

CMU Wearable Computers for Real - Time Speech Translation

Asim Smailagic, Dan Siewiorek, Richard Martin, Denis Reilly
Institute for Complex Engineered System, Carnegie Mellon University
Pittsburgh, PA 15213, USA
{asim, dps, martin+ }@cs.cmu.edu, dpr@andrew.cmu.edu,

Abstract

*Carnegie Mellon's Wearable Computers Laboratory has built four generations of real - time speech translation wearable computers, culminating in the Speech Translator Smart Module. Smart Modules are a family of interoperable modules supporting real-time speech recognition, language translation, and speech synthesis. In this paper, we examine the effect of various design factors on performance with emphasis on modularity and scalability. A system-level approach to power / performance optimization is described that improved the metric of (performance / (weight * volume * power)) by over a factor of 300 through the four generations.*

1. INTRODUCTION

The goal of CMU's Wearable Computer project is to develop a new class of computing systems with a small footprint that can be carried or worn by a human and interact with computer-augmented environments. By rapid prototyping of new artifacts and concepts, CMU has established a new design paradigm for wearable computers [1],[2]. Eighteen generations of wearable computers have been designed and built over the last seven and a half years, with most field-tested. One of the application domains is real-time speech recognition and language translation.

Bringing computing-intensive applications to a wearable platform means that users have mobile access to those applications at any time and any place. A well designed wearable computer should make using computing-intensive applications almost as easy and intuitive as using a hand tool.

There are several criteria that can be of use when designing a wearable system:

- Keep the latencies involved with running the operating system (OS) and the application low (as close to "instant response" as possible, such as a flashlight).
- Make the battery life as long as possible (reduce power consumption)

- Make the interface to the software as intuitive as possible.
- Make the form factor of the device as unobtrusive as possible, specifically lightweight and operable in multiple orientations.

The Smart Module project adds two more criteria to wearable computer design. These wearable devices must be modular; they should be usable in different configurations. They must also be scalable; existing code should be easily portable to the modules. By using a known OS, the modules have the potential to run a wide variety of applications supported by its hardware. The OS chosen was Red Hat Linux, because it is free, lightweight, scalable, customizable, and a variety of applications already ran on the Linux platform.

This paper will focus on the first two goals of improving performance and reducing power consumption. These goals seem to be inherently contradictory at first glance: any computing device that runs at a high clock frequency will tend to consume more power. This paper measures how close the Smart Module project is to achieving these goals.

The use of speech and auditory interaction on wearable computers can provide hands-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 a series of wearable computers built by CMU, including: 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],[5].

There have been several explorations into wearable auditory displays, such as using them to enhance one's environment with timely information [6], and providing a sense of peripheral awareness [7] of people and background events. Nomadic radio has been developed as a messaging system on a wearable audio platform [8], allowing messages such as hourly news broadcast or voicemail to be downloaded to the device.

Most of these prior systems have focused on speech recognition and speech synthesis. Language translation presents one additional challenge for wearable computers.

2. EVOLUTIONARY METHODOLOGY

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 helping define this new class of computers.

The four generations of real - time speech translation wearable computers span from general purpose to dedicated computers: TIA-P, TIA-0, Speech Translator Functional Prototype Smart Module, and Optimized Speech Translator Smart Module. This evolution was based on lessons learned from their field tests and deployment. These four systems were developed as two related pairs. The first member of each pair was a functional prototype that was suitable for field evaluation. The second member was optimized for power consumption, size, weight, and performance. The feedback from field tests guided the design of the next version.

These systems had attributes which were the same for all four as well as attributes which were varied to achieve improved designs:

Constants

- Speech Recognition (SR) / Language Translation (LT) Application
- Cardio Processor Subsystem

Variables

- System and Software Architecture
- User Interface

2.1 SR / LT Application

The SR / LT application is a speech translation process which consists of three phases: speech to text language recognition, text to text language translation, and text to speech synthesis. The application running on TIA-P and TIA-0 is the Dragon Multilingual Interview System (MIS). It is a keyword-triggered multilingual playback system, which 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. Therefore, a key characteristic of this application is that it deals with a fixed set of phrases, and includes one-way communication.

