAC is a technology feasibility prototype which explores capabilities in the low weight, low

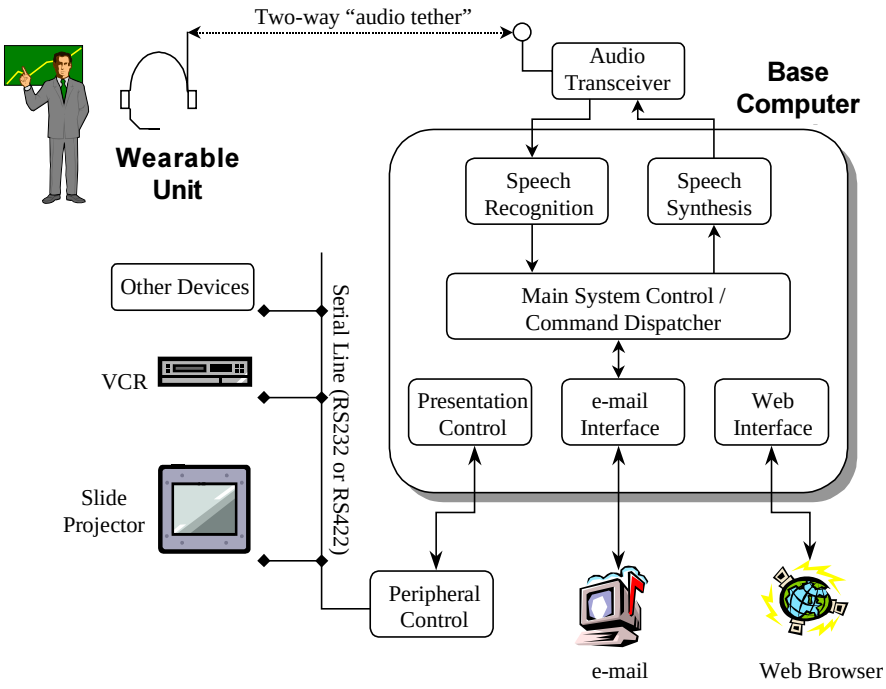

Figure 3 : ISAAC: Architecture of the System

energy design. The goal is to provide users with hands-free control of common office systems from anywhere in an office building. The system supports interaction with a wide variety of applications including email, multimedia presentation control, and web browsing. Each user has a wearable unit which communicates with a base computer via a wireless “audio tether.”

The architecture of ISAAC is shown in Figure 3. Note that the base computer provides the user with all services including peripheral control, e-mail handling, and other functions. The wearable unit has sufficient power to function continuously throughout a typical workday and be able to recharge at night. The base computer provides computational and storage resources to interpret all dialog from the user, process user requests, and generate appropriate responses.

In order to achieve greater freedom and ease of use for the wearable computer, a wireless link between headset and the mobile computer is included. The wireless link is constructed using an infrared transmitter, which alone is normally short ranged (within 2 meters), but when used in conjunction with infrared extenders the range is extended to around 25 meters. The transmitting extenders convert the infrared signal to an RF signal while the receiving extender reverses the process. ISAAC is a voice activated/speech response system. Speech synthesis is delivered over IR to a headset for output. ISAAC weighs less than a pound and consumes less than one watt of power.

Speech Recognition is based on the Speech System Inc. (SSI) board PE1000, which supports speaker independent speech recognition [8]. Table 1 shows the accuracy of this speech recognition during evaluation. In manual operation (push to talk), whenever the success rate was not 100 percent, it was due to an incorrect word, namely the word not being in the vocabulary. However, in voice activation (onset of speech), over half of the errors were due to the lack of recognizing a spoken word in real-time. This was largely caused by the first word spoken being clipped.

Speech synthesis is based on Centigram’s TruVoice software systems, which takes a string of characters and creates spoken output.

	Correct	Nothing Recognized	Recognized Incorrect Word	Percentage Error Caused By A Single Word
Television				
Manual	100.00%			
Voice Activated	71.67%	18.33%	10%	10%
VCR				
Manual	98.00%			
Voice Activated	72.50%	15%	12.50%	10%
E-Mail				
Manual	97.50%			
Voice Activated	81.25%	10%	8.75%	2.50%
Standby				
Manual	95.00%			
Voice Activated	95.00%	5%	0%	5%

Table 1: Speech Accuracy Achieved

Figure 4: TIA-P Wearable Computer

3.2 TIA-P and TIA-0

TIA-P is a commercially available system, developed by CMU, incorporating a 100 MHz 486 processor, 32MB DRAM, 2 GB IDE Disk, full-duplex sound chip, and spread spectrum radio (2Mbps, 2.4 GHz) in a ruggedized, hand-held, pen-based system designed to support speech translation applications. TIA-P is shown in Figure 4. TIA-P supports the Multilingual Interview System/ Language Translation that has been jointly developed by Dragon Systems and the Naval Aerospace and Operational Medical Institute (NAOMI).

