# The Audio-Graphical Interface to a Personal Integrated Telecommunications System

by
Barry Michael Arons

B.S. Civil Engineering
Massachusetts Institute of Technology
1980

Submitted to the Department of Architecture
in partial fulfillment of the requirements for the degree of
**Master of Science in Visual Studies**
at the
**Massachusetts Institute of Technology**

June 1984



Signature of author . . . . . .
Barry Michael Arons
Department of Architecture
May 11, 1984

Certified by . . . . . .
Professor Andrew Lippman
Associate Professor of Media Technology
Thesis Supervisor

Accepted by . . . . . .
Professor Nicholas Negroponte
Chairman, Departmental Committee for Graduate Students

# DISCLAIMER NOTICE

The accompanying media item for this thesis is available in the MIT Libraries or Institute Archives.

Thank you.

# The Audio-Graphical Interface to a Personal Integrated Telecommunications System

by
Barry Michael Arons

Submitted to the Department of Architecture on May 11, 1984
in partial fulfillment of the requirements for the degree of
Master of Science in Visual Studies.

## Abstract

The telephone is proposed as an environment for exploring conversational computer systems. A personal communications system is developed which supports multi-modal access to multi-media mail. It is a testbed for developing novel methods of interactive information retrieval that are as intuitive and useful as the spoken word.

A personalized telecommunications management system that handles both voice and electronic mail messages through a unified user interface is described. Incoming voice messages are gathered via a conversational answering machine. Known callers are identified with a speech recognition unit so they can receive personal outgoing recordings. The system's owner accesses messages over the telephone by voice using natural language queries, or with the telephone keypad. Electronic mail messages and system status are transmitted by a text-to-speech synthesizer. Local access is provided by a touch sensitive screen and color raster display. Text and digitized voice messages are randomly accessible through graphical ideograms. A Rolodex-style directory permits dialing-by-name and the creation of outgoing recordings for individuals or mailing lists.

Note: A 3/4 inch color U-matic video cassette accompanies this thesis, it is five minutes in length, and has an English narrative.

Thesis Supervisor: Andrew B. Lippman
Title: Associate Professor of Media Technology

The work reported herein was supported by a grant from Atari, Inc. and NTT, the Nippon Telegraph and Telephone Company.

# Table of Contents

In memory of my mother and father.

# Introduction: The Personal Telecommunications System

> *But a major question remains: will these terminals be easy to use? Will these electronic interfaces between humans and computers be compatible with human beings as well as computers? Although human-computer interfaces are becoming "friendlier," the potential for improving their usability is enormous. The ideal interface should be usable immediately by people approaching it for the first time, or by those who use only occasionally. It should allow people to tap the rich resources of electronic information technology with a minimum of effort.* [Klapman 82]

This thesis describes the man-machine interface for a prototype telecommunications system developed at the Architecture Machine Group. The system owner's methods of interaction and access are explored and developed. The system is personalized and integrated in that a personal computer becomes one's total telecommunications manager, handling both incoming and outgoing communications of various types. The computer acts as its owner's telephone directory, mail box, and personal secretary. The machine recognizes its owner and his acquaintances, delivering specialized greetings and messages to each. It is a system which may be used without realizing that you are conversing with a machine; it is not necessary to know anything about computers before you use it.

Two common forms of inter-personal communication, voice and text, are merged. The methods of access and presentation are identical; the differences between types of messages are made transparent. Voice messages, conventionally recorded sequentially or transcribed, are interactively gathered during a dialogue with the computer in which the maximum amount of relevant information is interchanged. Text messages, in the form of electronic mail, are similarly collected, sorted, and distributed. The owner no longer has to obtain his messages from different sources, waste time playing *telephone tag*, or worry

about missing an important call. The owner may call in, hear a message from Mr. X, then create a personalized outgoing recording for Mr. X, making a dialogue possible even though the parties never speak to each other directly.

Access is multi-modal: by voice, Touch-Tones, or touch sensitive screen. Each method provides equal capabilities, they may be used individually or in concert. The primary remote interface is by voice over the existing switched telephone network. The owner makes verbal requests and is similarly answered by voice; speech either previously recorded by a human or generated on-the-fly with a text-to-speech synthesizer. At home or office the owner can interact with a color raster display outfitted with a touch screen. Touch sensitive ideograms in various screen images enable the viewing of messages, creation of outgoing recordings, and dial-by-name capabilities.

Speech, to be an effective bidirectional communication medium, must be intimately tied to the application [Schmandt 82a, Schmandt 82b]. A speech recognizer or synthesizer is not a black box that simply gets connected between a human and a computer instead of a keyboard or computer terminal.

> Successful approaches can be broadly classified as *systemic* or *holistic*. The solution is not to make speech i/o replace a few buttons or indicator lamps, but rather to fully integrate speech into the whole context of communication, i.e. exchange of information, between the operator and computer. The tools are not so much recognition as understanding, with the implication of an intelligent system interacting with an intelligent user. [Schmandt 82b]

While this work is currently embodied in a hypothetical *teleterminal*[1], the underlying principles investigated in this project range far beyond the telephone. It is anticipated that speech communication with computers will become more prevalent in the future, and to this end conversational computers must be

---

[1] A teleterminal is defined to be a piece of equipment that merges the functionality of a traditional *tele*phone with that of a computer *terminal* [Bayer 83].

explored. An area where speech is already a natural mode of communication is the field of telecommunications. People are quite accustomed to speaking into the mouthpiece of a telephone and receiving verbal replies. This existing link is used as a means to investigate machines that speak and listen.

# Chapter One

# Telephone Perspective

> *The conventions of telephone use are deeply established. Although people are often annoyed by the present arrangements, they also tend to be quite conservative and to resent changes in the system.* [Swinehart 83]

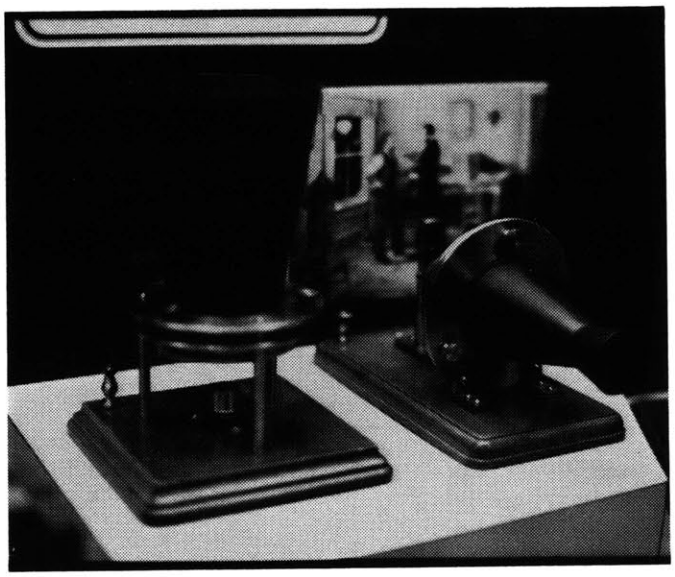
In the most simple terms possible this work can be described as an *intelligent telephone*, and this is the context in which the project was developed. This chapter traces the evolutionary path of the telephone and introduces various technologies upon which the personal integrated telecommunications system is built. A survey of related research and commercially available products is included to characterize the current state-of-the-art in teleterminals and interactive voice messaging systems.

## 1.1 A Brief History of the Telephone

Samuel F. B. Morse completed his first working model of the telegraph in 1832. Morse and Alfred Vail, his financial backer, exchanged the first long distance telegraph message between Washington and Baltimore in 1844, President Lincoln received the first transcontinental telegram seventeen years later. By 1873 the Western Union Company was transmitting more than 90% of the telegrams in the U.S., over a network consisting of more than 150,000 miles of wire. The costs for installing and maintaining the wires, poles, and insulators for the booming telegraphy market were high, and many enterprising individuals were trying to develop ways to make multiple use of the existing telegraph lines.