The Speech Translator Smart modules (Functional Prototype and Optimized) run a freeform, continuous speech translation application, including two-way communication. The modules use CMU language translation and speech recognition software that was profiled to identify "hotspots" for software and hardware acceleration. TIA-P and TIA-0, as uniprocessor units, would not be appropriate for this application and we decided to proceed with a dual processor dedicated architecture (smart modules) to decrease size and response time. The first module incorporates speech to text language recognition and text to speech synthesis. The second module performs text to text language translation.

2.2 Cardio Processor Subsystem

The core of all four speech translation wearable computers is the Cardio processor card, which combines the processor and many of the motherboard chips into one package, about the size of a PCMCIA card [12]. The hardware architecture of the modules is illustrated in Figure 1.

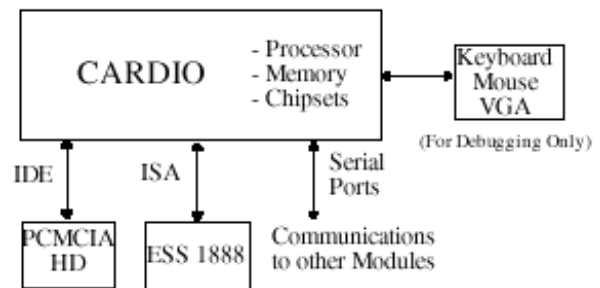


Fig. 1. Smart Module Hardware Diagram

All the necessary signals for the ISA and IDE buses come out of the Cardio card. The Cardio also supports two serial ports, which are used for communication between the modules, and a VGA interface. The ISA and IDE buses both typically operate at 8 MHz, with a width of 16 bits. The ISA bus is limited to 8 MB/s throughput, while the IDE interface can achieve up to 13 MB/s throughput. Main memory is significantly faster – although the Cardio data sheet [9] does not have complete information on the internal memory bus of the Cardio, a reasonable estimate is that the 133 MHz 586-based Cardio has at least a 33 MHz system bus with a width of 32 bits. Even with a wait

state, the memory architecture is speculated to move 66 MB per second.

2.3 System and Software Architecture

The main difference in the system architecture is due to TIA-P and TIA-0 speech translation application being one-way, and smart modules perform two-way speech translation.

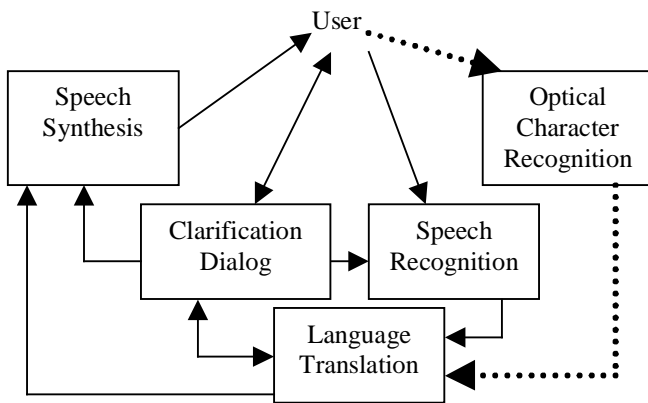


Fig. 2. Two way Speech Translator System Structure

Figure 2 depicts the structure of the free form, two way speech translator, from English to a foreign language, and vice versa. The speech is input into the system through the Speech Recognition subsystem. A user wears a microphone as an input device, and background noise is eliminated using filtering procedures. The Language Translation module includes a language model, glossary, and machine translation engine. The language model, generated from a variety of audio recordings and data, provides a knowledge source about the language properties. The Example-Based Machine Translation (EBMT) engine translates individual "chunks" of the sentence using the source language model and then combines them with a model of the target language to ensure correct syntax. The glossary is used for any final lookups of individual words that could not be translated by the EBMT engine. When reading from the EBMT corpus, the system makes several random-access reads while searching for the appropriate phrase. In small wearable systems the language corpus is stored on disk. Since random reads are done multiple times, instead of loading large, continuous chunks of the corpus into memory, the disk latency times are more important than the disk bandwidth.

The Speech Synthesis subsystem performs text to speech conversion at the output stage. To make sure that misrecognized words are corrected, a Clarification Dialog takes place on-screen. It includes the option to

speak the word again, or to write it. As indicated in Figure 2, an alternative input modality could be the text from an Optical Character Recognition subsystem (such as scanned documents in a foreign language), which is fed into the Language Translation subsystem. The Smart Modules software architecture is described in section 5.