The Dragon Multilingual Interview System (MIS) is a keyword-triggered multilingual playback system. It listens to a spoken phrase in English, proceeds through a speech recognition front-end, plays back the recognized phrase in English, and after some delay (~8-10 secs) synthesizes the phrase in a foreign language (Croatian). The other, local person can answer with Yes, No, and some pointing gestures. The Dragon MIS has about 45,000 active phrases, in the following domains: medical examination, mine fields, road checkpoints, and interrogation.

Dragon loads into memory and stays memory resident. The translation uses uncompressed ~20 KB of .WAV files per phrase. There are two channels of output: the first plays in English, and second in Croatian. A stereo signal can be split and one channel directed to an earphone, and the second to a speaker. This is done in hardware attached to the external speaker. An Andrea noise canceling microphone is used with an on-off switch.

Speech translation for one language (Croatian) requires a total of 60MB disk space. The speech recognition requires an additional 20-30MB of disk space.

TIA-P has been demonstrated with the Dragon speech translation system in several foreign countries. TIA-P has supported speech translation demonstrations, human intelligence data collection, and experimentation with the use of electronic maintenance manuals for F-16 maintenance.

TIA-0, show in Figure 5, is a small form factor system using the electronics of TIA-P. The entire system including batteries weighs less than

three pounds and can be mission-configurable for sparse and no communications infrastructures. A spread-spectrum radio and small electronic disk drive provide communications and storage in the case of sparse communications infrastructure whereas a large disk drive provides self-contained stand-alone operation when there is no communication infrastructure. A full duplex sound chip supports speech recognition. TIA-0 is equivalent to a Pentium workstation in a softball sized packaging. The very sophisticated housing includes an embedded joypad as an alternative input device to speech.

Figure 5: TIA-0 Wearable Computer

3.3 Smart Module for Speech Translation

The smart modules are a family of wearable computers dedicated to the speech processing application. A smart module provides a service almost instantaneously and is configurable for different applications. The speech recognition module uses CMU's Sphinx 2 continuous, speaker independent system [9]. The speech recognition code was profiled and tuned. Profiling was done to identify "hot spots" for hardware and software acceleration and to reduce the required computational and storage resources of software. Input to the module is audio and output is ASCII text. The speech recognition module is augmented with speech synthesis. Figure 6 illustrates a combination of the language translation module (LT), and speech recognizer (SR) module, forming a complete stand-alone audio-based interactive dialogue system for speech translation.

Figure 7 depicts the structure of the speech translator, from English (L1) to a foreign language (L2), and vice versa. The speech is input into the system through a microphone, and background noise is eliminated using filtering procedures. Next, the sound is converted into its corresponding phonemes. The list of phonemes is then converted into words using speaker models, dictionaries, and syntactical models. The speaker models are used to determine the linguistic characteristics of the individual users. Dictionaries are used to convert

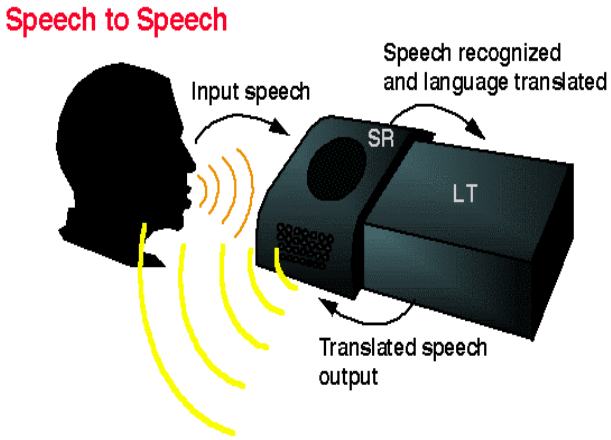

Figure 6: Speech Recognizer (SR) and Language Translator (LT) Smart Module

the phonemes into possible words. Then, the syntactical models are used to decide which of the possible words that could be made is correct. The resulting text is then fed to the Translation module which performs a text to text translation. The clarification dialogue takes place on-screen in the Edit operation, making sure that misrecognized words are corrected. A speech synthesizer performs text to speech conversion at the output stage.