Elisha Gray, one of the eventual founders of the Western Electric Company; Thomas Alva Edison, the inventor of the incandescent lamp and numerous other

devices; and Alexander Graham Bell were all attracted to the lucrative possibility of giving Western Union the ability to multiply its system capacity without adding more miles of wire. In 1874 Edison invented the quadruplex telegraph which allowed two messages to be sent in both directions on a single telegraph wire. Bell and his assistant, Thomas Watson, were working on a harmonic telegraph scheme which would permit 30 or 40 messages to be sent simultaneously.



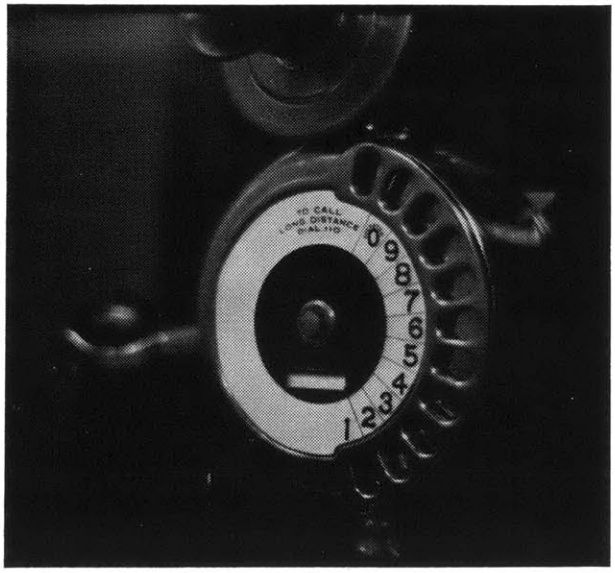
**Bell's first telephone.**

An accidental discovery in June 1875 excited Bell so much that he stopped work on the harmonic telegraph and began working almost exclusively on the transmission of human voice over wires. This work culminated with the invention of the telephone for which Bell was granted a patent in early 1876. By the fall of 1877 the newly formed Bell Company had over 1000 telephones in operation. In the following years many legal battles were fought between Bell and the telegraph giant Western Union. Edison developed and patented many telephone devices; most notable was his invention of the pressure sensitive carbon transmitter which amplified the energy extracted from the sound wave and made the telephone commercially feasible. In late 1879 Western Union admitted the validity of the Bell patents, and agreed to retire from the telephone business.

In the early days of the telephone, the operator on a manual switchboard depended upon tone signals, verbal instructions, and lamps associated with cords for the handling of calls. Subscribers had to verbally transmit the called number to the operator despite technical advances in the handling of calls. Operators also informally served as receivers and transmitters of messages for their customers. In many small towns they simply jotted down notes and periodically tried to deliver them. In all but the smallest exchanges, this quickly became an unbearable task. In self-defense, the operators reverted back to simply making connections. In light of this development and the advent of automated switching, a new method for the storage and distribution of messages had to be created.

**An early manual switchboard.**

The rotary dial was introduced around 1895; this event marked a major technical advance in telephony. The dial, as it rotates back to its stop position, generates a train of pulses that correspond to the number of the selected digit. These pulsed digits are used to directly position switches in the associated central office so that no operator is involved in the connection of a call.




**An early rotary dial telephone.**

By the 1930s telephone answering bureaus sprang up; a specialized operator would answer the phone, and manually transcribe and deliver messages. In the early 1950s an attempt was made to mechanize the storage and playback of one-way voice messages through answer-and-record devices on the customer's premises. Initially these devices were quite expensive and their major use was limited to providing universal announcements which could be accessed by many people. These machines proved to be more economical when a single device was used to store the messages of many people then when they were used for individual customers. In recent years telephone answering machines have become quite common in homes and offices; a dual-cassette system is generally employed which permits a single announcement to be played to all calling parties and the sequential recording of a large number of incoming messages.

The higher switching speeds of electronic switching systems brought about a need for a more rapid and accurate means of dialing. Touch-Tone dialing, introduced by the Bell System, uses tones in the voice frequency range to transmit dialing information to the central office. These dual-tone multifrequency

(DTMF) signals can be transmitted worldwide, extending new services to the general public.

The telephone monopoly was controlled by the Bell System until 1968 when the FCC allowed the Carterfone Company to interconnect its mobile radio-telephones to the Bell System. The Carterfone decision, subsequently upheld by the Supreme Court, broke AT&T's equipment monopoly, allowing other vendors to enter the telephone market. In 1982, an agreement was reached between AT&T and the U.S. Justice Department, that led to the deregulation of the phone industry and the divestiture of the Bell system in 1984.

The goal of these moves was to stimulate competition in the communications industry so that new and improved products and services could become available for consumers more quickly than when the industry was under the near-total control of a regulated monopoly. The long term effects of the AT&T breakup are still uncertain, but many advanced telecommunications products such as those described in this document would not have been possible without the Carterfone decision.

## 1.2 Speech Recognition

> *Although I have not been able to track it down, I have heard there is a paper about one of the hardest problems in Artificial Intelligence and Signal Processing entitled "How to wreck a nice beach." [Hint: say it aloud.]*
>
> Computer Humor, Communications of the ACM, April 1984

People often take their ability to understand speech for granted. Speech recognition by machines is a very complicated task which may involve signal

processing, pattern matching, and syntactic and semantic constraints [Klatt 77, Reddy 76]. Automatic speech recognition is currently being used in the office for information entry and retrieval and in industry where jobs require hands-busy/eyes-busy activities. Speech recognition may both reduce expensive manpower requirements while substantially increasing the functionality of computers in such environments.

There are several general classifications of automatic speech recognition systems including: speaker-dependent vs. speaker-independent, isolated speech vs. connected speech, and limited vs. virtual vocabulary [Pathe 83]. Most recognizers must be trained to a particular word set, for speaker-independent recognizers speech samples from hundreds of people must be gathered and processed. The computation necessary for connected speech recognition is significantly more complex than for isolated word recognition. The task is made difficult by the dropping of inter-word pauses and the coarticulation of adjacent words in continuous speech. For example, when spoken quickly in a sentence, the words "did you" are usually pronounced as "didja".

Although specific speech recognition systems differ in the details of implementation, all existing systems go through three essential steps in performing recognition: feature extraction, similarity determination, and response decision.

Feature extraction consists of processing the incoming acoustic signal to determine the beginning and ending of an utterance, and to yield a set of features. Less expensive systems base their features on a simple measure of energy and zero crossing rates. The most common technique is to use the output of a bandpass filter bank. A third method bases the feature set on a linear predictive coding (LPC) analysis of short overlapping segments of the digitized signal. Commercial systems often use proprietary techniques instead of, or in combination with these methods.

The features extracted from the speech signal are then compared with previously stored feature sets or *templates* to produce a similarity metric. The features must be time aligned with the templates using a technique such as dynamic time warping or linear time alignment. Dynamic time warping involves a nonlinear compression or expansion of the input signal to maximize its similarity measure against the template. The exact method of similarity determination depends upon the feature set used, but may simply be the number of bits in the input signal that match corresponding bits in the template. Speaker-independent recognition systems usually contain multiple templates for each vocabulary item, characterizing variation in a word across speakers and speaking conditions. In sophisticated speaker-independent systems, the templates may be adapted automatically to provide better recognition performance as a speaker continues to use the system.

The system must make a decision as to which template most closely represents the spoken input. Most recognition systems allow the setting of an absolute threshold; if the highest score doesn't exceed this limit, no recognition decision will be made. Some systems also provide the option of setting a relative threshold based on the ratio of the highest and second highest scores. If both thresholds are exceeded, the utterance is recognized.

Higher level information can also be used to syntactically constrain or partition the vocabulary. It may, for example, be known that the utterance must be a one or zero, as in the second digit position of an area code. To speed the similarity determination process, the pattern matching may only take place between these two templates. Similarly, the response decision may also be limited by this knowledge.