Figure 3 illustrates the one way speech translator, based on the Multilingual Interview System (MIS) that has been jointly developed by Dragon Systems and the Naval Aerospace and Operational Medical Institute (NAOMI), and runs on TIA-P and TIA-0. The user (interviewer) selects a domain module and target

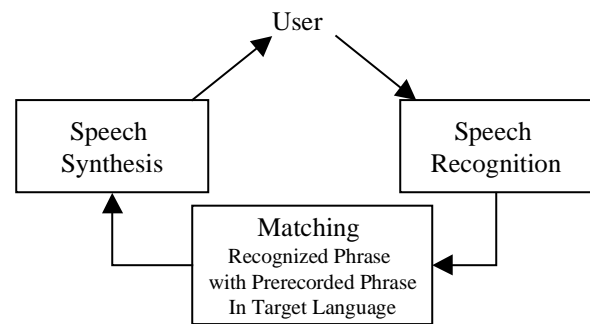


Fig. 3. One way Speech Translator System Structure

language, then selects and speaks phrases from a set of prerecorded phrases. The speech recognition system uses Dragon Dictate. In the next step, matching of a recognized phrase with the prerecorded phrase in a target language is performed (phrase book lookup), and the prerecorded phrase is played back at the output stage (speakers). The phrases are designed to elicit brief responses (yes or no) or gestures.

2.4 User Interface

User interface design went through several iterations based on feedback received during field tests. The emphasis was on correct two - way speech translation, and having an easy to use, straightforward interface for the clarification dialogue.

3. TIA-P AND TIA-0

Our first two systems built in a family of wearable computers dedicated to speech translation applications were TIA-P and TIA-0.

3.1 TIA-P

TIA-P is a commercially available system, developed by CMU, incorporating a 133 MHz 586 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



Fig. 4 TIA-P Wearable Computer

to support speech translation applications. TIA-P is shown in Figure 4. TIA-P supports the Multilingual Interview System.

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

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.

TIA-P has been tested with the Dragon speech translation system in several foreign countries: Bosnia



Fig. 5 U.S. Soldier in Balkans Using TIA-P

(Figure 5), Korea, and Guantanamo Bay, Cuba. TIA-P has also been used in human intelligence data collection and experimentation with the use of electronic maintenance manuals for F-16 maintenance.

Operational Experience

The following lessons were learned during the TIA-P field tests: wires should be kept to a minimum; handheld display was convenient for checking the translated text; standard external electrical power should be available for use internationally; battery lifetime should be extended; ruggedness is important. All these lessons were used as an input into the design of the optimized version, TIA-0.

3.2 TIA-0

The main design goals for the TIA-0 computer were shrinking the size, reducing the weight, and incorporating the lessons learned from the TIA-P field tests. TIA-0, shown in Figure 6, is a smaller form factor system using the electronics of TIA-P. The entire system



Fig. 6 TIA-0 Wearable Computer

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

sophisticated housing includes an embedded joypad as an alternative input device to speech.

4. SMART MODULE APPROACH

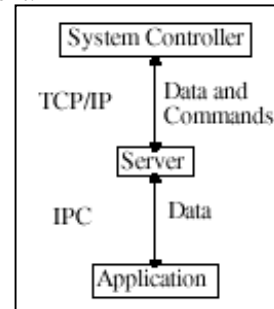
Smart modules are a family of wearable computers dedicated to speech processing application. A smart module provides a service almost instantaneously and is configurable for different applications. The design goals also included: reduce latency, eliminate memory context swaps, and minimize weight, volume, and power consumption. The functional prototype consists of two specialized modules, performing language translation and speech recognition. The speech recognition module uses CMU's Sphinx 2 continuous, speaker independent system [10],[11]. The speech recognition code was profiled and tuned. Profiling identified "hot spots" for hardware and software acceleration and places to reduce computational and storage resources. Input to the module is audio and output is ASCII text. The speech recognition module also supports text to speech synthesis. Figure 7 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. As a result of the profiling, we have achieved a five times smaller memory requirement in comparison to the software desktop version.