4. Collaboration

Two wearable computers developed for collaboration applications, Adtranz and MoCCA, will be briefly described in this section.

4.1 Adtranz

Adtranz is a mobile pen-based computer enhanced with voice transmission, capability for

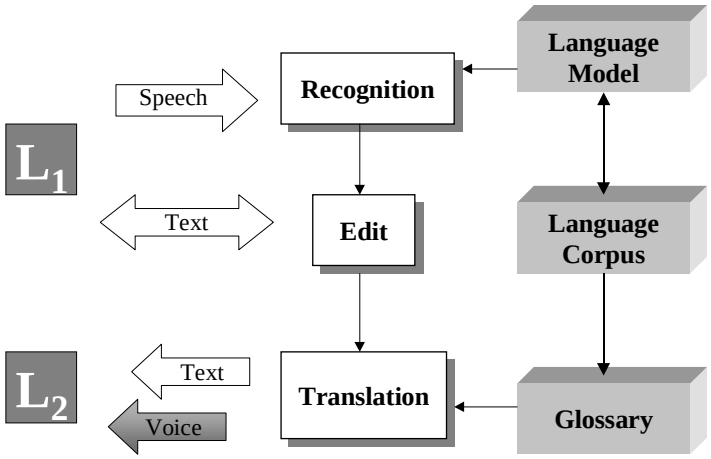

Figure 7: Speech Translator System Structure

collaboration with remote experts, a spread spectrum radio, image capture, and support for a VGA head mounted display. The computer unit includes a 50 MHz 486DX2 processor, 12 MB RAM memory, 170 MB hard disk, two PCMCIA Type II slots, one serial port, one parallel port, one Infra-Red port, and grayscale 640x480 display. The PCMCIA slots are occupied by an AT&T WaveLAN or CDPD (Cellular Digital Packet Data) card, and Wave Jammer sound card used for voice communication, as shown in Figure 8.

Full duplex voice communication is accomplished using a real-time audio program. Voice is digitized using a sound card, compressed, and sent over the WaveLAN network as files using the TCP/IP protocol.

The mobile computer provides access to drawings, schematics, checklists, legacy databases, and manuals for technicians while on or off the train. The content area of the computer display, occupying the left two-thirds of the screen, provides access to documentation and user collaboration capabilities. In collaboration mode, the technician can seek help from a remote expert in repairing a problem. They can collaborate by a whiteboard which allows all members of a session