> In fact a straightforward digit recognition task with only 10 words total (0 through 9), but with each word equally likely, is far more difficult then most tasks that have actually been tested using laboratory ASR systems with total vocabularies of up to 1000 words...Performance error rates often increase by a factor of 3 to 10 when a laboratory system, tested experimentally on 100, 200, or more voices, moves into a genuine, commercial, field operation. [Baker 81]

Speech recognition over the phone is aggravated by several factors: 1) human speech ranges in frequency from approximately 100-8000 Hz, while the telephone is band is limited to 300-3000 Hz, therefore some information carrying parts of speech spectrum are lost, 2) frequency characteristics vary depending upon the telephone line and telephone set being used, 3) a carbon granule telephone transmitter produces nonlinear distortion (more distortion with higher input level), 4) input sound levels may vary over a broad range, and 5) many types of line and room noise, with different frequency characteristics, overlap the input speech signal.

### 1.2.1 Bell Laboratories

Bell Laboratories produced an isolated word recognition system based on equally spaced frames of LPC coefficients [Itakura 75]. This recognizer was used over dial-up phone lines for several experiments dealing with a simple airline information and reservation system. An 84 word vocabulary was initially used in a question and answer dialogue between the caller and the computer. This type of dialogue often resulted in a long series of questions in order to completely specify a request.

The vocabulary was expanded to 127 words, including many auxiliary and function type words, so that reasonably natural English sentences could be formed. The effects of syntactic constraints on this finite-state grammar (144 states, 450 transitions, and $6 \times 10^9$ possible sentences) were investigated through a computer simulation [Levinson 78a] and an experiment [Levinson 78b]. The

simulation consisted of generating 1,000 sentences with an average length of 10.3 words. An assumed word error rate of 10% was reduced to 0.2% by the syntactic constraints of the task language. In the experiment, speakers prompted by a computer terminal spoke sentences containing an average of 8.7 isolated words. Recognition was carried out off-line; the best five word candidates from the acoustic recognizer were input to the syntax analyzer. As indicated in the computer simulation, the syntactic analysis had a powerful correcting influence on acoustic word errors. The word error rate was reduced from 11.7% to only 0.4% and the overall sentence error rate was 3.9%.

A third level, in the form of a semantic processor [Levinson 80], was added to the existing recognizer/semantic analysis system. An audio response system was controlled by this processor so that a natural language conversation could be held with the machine. With one set of test sentences, 6 of 21 sentences were corrected by the semantic processor without intervention by the user. The remaining 15 sentences caused the system to responded with "What did you say?", the error was corrected by the user on his next input sentence.

> In no case was communication seriously disrupted. This phenomenon has a profound effect on the user of the system. His attention is drawn away from speech recognition accuracy and sharply focused on the exchange of information between himself and the machine. This points very strongly to the conclusion that progress in speech recognition can be made by studying it in the context of communication rather than in a vacuum or as part of a one-way channel. [Levinson 80]

An automatic directory assistance system that permits users to spell out the last name and initials of the desired party with Touch-Tones has been available for the 18,000 entry Bell Laboratories telephone directory since 1976 [Rabiner 76]. Multiple matches occur approximately 25% of the time due to identically spelled names (only the first six letters are used) and ambiguities arising from the multiple letters assigned to each button on the keypad. These conflicts are

resolved by using a voice-response system which asks the caller to supply additional information. Related work has shown that carefully worded prompts can elicit isolated speech even from first time users [Holmgren 83].

The previously described isolated word recognizer was used in conjunction with this database to provide a speaker-dependent [Rosenberg 79], and subsequently a speaker-independent [Rosenberg 80], directory system. The spoken alphabet is a notoriously poor vocabulary for a word recognizer; large groups of utterances within the vocabulary are easily confused because they have minimal acoustic differences (e.g. the A-J-K and B-D... families). An individual letter error rate of approximately 20% was reduced to 4% for the entire name in both the dependent and independent training cases by the constraints imposed by the spellings of the names.

### 1.2.2 Nippon Electric Company

Nippon Electric Company (NEC) has produced several speech recognition products ranging upward in complexity from a single board for a personal computer to the NEC SR-1000. The SR-1000 series speaker-independent isolated word recognizers are designed to be used over the phone line as part of an integrated voice recognition and voice response (V&V) system [NEC 82]. The SR-1201 can be used to recognize ten numerals and six functional words (e.g. yes, no, cancel) when a push-button telephone is not available[2].

Each V&V system can be connected to as many as 128 telephones. To use the recognition units efficiently, they are time-shared by dynamically connecting them to a telephone line for each input word according to instruction of the system controller. This system has been used in a banking application allowing customers to obtain transaction information from any telephone.

---

[2]Over 90% of the phones in Japan still use rotary dials.

### 1.2.3 Nippon Telegraph and Telephone

Nippon Telegraph and Telephone (NTT) has developed a speaker-independent recognition unit specifically for telephone line use [Ishii 82]. The design of the system was based on an analysis of a large amount of speech data gathered from many speakers over various telephones and lines. The unit has 32 input channels and a vocabulary of 16 words. The speech detection threshold is adaptively determined by sampling the background noise on the line. Detailed recognition results and second and third guesses with confidence values, can be transmitted to allow the host to perform higher level processing. At any stage of a recognition sequence the accepted words can be limited to a subset of the given vocabulary.

### 1.2.4 Verbex

The Verbex Model 1800 speaker-independent isolated word recognizer is designed to be used over the phone. A 32-bit array processor supports eight channels of simultaneous recognition in real-time. Some machines in commercial use handle over 4000 calls per day. The model 1800 dynamically adapts to noise and to each speaker's voice during the course of an interaction (transparent training).

### 1.2.5 Votan

Votan produces a series of modular voice products including recognizers tailored for the telephone bandwidth, and voice response and speaker verification units. A new system based around a board for the IBM Personal Computer features a speaker-dependent word recognizer with high noise immunity and a voice response and storage option. An inexpensive speaker-independent recognizer tailored for phone line use is expected soon.

## 1.3 Voice Dialing

> *To converse with a computer in a natural manner it must, as a minimum, be able to speak, listen and understand conversational English.* [Bergland 82a]

A seemingly natural application of speech recognition is in the area of automated dialing. Several experimental and production automatic dialing systems based on speech recognition have been built on the premise that it is generally easier to remember a person's name than his telephone number.

### 1.3.1 Bell Laboratories Repertory Dialer

Bell Labs developed a speaker-dependent dialing system with a vocabulary of the ten digits, seven commands (e.g. hangup, error) and a list of names [Rabiner 80]. Once trained, it could dial the telephone number corresponding to any name in the repertory, or dial a 4-digit extension when spoken as a string of isolated digits.

All communication between the user and the system is by voice; there is no visual display needed to train or operate the system. A digitized voice response system is used to provide feedback and training cues. If the speech analyzer detects any recording problems (e.g. level too low), a request is made to repeat the word. The recognizer only responds to isolated words, so the user may hold a conversation while the dialer is operating. The system will not be triggered unless one of the command words, spoken in isolation, has two distance scores within prescribed limits. The vocabulary is partitioned such that at any time the recognizer must choose among only a subset of the candidates. Due to the tight recognition constraints, no recognition errors and only a small number of requests for repeats occurred in over 4500 trials for six speakers.