The LT module runs the PANLITE language translation software [12], and the SR module runs CMU's Sphinx II Speech Recognition Software and Phonebox Speech Synthesis software. Target languages included Serbo-Croatian, Korean, Creole French, and Arabic. Average language translation performance was one second per sentence.

5. SMART MODULE ARCHITECTURE

The Smart Module system has two distinct kinds of processes: the Server-Application Group and the System Controller. A Server-Application Group consists of a UNIX background process which communicates with an application, such as PANLITE, via Inter-Process Communication within a module. The server process also communicates with the System Controller over the TCP/IP Network. The System Controller keeps track of what servers are present on which modules, and coordinates the flow of information between the servers. It is possible to interface any number of applications with one server process. This architecture makes it easy to add new modules to the system. Figure 8 illustrates how data flows between the System Controller and a Service-Application Group. The System Controller operates on a Newton MessagePad 2000 to give the user a chance to correct misrecognized words. The Newton is primarily used as a display device.

The key factors that determine how many processes can be run on a module are memory, storage space, and available CPU cycles. To minimize latency, the entirety of an application's working dataset should be memory resident.



e



Fig. 7. Speech Recognizer (SR) and Language Translator (LT) Smart Module

The intermodule communications infrastructure

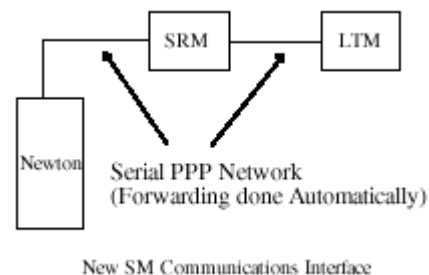


Fig. 9. Serial PPP Communication

is a TCP/IP based network running over serial PPP links, as detailed in Figure 9 [13]. TCP/IP can be built directly into the Linux kernel, eliminating the need to deal with the network in the Server software. It also supports packet forwarding directly in the kernel. Finally, it can be utilized over a variety of communications media, supporting several wired as well as wireless connections. It is even possible for the system to communicate with

any TCP/IP based intranet or the Internet, if a module is configured as a gateway with a connection to an outside network.

The position of each module in the physical network does not matter; the System Controller simply sends out all communications for all modules over the same link, creating a virtual network as shown in Figure 10. The modules themselves handle routing. New modules added to the system can have the capability to modify each others' routing tables automatically. Currently, because all of the modules used are physically connected with each other, the Linux PPP server automatically configures the routing tables of the modules. But if more modules are added to the system, a dynamic routing protocol must be used to modify the tables of a module that may not be physically connected to the module that is added.

The secondary storage drives are of Type II and Type III PCMCIA form factor, but these drives also support an IDE interface. The PCMCIA socket that is on the Smart Modules is wired directly into the IDE bus, and there is no PCMCIA controller in the hardware design. While this precludes the use of anything other than hard disks in the PCMCIA slots, it saves space in the overall design.

Figure 11 depicts the functional prototype of the Speech Translator Smart Module, with one module performing language translation, and another one speech recognition and synthesis. The optimized version of the Speech Translator Smart Module is shown in Figure 12.

Operational Experience

The lessons learned from tests and demonstrations include: the manual intervention process to correct misrecognized words incurs some delay; swapping can diminish the performance of the language translation module; the size of display can be as small as a deck of cards.

The required system resources for speech translator software are shown in Table 1. We achieved a six times speedup over the original desktop PC system implementation of language translation, and five times smaller memory requirements.

6. PERFORMANCE EVALUATION

Figure 13 illustrates the response time for

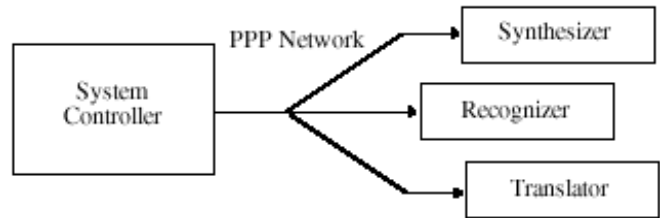


Fig. 10. The Smart Module Virtual Network



Fig. 11. Speech Translator SM Functional Prototype



Fig. 12. Optimized Speech Translator SM

	Laptop / Workstation	Functional Module SR / LT	Optimized Module SR / LT
Memory Size	195 MB	53 MB	41 MB
Disk Space	1GB	350 MB	200 MB