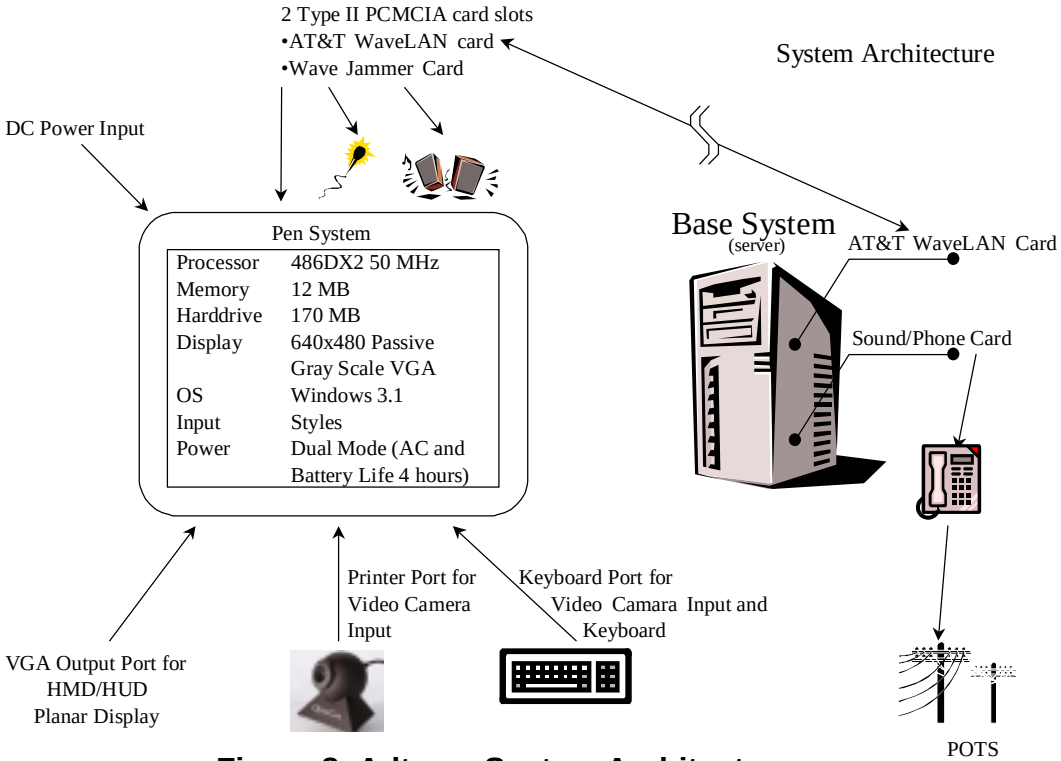

Figure 8: Adtranz System Architecture

to view the same content area, including a picture captured by an on-site camera. The control area allows the user to select dicuments, set bookmarks, enter alarms, etc. The bottom of the display contains a menu bar that allows access to the major usage modes at any time. The major usage modes include: login/out, reference, bookmark, troubleshoot, annotate, and collaboration. Information, annotation, and speech are transmitted over the wireless network. The network can simultaneously transmit high quality digital images, photographs, and two-way full-duplex telephone quality voice communication to a remote expert.

4.2 MoCCA: The Mobile Communication and Computing Architecture

The Mobile Communication and Computing Architecture (MoCCA) was designed to support a group of geographically distributed mobile field service engineers (FSE), using a voice bulletin board for audio transmission. There is no physical contact among the FSEs to share/build corporate memory. The FSEs , working for Digital spend up to 30% to 40% of their time in a car driving to customer sites. Half of what they service is third party equipment for which they may not have written documentation. The challenge is to provide a system that allows the FSEs to access information and advice from other FSEs while on customer sites and while commuting between sites. An additional challenge arose from user interviews which suggested that the FSEs desired all of the functionality of a laptop computer including a larger color display with an operational cycle of at least eight hours. In addition the system should be very lightweight, preferably less than one pound in weight, and have access to several legacy databases that existed on different corporate computing systems.

Figure 9 depicts the components of the MoCCA system, with an FSE shown in the center of the figure:

1. A base unit, about the size of a small laptop computer which is connected to a remote server (located at the home office) wirelessly through a CDPD connection.

2. A cellular phone associated with the base unit and tethered to it through a PCMCIA port. The cellular phone communicates wirelessly with the local cellular provider and thus has access to the telephone network.

3. The FSE holds a smaller satellite unit which is connected to the base unit. The satellite unit shows the contents of the base unit screen and its keyboard input links directly to the base unit keyboard.

4. The FSE wears a microphone and headset that are wirelessly linked to the cellular phone.

The MoCCA system supports audio

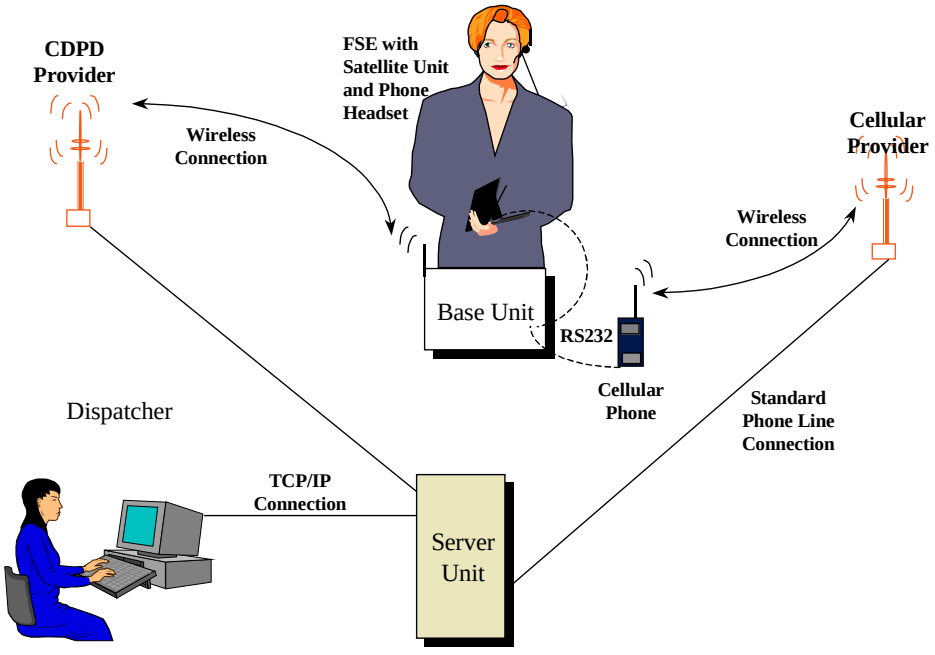

Figure 9: MoCCA System Architecture

transmission via voice bulletin board. The original motivation for the voice bulletin board was to provide a means of asynchronous communication for FSEs, with voice being the medium of interaction. The concept of a bulletin board is equivalent to a storehouse for voice clips describing the problems that Digital's FSEs encounter while on the job. Each "trouble" topic contains a list of voice responses from other FSEs on possible solutions. It differs from a conventional voice mail system because the user interactively creates new topics, listens to topics, selects topics, listens to responses, leaves and archives responses.