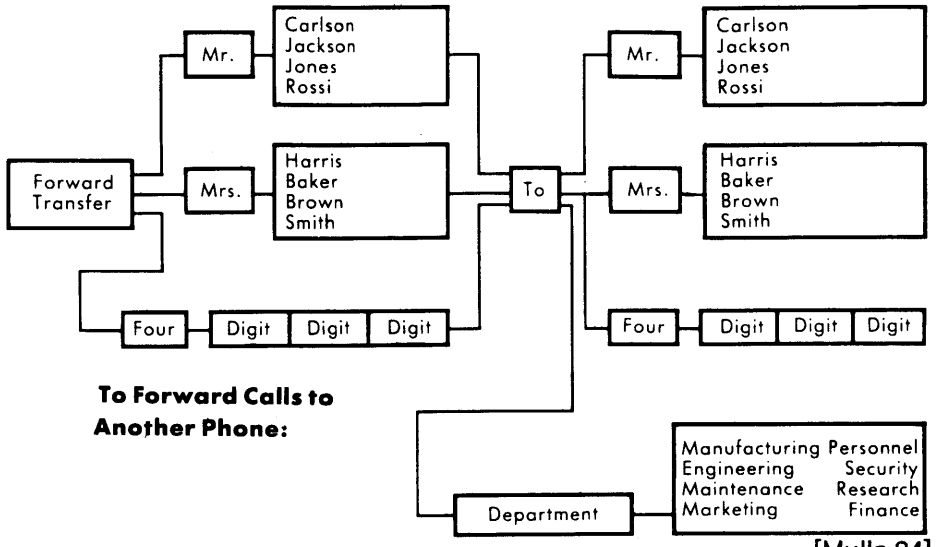
### 1.3.2 Ericsson Voice Controlled Intercom

A speaker-independent voice dialed intercom system developed at LM Ericsson Telematerial AB [Lundin 83] permitted access to a list of 30 people without having to individually train each user. Pushing the 9 button on the keypad lit an LED indicating that the recognition unit was ready and that the user had ten seconds to speak an utterance. A tone indicated recognition processing after the name was pronounced, an audio response system confirmed the recognized name. The facility was generally found useful, but the recognition accuracy and response time were too low and the training procedure too involved for commercial applicability.

A new speaker-dependent device was developed which eliminated the complex speaker-independent training procedure, the effects of varying acoustical environments, and permitted each user to have a personalized directory. This work led to the manufacture of a commercial product containing a Motorola MC68000 microprocessor and a NEC 7720 digital signal processing chip. Each person can create his own directory by entering the desired extension number and training the name five times.

### 1.3.3 ITT Experimental Voice Dialing PABX

A private automatic branch exchange (PABX) is often equipped with features such as call forwarding and call transfer. These utilities, however, are often not used because the sequences of keystrokes are difficult to memorize. ITT developed a system in which a PABX is tied to a speaker-independent recognizer and speech synthesizer [Mulla 84]. In this configuration the speech processors are a shared resource serving many users, and hence can justify higher cost and performance equipment.




[Mulla 84]

**Sample syntax diagram from the ITT PABX.**

The system has a well defined syntax and rather limited vocabulary permitting the use of syntactic constraints to improve recognition performance. The aggregate score of an utterance is checked against a threshold which is based on how disastrous the consequences of taking an incorrect action would be.

> The setting of thresholds has a very strong influence on the user's perception of the system's performance. With the thresholds set too low, the number of incorrect actions is large, whereas with the thresholds set too conservatively, the system gives the impression of being hard of hearing. [Mulla 84]

### 1.3.4 Audec Command Dialer

Audec Corporation currently markets a consumer product billed as "the world's first telephone you dial with your voice." In addition to automatic re-dial and hands free talking, the phone allows sixteen numbers to be dialed by speaking a single word command such as "school" or "office".

**Audec Command Dialer.**

## 1.4 Speech Synthesizers

Computer controlled voice synthesizers are becoming more prevalent in our society. Stored voice response systems are regularly used by the phone company for information about misdialed or disconnected numbers, and in automobiles for reminders about fastening seatbelts or getting an oil change. These systems contain a specialized and very limited vocabulary consisting of short segments of prerecorded speech. Different utterances can be produced by playing the words in a different order. Sentences often sound unnatural because the individual words are recorded with little inflection so they can be pieced together in any order. These systems are relatively inexpensive but with a high bit rate can produce good quality speech.

Speech can be stored in analog or digital form on a variety of magnetic or optical media for random access. The large data rate for PCM (see section 2.1.1) can be reduced through a variety of data compression techniques which generally take advantage of the relatively slowly varying nature of speech [Rabiner 78, Flanagan 79]. As the data rate is reduced, the quality and

intelligibility of the synthesized speech tend to decrease. Synthesis-by-rule is necessary for applications where speech must be generated from free form text, as in the reading of electronic mail messages.

Waveform coders use alternative descriptions of the speech waveform, such as the difference between samples (delta modulation), pause removal, or the run-length encoding of invariant portions of the speech signal.

Parametric coding provides greater compression through a reversible description of the signal, often based on the vocal tract. Parameters are extracted and later used as input to the inverse operation, to reproduce the spectrum of the original, rather than matching it sample by sample. Linear Predictive Coding (LPC) produces coefficients which describe the sound as a linear function of previous samples.

Text-to-speech synthesizers are advantageous because they are not restricted to a small vocabulary and do not require vast storage or high transmission bandwidth. There are two popular approaches to a text-to-speech synthesizer model; one is to generate LPC coefficients, the other is to use the resonant formants of the vocal tract. To a first approximation, these two techniques produce similar results, at a data rate about one-tenth that of PCM encoded speech [Flanagan 81].

Standard ASCII text is first converted into a sequence of symbols representing the distinctive phonemes for the utterance. This process can be performed by a letter-to-sound conversion algorithm or by looking it up in a dictionary [Allen 76, Allen 81], most text-to-speech synthesizers use a combination of these techniques. Advanced approaches attempt to model clause and sentence level prosody, including pauses, pitch contours, and stress.

### 1.4.1 Federal Screw Works

The Vocal Interface Division, of Federal Screw Works produced several early speech synthesis products, notably the Votrax and Type'n'Talk. The Votrax did not contain any text-to-speech rules, and is representative of a synthesizer where the text to phoneme conversion is placed in the hands of the user. Sixty-three phonemes with four choices of inflection are available to the user, for example the word "America" would be transmitted to the Votrax as:

| phoneme | intonation |
|---|---|
| UH1 | IN2 |
| M | IN2 |
| EH1 | IN3 |
| R | IN2 |
| I2 | IN2 |
| K | IN1 |
| UH1 | IN1 |

### 1.4.2 Digital Equipment Corporation

DEC has recently announced DECtalk, a new text-to-speech synthesis product that features multiple vocal tract models which allow different voices to be synthesized (e.g. men, women, and children). A user definable dictionary (150 words) and a large (6000 word) built-in exception dictionary in conjunction with a set of letter-to-sound rules permits DECtalk to produce highly intelligible speech [Bruckert 84].

MCI Communications, an alternative long-distance telephone carrier, has recently announced [Zientara 84] that they will be using the DECtalk system in conjunction with MCI Mail, their electronic mail service. Subscribers call a toll free number and enter an identification code to have their text messages read to them.

### 1.4.3 Speech Plus

Several text-to-speech synthesizers are available from Speech Plus including the CallText 5000, a board level product which fits in the communications bus of an IBM Personal Computer. A dictionary is augmented by a set of contextual rules which disambiguate abbreviations such as Memorial Dr. versus Dr. Backer. The unit for the IBM PC has a telephone interface, with a DTMF generator and decoder, which permits the synthesizer to be directly connected to a telephone line.

## 1.5 Voice Storage and Forwarding

> *The 1A VSS brings together a unified communication system incorporating the required transmission paths and storage media to permit voice communication between individuals who do not coexist in either space or time.* [Bergland 82b]

Electronic messaging systems, such as voice store and forward (VSF) have many advantages over more traditional forms of communication. They provide the ability to get information to and from a person without having to locate him, worry about time zone differences, or be inconvenienced by interruptions. Messaging systems can also provide a detailed record of all correspondence, allow routing of messages to many parties, and prevent the obligatory discussion of meteorological conditions.