Table 1. Comparison of Required System Resources

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

The performance of the family of Speech Translation modules is summarized in Figure 14. The metric for comparison in Figure 14 is proportional to the

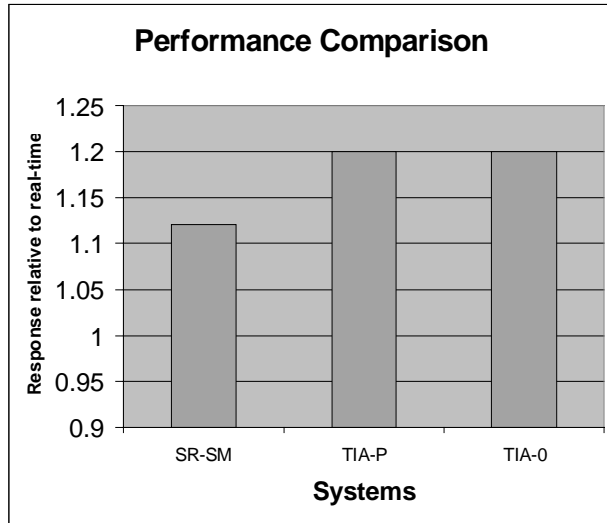


Fig. 13. Response Time Comparison

Name	SpecInt	Volume (in ³)	Weight (lbs)	Power (watts)	R (V*W*P)	Normalized - SpecInt/R	Log of Normalized
TI 6030	175.00	260.00	7.50	36.00	70200.00	1.00	0.00
TIA-P	55.00	88.00	3.00	6.50	1716.00	12.86	1.11
TIA-O	55.00	45.00	2.50	4.50	506.25	43.58	1.64
SR-SM	175.00	45.00	2.13	4.00	382.50	183.53	2.26
OPT-SM	175.00	33.00	1.50	4.00	198.00	354.55	2.55

Table 2. Performance Values Measured and Calculated for Wearable Computers

processing power (SpecInt), representing performance, and inversely proportional to the product of volume, weight, and power consumption (R), representing physical attributes. Figure 14 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.

7. CONCLUSIONS

Four generations of CMU wearable computers have been built for real-time speech translation applications, culminating in the Speech Translator Smart Module. Our results show that there are orders of magnitude improvement in performance as we proceed from one generation of Wearable Computers performing speech recognition to the next one. To our knowledge, Speech Translator Smart Modules are the only wearable computers capable of performing two-way speech translation (involving speech recognition and language translation).

A system-level approach to power / performance optimization improved the metric of (performance / (weight * volume * power)) by over a factor of 300 through the four generations.

8. ACKNOWLEDGMENT

This work was supported by Defense Advanced Research Project Agency Contract # DABT63-95-C-0026 and Institute for Complex Engineered System at Carnegie Mellon University.

9. 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. Siewiorek, A. Smailagic, and J.C. Lee, "An Interdisciplinary Concurrent Design Methodology 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] D. Reilly, "Power Consumption and Performance of a Wearable Computing System," Masters Thesis, Carnegie Mellon University, Electrical and Computer Engineering Department, 1998.