Figure 10 depicts the control structure of the voice bulletin board. It is a menu interaction diagram representing the choices presented to the user logged into the MoCCA voice bulletin board system. Each branch in the tree is a decision point for the user. The FSE listens to the available choices (which are numbered) and then makes a selection by pressing the appropriate digit on the phone.

Figure 11 illustrates the use of the MoCCA protoype, showing its display and overall form factor.

5. Experiments and Results

Figure 12 illustrates the response time for speech recognition applications running on TIA-P, TIA-0, and SR Smart Module. As SR is using a lightweight operating system (Linux) versus Windows 95 on TIA-P and TIA-0, and the speech recognition code is more customized, it has a shorter response time. An efficient mapping of the speech

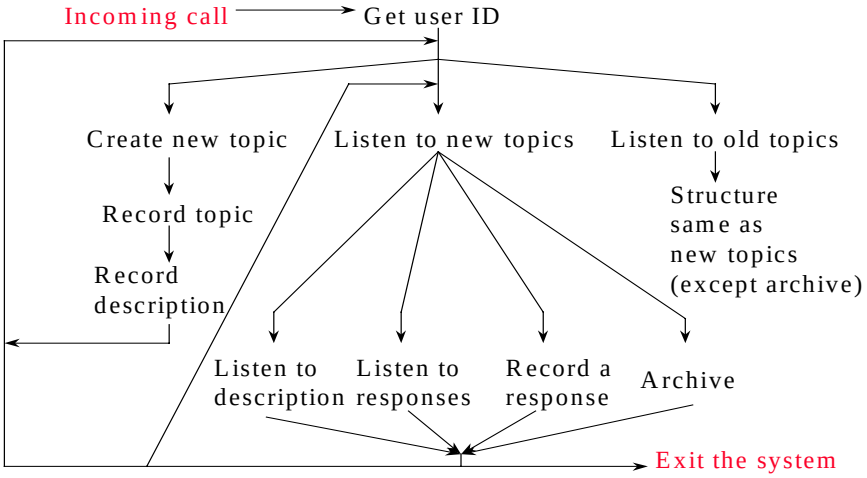

Figure 10: Voice Bulletin Board Control Stucture

Figure 11: MoCCA In Use

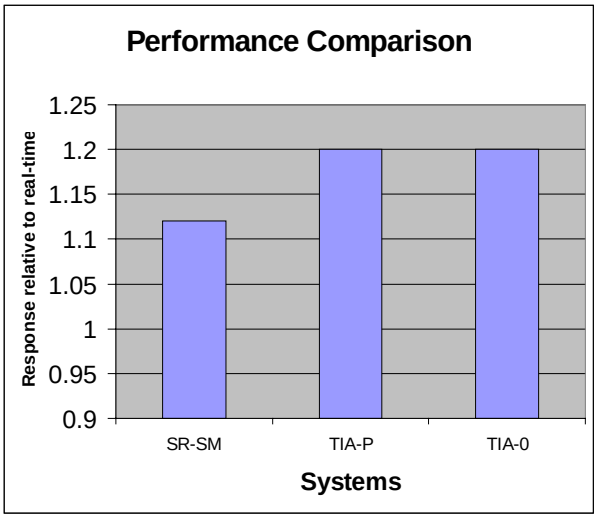

Figure 12: Response Time Comparison

recognition application onto the SR Smart Module architecture provided a response time very close to real-time.

The metric for comparison on Figure 13 is proportional to the processing power (SpecInt), representing performance, and inversely proportional to the product of volume, weight, and power consumption (R), representing resource metrics. It shows the normalized performance scaled by volume, weight, and power consumption. The diagram was constructed based on the data shown in Table 2. A TI 6030 laptop is taken as a baseline for comparison, and its associated value is one. TIA-0 is a factor of 44 better than the laptop while SR Smart Module is over 355 times better than the laptop (i.e., at least a factor of five better in each dimension). Therefore there are orders of magnitude improvement in performance as we proceed from more general purpose to more special purpose wearable computers, as defined in Figure 1. It can be seen that functional specification can yield over two orders of magnitude improvement in composite weight, volume, power, and performance.