Studies show that "white collar" professionals spend about 25% of their time in non-interactive communication (i.e. reading and writing), and over 40% of their time in interactive communications such as meetings, face-to-face conversations, and telephone conversations [Klemmer 71]. Much of this information exchange is basically asynchronous; the executive does not require (or sometimes even desire) interaction, he merely seeks to obtain or distribute information. Non-

interactive communications provide a permanent record and have traditional and usually slow distribution paths (e.g. U.S. Mail, libraries). Interactive communications not only have the advantage of speed, but allow voice intonation to carry additional 'meaning. Only 75% of telephone calls are successfully connected to the desired party due to a busy line, no answer, etc. This results in an average of three calls being placed before the game of telephone tag is completed.

### 1.5.1 Bell Custom Calling Services

Bell Laboratories began research in the 1940s on a centralized method of storing voice communications. Research into the specific component technologies, architectures, and service definitions continued into the early 1970s when the required technologies had matured sufficiently to enable a cost effective realization of the Voice Storage System (VSS) [Cornell 82, Gates 82]. The development of a VSS was begun in 1975 with the specific design of the 1A VSS reaching completion in 1976. The first 1A VSS was shipped to the Bell Telephone Company of Pennsylvania in late 1978, with the expectation that the Custom Calling Services II (CCS II) features would be placed in service in 1980.

Custom calling services (Call Waiting, Call Forwarding, etc.) became available with the first electronic switching system (no. 1 ESS) installed in 1965. The VSS provided several new messaging features to the telephone customer [Worral 82]. *Call Answering* offers the general capability to answer a call, deliver a customer recorded greeting, and then record a message from the caller. *Advance Calling* allows a customer to record a message and have it sent to a designated number at any designated time. *Custom Announcement Service* permits a recorded announcement to be delivered to anyone who calls the customer's telephone number. All of these services can be controlled remotely by the customer via Touch-Tones or a specially designated telephone number.

Call Answering customers can screen incoming calls as they are being recorded by dialing a monitor access code. If desired, the customer can flash the switchhook to be connected to the calling party. Recordings are terminated when they attain a maximum message length or after three seconds of silence. The voice-presence detector permits the recording of continuous signals at high levels, thus allowing for the storage of encoded data that was generated by frequency shift keying. Once recording has stopped, the entire message is then duplicated for enhanced reliability.

| | Code† |
|---|---|
| Call Answering | |
| Activation | 1151 |
| Message retrieval | 1152 |
| Deactivation | 1153 |
| Monitor | 1154 |
| Advance Calling | |
| Message recording | 1141 |
| Status check | 1145 |
| Custom Announcement | |
| Activation | 1158 |
| Deactivation | 1159 |
| Remote Access | Seven-digit telephone number |
| Privacy Code Change | 1161 |

† Customers with DTMF signalling telephones may use * instead of the digits 11 (i.e., *51 instead of 1151).

[Worrol 82]

**Digit scheme for using CCS II features.**

A Call Answering customer is notified of pending messages by an interrupted dial tone when the phone is taken off-hook and a short ringing signal when the phone is hung up. To hear messages the customer dials a retrieval code, the CA service responds with the prompting statement; *"You have had M calls since you last played back your messages and you have N messages waiting."* After announcing the day and time of arrival the messages are played to the customer in the order that they were received. The customer has the ability to save, repeat, skip, or pause during the playback of a message.

Initial experiments indicated that the human factors aspects of the design

were generally good, but the reliability of service in terms of lost calls, integrity of the long-term message database and overall throughput of the system was below expectations. Software was modified and re-tested; three additional VSS systems were installed in New York, Dallas, and Chicago. In 1980 full-time "friendly user" service was available to selected telephone company employees at the test sites.

A tariff for the Philadelphia offering was filed with the Pennsylvania Public Utilities Commission [Nussbaum 82], but was not approved because of a pending antitrust suit filed by the Associated Telephone Answering Exchanges, Inc. The suit was rejected, but the FCC ordered that enhanced services encompass the area of voice storage, which could therefore not be offered as part of the regulated telephone network. The Bell system filed a petition for waiver of CCS II but this was rejected in late 1981, the offering was withdrawn and the VSS project terminated.

### 1.5.2 VMX Voice Message Exchange

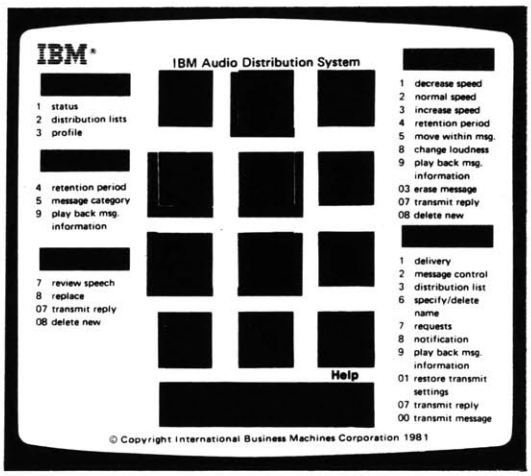
ECS Telecommunications Inc. (later changed to VMX Inc.), founded by Gordon Matthews in 1978 delivered the first commercial voice store and forward system in 1980. The VMX system uses a modified version of delta modulation and a parallel processing scheme; a fully configured system uses over 100 microprocessors to handle 3000 voice mailboxes. VMX holds a patent on 51 individual design features of its VSF system, and has settled a court case against Commterm Inc. by granting a royalty-bearing license on its patent (VMX currently has two outstanding suits for alleged infringements on the patent).

### 1.5.3 IBM Audio Distribution System

The Speech Filing System (SFS) was developed at the IBM Research Center during 1973-1975 [Gould 82, Gould 83]. In the subsequent six years approximately 750 IBM executives used the SFS in their daily work; the user

interface was significantly changed and improved during this period. The IBM Audio Distribution System (ADS), a direct outgrowth of SFS, was released as a product in late 1981. The SFS emphasizes spoken messages, but it has been used to compose and distribute handwritten and typed messages as well.

The ADS has a very extensive Touch-Tone based command set which permits audio documents to be created, edited, and forwarded [IBM 82]. It can be customized by the user to create distribution lists, allow others to hear limited parts of one's messages, or change the amount of prompting from the system. Distribution lists and individual messages are specified by keypressing the alphabetic representation of the name. Message playback speed is increased by automatically deleting pauses between words. An audio-response help facility is available which is conditioned by the context of the user's current location in the hierarchical command tree.




**ADS keypad overlay showing top level functions.**

ADS can automatically place outgoing calls. When the phone is answered ADS prompts; *"Hi, IBM Audio Distribution System calling John Smith. Please keypress your password."* A person can send messages to himself which not only serve as personal reminders, but as wakeup calls as well.

### 1.5.4 PABX Based Voice Storage Systems

A alternative to a stand alone voice mail system is to use a service offered by an existing telecommunication firm, providing the advantages of voice mail without any hardware purchases. Unfortunately since the voice storage system is not tied to a local PABX the caller will most likely have to make two calls, one to the desired party and one to the service bureau. An advantage of having a voice mail system that is closely tied to a PABX is that the called party can be notified, via a lamp or other signal, when a message is waiting or can even identify the calling party.