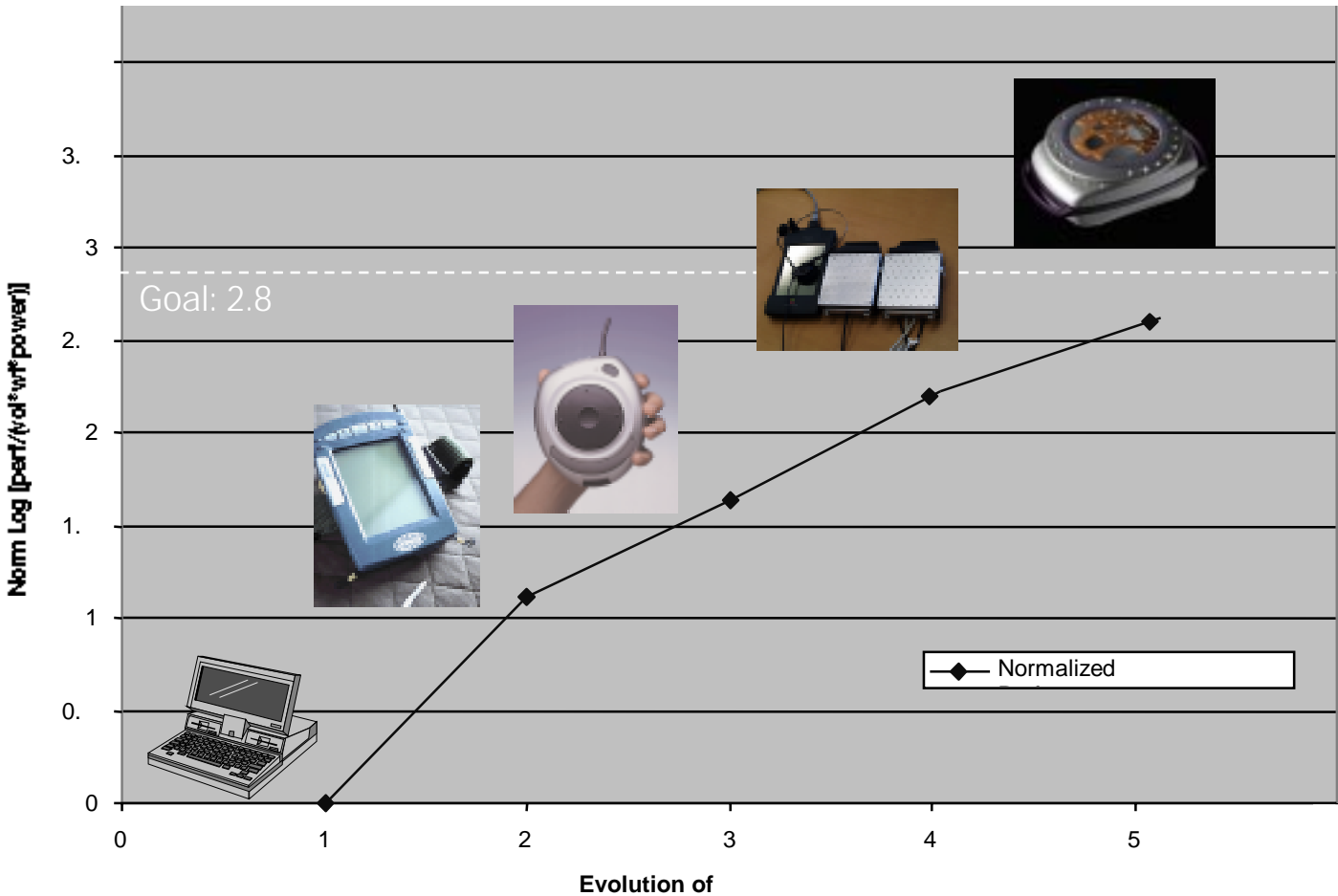


Fig. 14. Composite Performance of Speech Recognition Wearable Computers

- [5] D.P. Siewiorek, A. Smailagic, L. Bass, J. Siegel, R. Martin, B. Bennington, "Adtranz: A Mobile Computing System for Maintenance and Collaboration," Proceedings 2nd IEEE International Conference on Wearable Computers, Pittsburgh, PA, 1998.
- [6] B. Bederson, "Audio Augmented Reality: A Prototype Automated Tour Guide," Proc. of CHI '95, May 1996, pp. 210-211.
- [7] 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
- [8] 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.
- [9] Epson Corporation, Epson CARDIO 486-D4 Data Sheet, 1997.
- [10] M. Ravishankar, "Efficient Algorithms for Speech Recognition," Ph.D Thesis, Carnegie Mellon University, Tech. Report. CMU-CS-96-143, May 1996.
- [11] K.F. Li, H.W. Hon, M.J. Hwang, and R. Reddy, "The Sphinx Speech Recognition System," Proc. IEEE ICASSP, Glasgow, UK, May 1989.
- [12] R. E. Frederking, R. Bown, "The Pangloss-lite machine translation system. Expanding MT Horizons," Proceedings of the Second Conference of the Association for Machine translation in the Americas, 1996, pp. 268-272.
- [13] J. Dorsey, "Smart Module Networking," Personal Communication, 1998.