Figure 14 presents a comparison of speech

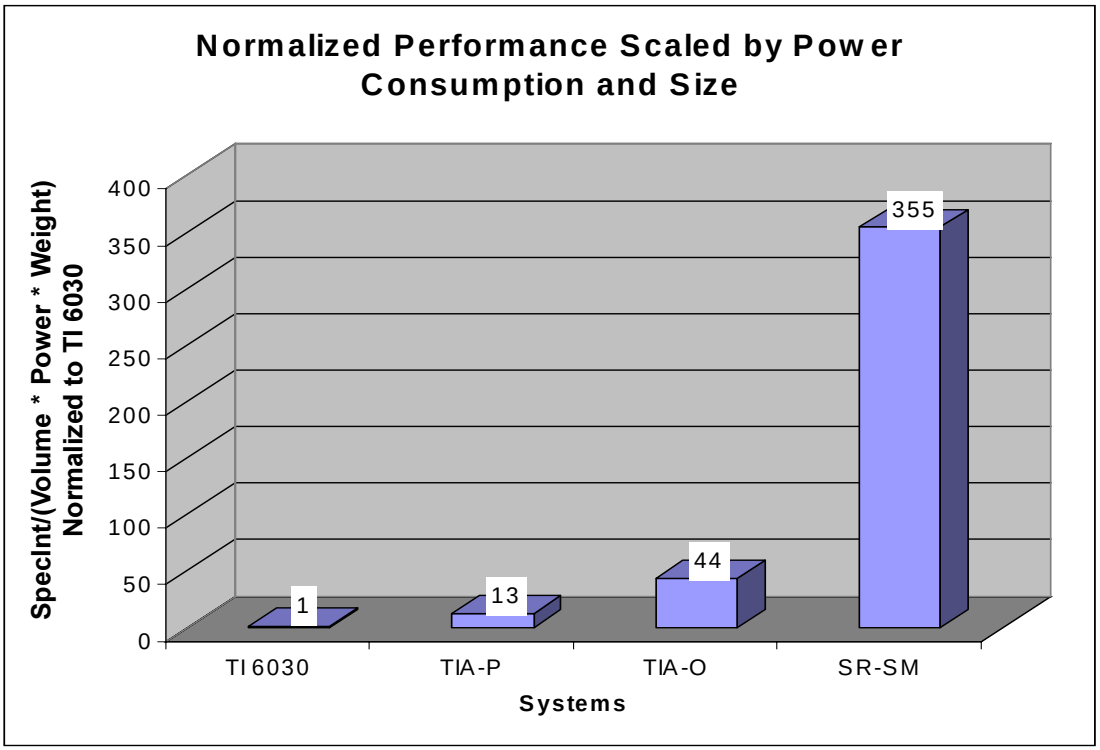

Figure 13: Composite Performance of Speech Recognition Computer Systems

Name	SpecInt	Volume (in3)	Weight (lbs)	Power (watts)	R (V*W-*P)	SpecInt/R	-Log[SpecInt/R]	Normalized - SpecInt/R
TI 6030	175	260	7.5	36	70200	0.002	2.603	1.000
TIA-P	55	88	3	6.5	1716	0.032	1.494	12.857
TIA-O	55	45	2.5	4.5	506.25	0.109	0.964	43.581
SR-SM	175	33	1.5	4	198	0.884	0.054	354.545