## 1.6 Integrated Telecommunication Workstations

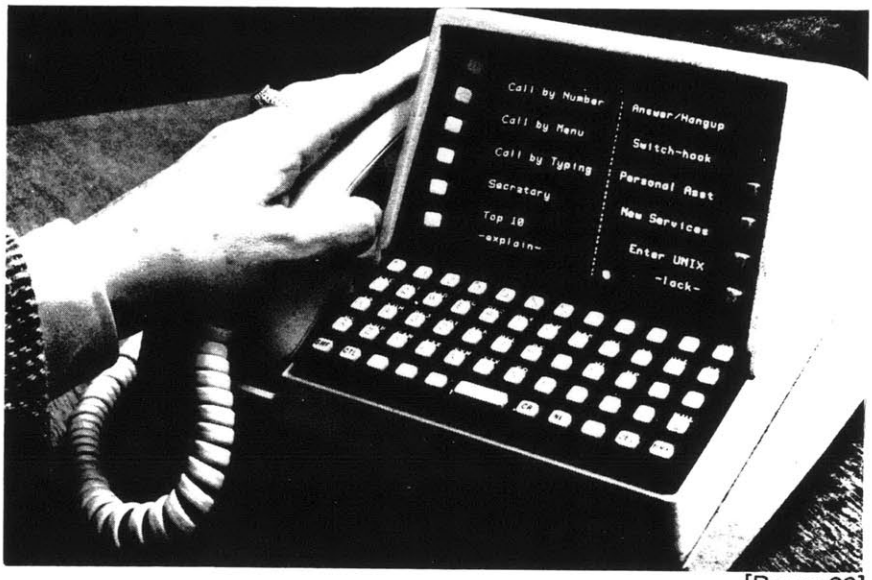
> *In the future it will no doubt be possible to send mixed voice and data over a digital network; however to experiment today we must use the existing analog network.* [Hagelbarger 83]

Telecommunication workstations which are integrated with personal computers and text mail systems have been appearing in the marketplace with increasing frequency. Ranging in size from small portable terminals to desktop workstations closely tied to a digital PABX, these teleterminals are slowly replacing conventional phones and becoming personalized telecommunications nodes.

### 1.6.1 Bell Experimental Teleterminals

Bell Laboratories has developed a series of experimental teleterminals that merge the functions of a conventional telephone with that of a computer terminal [Bayer 83, Hagelbarger 83]. The teleterminal consists of 1) a traditional telephone facility, 2) a microprocessor for internal intelligence, 3) a data communications facility, 4) a general purpose display, and 5) dynamic labeling of

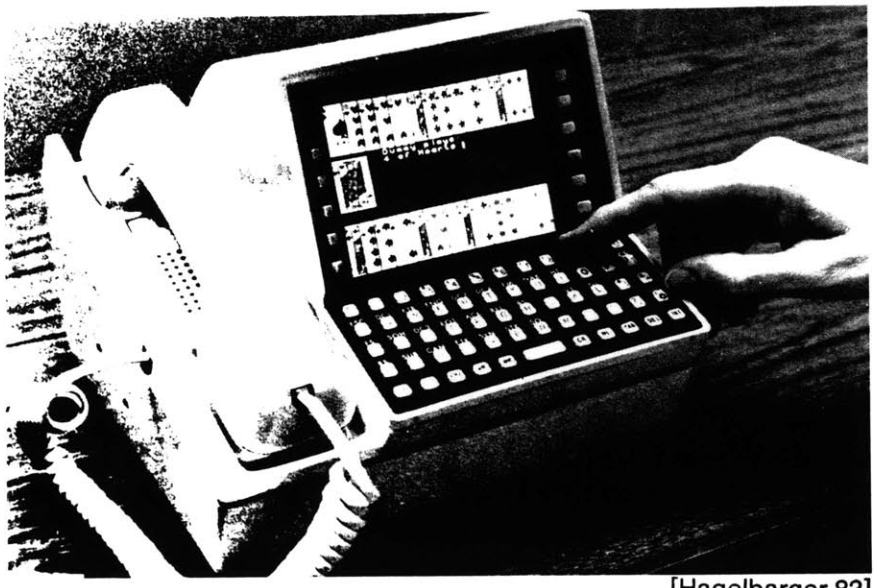
buttons (soft keys). The first terminal contained a two-finger keyboard and a small (16 rows by 32 column) video display with a row of six soft keys on either side. A unique feature of this terminal is that the functions of the keys are definable by the user; there is a tree structured profile file that can easily be changed and updated to suit the user's preferences.




[Bayer 83]

**The first Bell teleterminal.**

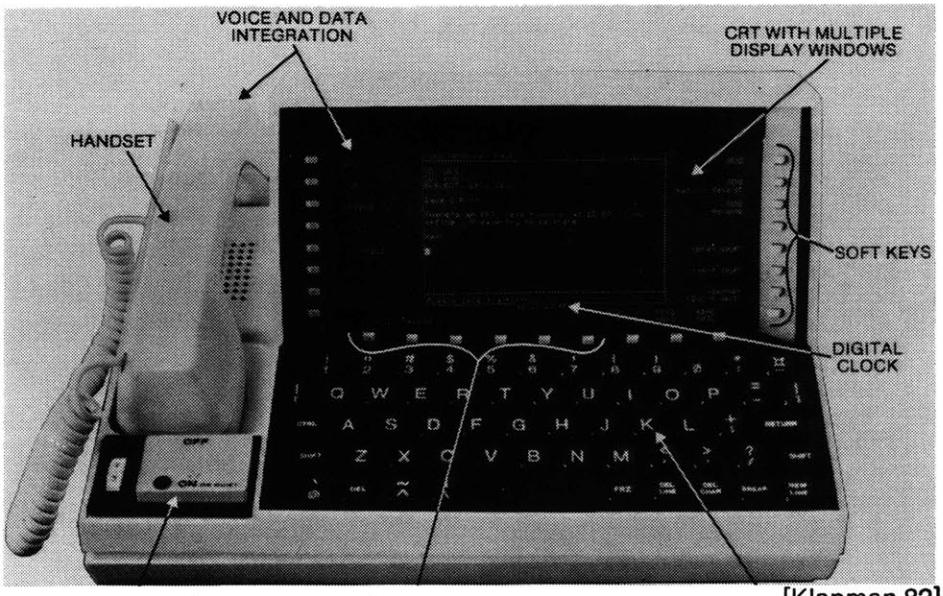
The operation of the teleterminal is perhaps best described by an example (as in [Bayer 83]). Suppose we wish to call Ellen, the department secretary. Customization of the tree-file places Ellen's name on the top level screen, as she is frequently called. Touching the ELLEN button causes her telephone number to be dialed, her mail address to be temporarily stored, and the *Personal Assistant* screen to be displayed. This menu allows access to personal appointment calendars, an electronic mail facility or UNIX on a local host, specialized directories, etc. If Ellen doesn't answer her phone, a message of the form; "I tried to call you. Please call me back at x5156" can be sent with a single button. When Ellen returns and reads her mail she can respond by touching a RETURN CALL button which will automatically dial the extension.



[Hagelbarger 83]

**A color GETSET.**

Several styles of these terminals, known as GETSETS (General-Purpose Electronic Telephone Sets), which transmit voice and data over separate lines have been developed [Bergland 82a]. More recent prototypes included models with higher resolution and color raster displays. The most recent prototype has a 24 line by 80 column display, almost full-sized keyboard, a speakerphone, telephone interface that imitates a six-button keyset, and a microprocessor running a UNIX-like operating system. There are eight soft keys at each side of the display and nine along the bottom. This teleterminal is part of an experimental communications service called EPIC (Executive Planning, Information and Communications) which is currently used by many Bell Labs executives [Klapman 82].




[Klapman 82]

**The EPIC GETSET teleterminal.**

### 1.6.2 Zaisan Voice/Data Workstation

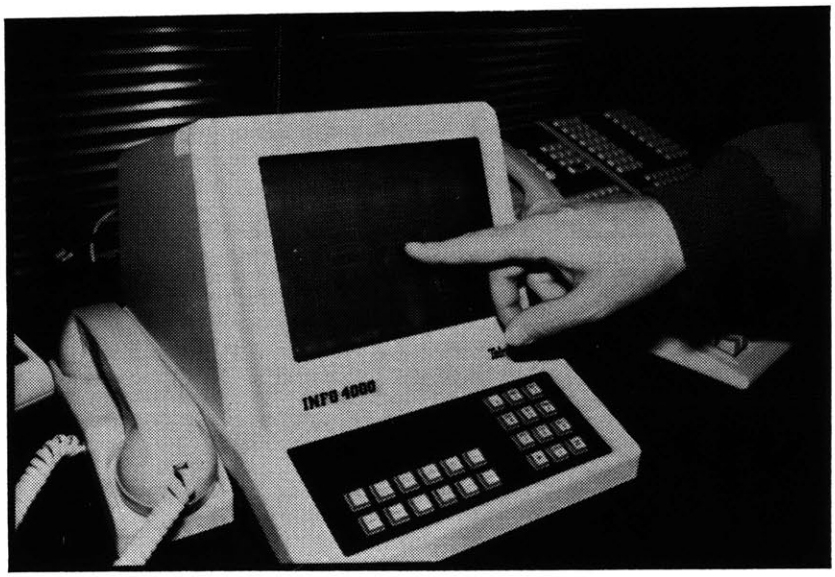
Zaisan's ES.1 teleterminal combines local computing (32K ROM + 32K CMOS RAM with battery backup) with standard telephone features and a built in modem. There are eight soft keys that permit dialing and menu selection from the monochrome display, 13 user programmable telephony keys, and a detachable keyboard. Software features include a directory that can be annotated during a call, calendar/alarm functions, and an electronic mail facility.