Table 2: Performance Values Measured and Calculated for Wearable Computers

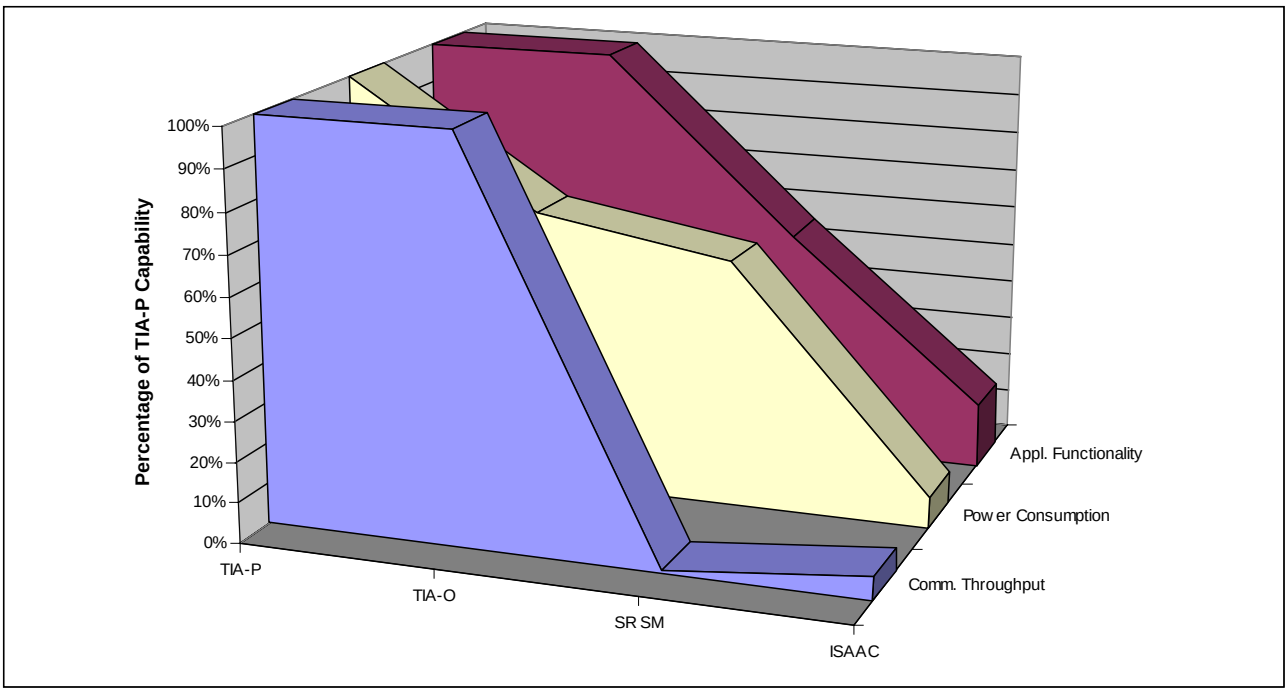

Figure 14: System Comparison Relative to TIA-P

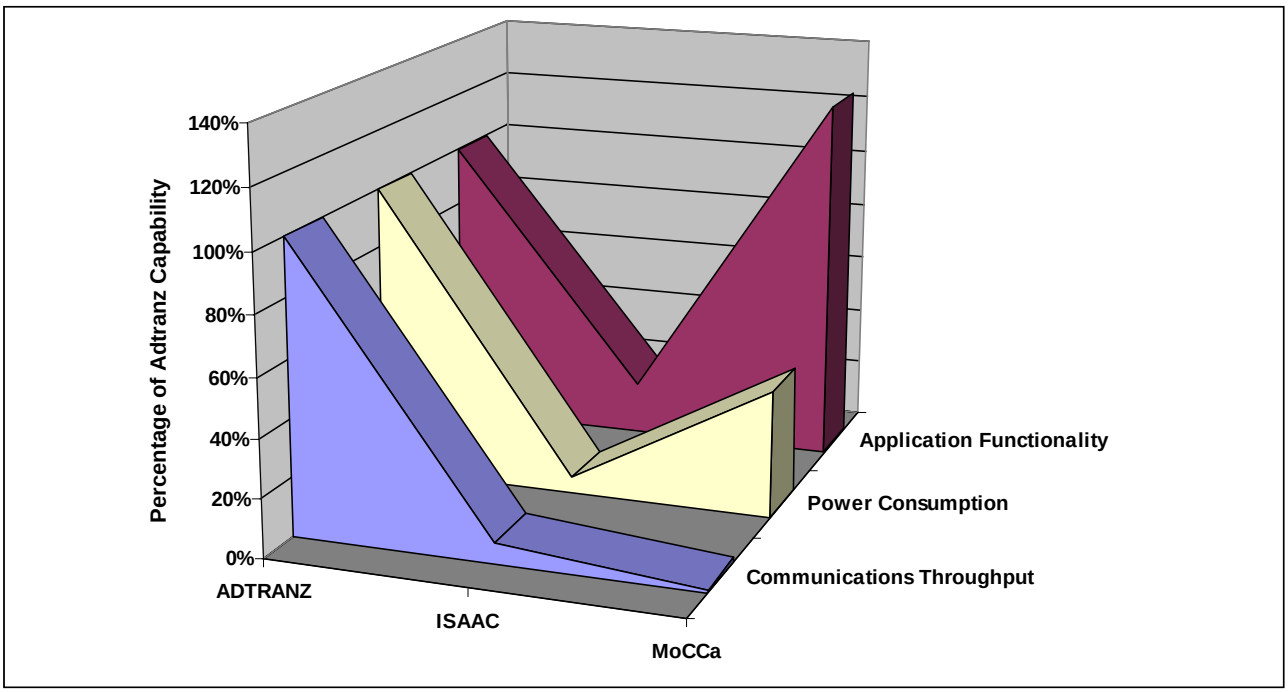

Figure 15: System Comparison Relative to Adtranz

recognition wearable computers (TIA-P, TIA-0, SR SM, ISAAC) in respect to application functionality, power consumption, and communication throughput. All results are displayed relative to the TIA-P system. Functionality determines what the system performs, including services provided [10]. The functionality graph indicates a system's ability to handle audio, text, diagrams, photographs, and full motion digital video. The diagram shows a continuous trade-off between functionality and resources consumed. General purpose speech recognition computers (TIA-P, TIA-0) provide the highest functionality, but their power consumption is also high. For example, relative to the TIA-P power consumption, TIA-0 consumes 69%, Smart Module 61%, and ISAAC 7.7%.

Figure 15 compares the collaboration wearable computers (as defined in Figure 2) in

respect to power consumption, functionality, and communications throughput. All results are displayed relative to the Adtranz. The diagram nicely illustrates what the cost is for going with minimum functionality and getting more functionality (MoCCa and Adtranz), where cost is represented by power consumption. Also, ISAAC and MoCCa demonstrate the advantage of using a satellite unit to reduce the worn part of the computer and still be able to access the desired functionality.

6. Conclusions

The fifteen generations of wearable computers have been designed and built over the last five and a half years. Most of them were field-tested. The interdisciplinary, user centered, rapid prototyping methodology has led to an increase in the complexity of the artifacts by a factor of over 200 while essentially holding design effort constant. The complexity of the wearable computer applications has also increased significantly. Two classes of wearable computer applications employing audio processing have been defined: Speech Recognition and Collaboration. In the speech recognition application, there are orders of magnitude improvement in performance as we proceed from one generation of wearable computers to the next one.