### 1.6.3 Xerox Etherphone

Work at the Xerox has taken a somewhat different approach in integrating the telephone into the office environment. A specially designed processor called the Etherphone connects to a telephone instrument and transmits digitized voice, signalling and supervisory information in discrete packets over an Ethernet local area network [Swinehart 83]. The Etherphone processor provides the standard functions of a telephone, but can provide many other services when combined with the power of a nearby workstation, a voice file server, or other shared resources such as databases.

### 1.6.4 Telrad Touchscreen Terminal

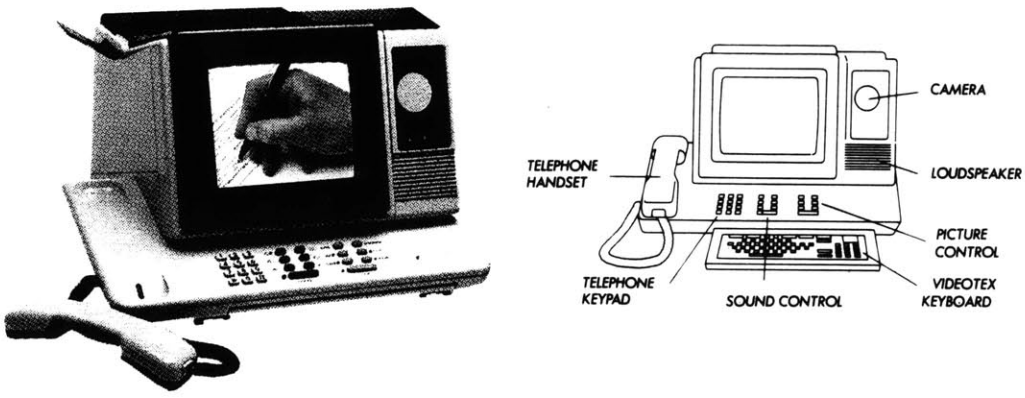
The Telrad Info 4000, an "Executive Voice/Data Touchscreen Terminal", allows touchscreen access to telephone directories and other on-line information. Entries or updates to a personal directory can be made through a typewriter style keyboard or by directly touching the screen. The terminal permits simultaneous voice and digital data transmission over the same line.




**Telrad 4000 with touch sensitive screen.**

### 1.6.5 French Telecommunications Videophone

The French PTT is installing an optical large fiber network in Bairritz France. A high bandwidth *videophone* is used to combine voice and image, as the Bell Picturephone attempted to do in the 1960's [Cagle 71]. The videophone can also be used as an enhanced videotex terminal which can access photographic still images, movies, and sound sequences stored on videodiscs at a central location.




**Videophone used for visual telephony and videotex.**

# Chapter Two

# The Phone Slave

*"I've frequently seen parents who were slaves to the phone and all the calls were for their teen-age sons or daughters..."*

Karen De Witt, The Washington Post, February 28 1977

This section presents an introduction to the hardware and software environment that were used in developing this project. Two predecessor systems are described upon which the telecommunications management system was built; a voiced electronic mail facility and an interactive answering machine. Collectively all of this telecommunications work has fallen under a project called the *Phone Slave*[3]. This work has branched in two main directions, a subscriber service that allows anyone in the lab to read their electronic mail over the phone, and a personalized integrated telecommunications system. For the sake of clarity these will be referred to as the Phone Slave and the PITS respectively.

## 2.1 Computing Environment

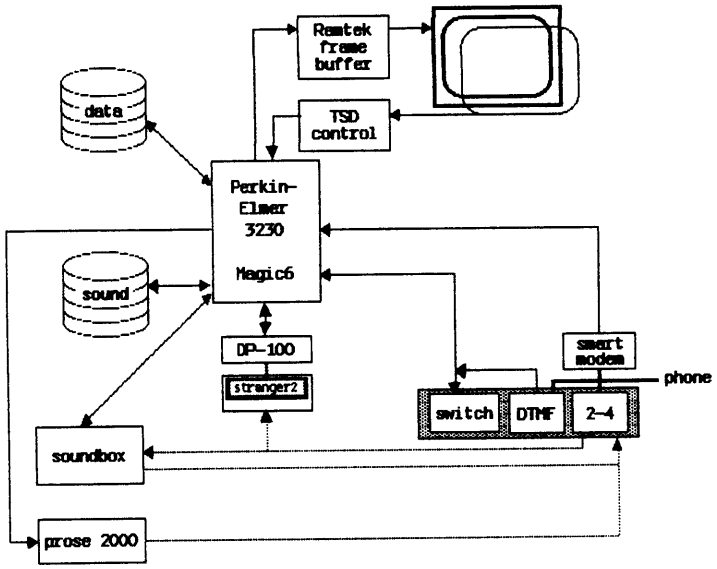
All of the devices described herein are peripherals to a Perkin-Elmer 3230, a 32-bit general purpose minicomputer, running MagicSix, a Multics-like operating system developed at the Architecture Machine Group [Steinberg 74, Parks 79, Kazar 78, Kazar 80, Boyle 81]. PL/1 is the language used for most systems and applications programs. MagicSix has a powerful screen oriented text editor (Emacs), a large body of support software for frame buffer graphics, including grayscale fonts, and device interfaces for speech recognition, synthesis, and digital sound equipment.

---

[3] At the objection of the author.

This is a flexible development system which permits projects such as this to be implemented in relatively short period of time. There is nothing inherent in the software or hardware design which requires a large processor or expensive peripherals. This work is seen as a prototype for a for a system that is integrated around a personal computer with a speech processing board, such as the Texas Instruments Professional Computer.

Several generations of hardware, particularly the telephone interface, have been built during the course of the project. The first phase used two heavily modified speakerphones as a 2-to-4 wire converter and a speech recognizer as a crude DTMF decoder! The following descriptions will discuss the hardware configuration in its current state.




**Hardware configuration.**

### 2.1.1 Sound System

The sound system is a set of software routines that manipulate data on a magnetic disk and a specialized hardware device known as the *soundbox* [Vershel 80]. It is an eight channel system that permits up to four different sounds to be played simultaneously.

The sound box, designed and built in-house, contains a group of analog-to-digital (A/D) and digital-to-analog (D/A) converters. Acoustic input is digitized at a rate of 8000 samples per second using 8-bit linear pulse-code modulation (PCM), providing a signal-to-noise ratio of about 50 dB. This sampling rate permits sounds below approximately 4000 Hz to be accurately recorded and reconstructed. Higher quality voice reproduction can be achieved at the same data rate with logarithmic rather than linear digital encoding. A United States standard for speech coding, known as mu-law encoding, has been developed which provides the equivalent dynamic range of 12-bit linear PCM using only eight bits [Henning 72].

At the rate of 64 kbits per second, storage of the digitized speech becomes expensive in many computing environments (this project has 85 megabytes, or 135 minutes of disk dedicated to sound storage). Many speech compression techniques exist (see section 1.4) for reducing this data rate to 1200-9600 bps while still maintaining intelligibility and the ability to perform speaker identification. This data rate reduction puts speech storage, reconstruction, and recognition tasks within the realm of commercially available personal computers.