To our knowledge, TIA-P, TIA-0, and SR-LT Smart Modules are the only wearable computers capable of performing two-way speech translation (involving speech recognition and language translation). As general purpose computers, TIA-P and TIA-0 can effectively be used for collaboration applications.

This paper presents a taxonomy and approach to evaluate performance of the wearable computers. The results of this research should allow us to set the design direction and make appropriate decisions for the future advanced wearable computer systems.

6. Acknowledgment

This work was supported by Defense Advanced Research Project Agency and Institute for Complex Engineered System at Carnegie Mellon University.

7. References

[1] A. Smailagic and D. P. Siewiorek, “The CMU Mobile Computers: A New Generation of Computer Systems,” Proceedings of the IEEE COMPCON 94, IEEE Computer Society Press, February 1994, pp. 467-473.

[2] D.P. Siewirorek, A. Smailagic, and J.C. Lee, “An Interdisciplinary Concurrent Design Methodoogy as Applied to the Navigator Wearable Computer System,” Journal of Computer and Software Engineering, Vol. 2, No. 2, 1994, pp 259-292.

[3] A. Smailagic, “ISAAC: A Voice Activated Speech Response System for Wearable Computers,” Proceedings of the IEEE International Conference on Wearable Computers, Cambridge MA, October 1997.

[4] A. Smailagic and D. P. Siewiorek, “Interacting with CMU Wearable Computers,” IEEE Personal Communications, vol. 3, no. 1, February 1996, pp. 14-25.

[5] B. Bederson, “Audio Augmented Reality: A Prototype Automated Tour Guide,” Proc. of CHI ‘95, May 1996, pp. 210-211.

[6] E.D. Mynatt, M. Back, R. Want, and R. Frederick, “Audio Aura: Light-Weight Audio Augmented Reality,” Proceedings of UIST ‘97 User Interface Software and Technology Symposium, Banff, Canada, October 15-17, 1997

[7] N, Sawhney and C Schmandt, “Design of Spatialized Audio in Nomadic Environments,” Proceedings of the International Conference on Auditory Display, November 2-5, 1997, Palo Alto, CA.

[8] PE1000 Documentation, Speech Systems Inc., Boulder, Colorado, 1996.

[9] K.F. Li, H.W. Hon, M.J. Hwang, and R. Reddy, “The Sphinx Speech Recognition System,” Proc. IEEE ICASSP, Glasgow, UK, May 1989.

[10] D. P. Siewiorek, “Multidisciplinary Design,” Private Communications, 1998.