### 2.1.2 Voice Synthesizer

A Prose 2000 text-to-speech synthesizer [Telesensory 82] is used to read text messages and provide feedback to the caller. This unlimited-text synthesizer was chosen because of its natural prosodics which are very important for the understanding of free form text [McPeters 84]. (A description of speech synthesizers occurs in section 1.4)

### 2.1.3 Speech Recognizer

The DP-100 Connected Speech Recognition System from NEC (Nippon Electric Company) is capable of recognizing up to five words or "utterances" per

spoken sentence. In this application the DP-100 is primarily used as an isolated word recognizer, yielding significantly higher recognition rates. (A general description of speech recognizers occurs in section 1.2)

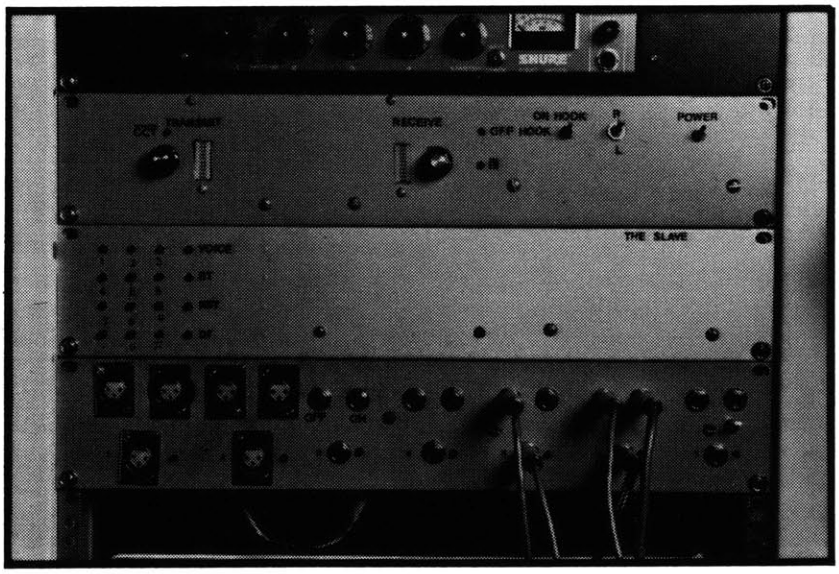
The recognition response time at the end of each sentence is about 300 milliseconds. Output is communicated to the host computer via a high speed serial interface and is also displayed on an alphanumeric plasma display. The device has a maximum vocabulary of 120 utterances, which are stored in the active memory of the recognizer as a set of reference patterns. Each word slot is trained only once and the reference patterns cannot be altered under software control, although individual word slots can be retrained.

The DP-100 speech analyzer performs spectrum analysis, and transforms the input speech signal into 16-dimensional spectrum vectors every 18ms using a digital filtering technique. The filter bank covers a frequency range up to 5900Hz [Tsuruta 79], while the standard voice frequency channel used in telephony is nominally limited to 3000Hz. The DP-100 is therefore unable to use its full range of spectrum analysis capabilities in differentiating utterances. This is particularly a problem for recognition of female voices where there is a larger energy content in the higher frequencies. In order to train the DP-100 in such a way as to minimize the effects of this limited bandwidth all training was performed over the telephone [Dautrich 83].

Statistics of the errors and rejections made by the DP-100 were not kept, however, we experienced a high recognition rate over local telephone lines (within MIT's CENTREX system). Recognition over long distance lines was next to impossible due to the varying types and amounts of background noise. Recognition rates could be increased if a speech recognizer was tailored to work within the limited bandwidth of the telecommunications channel and dynamically adapt to the background noise level. Note that the DP-100 is by no means the ideal recognizer for this application.

### 2.1.4 Telephone Interface

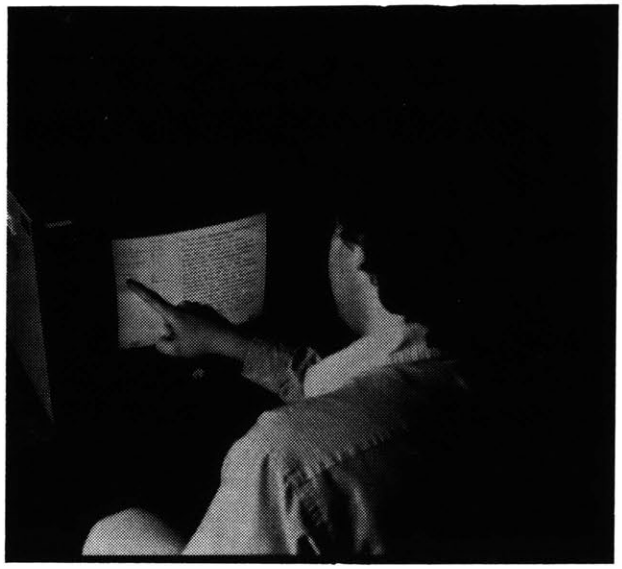
The *phonebox* is a specialized piece of hardware consisting of a subscriber line access circuit, 2-to-4 wire converter (hybrid circuit), DTMF decoder, and audio switch bank that communicates with the host computer over a single serial line. A Hayes Smartmodem is used as an auto-dialer and to detect an incoming ring signal on the phone line [Hayes 82]. This modem was chosen, rather than building the equivalent tone generation circuitry, because of its relatively robust software interface and a desire to explore mixed data and voice communications over the existing telephone network.




**The phonebox: 2-4 wire converter, DTMF decoder, and audio switches.**

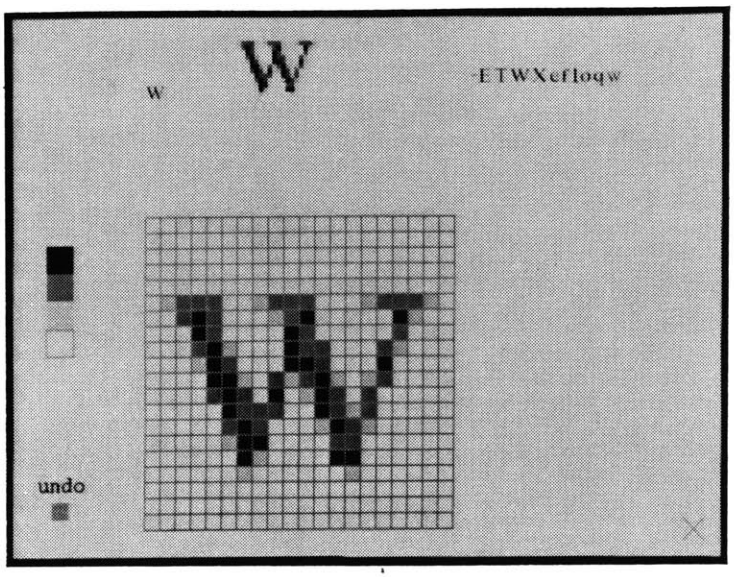
A 2-to-4 wire converter is needed to separate the transmit and receive signals from the telephone line, which is a two-wire circuit that provides full duplex operation. The 2-to-4 wire converter only provides 20dB of isolation, an insufficient amount to prevent outgoing audio (e.g synthesizer or digitized voice) from being fed back to the speech recognizer. In our current configuration it is not possible to perform recognition while the machine is speaking, so the DP-100 is temporarily disconnected from the audio circuit. A better hybrid circuit or a voice operated switch monitoring the line would permit the caller to interrupt the machine by voice.

### 2.1.5 Graphical Interface



**Color graphics workstation with touch screen.**

Graphics and text are generated in a Ramtek 9300 frame buffer with 640x480x9 bit resolution. The touch sensitive display (TSD) is a commercially available, visually transparent, digitizer that is overlayed on the face of a video monitor [Elographics 80]. The grayscale or "soft" fonts that are used for text display were developed at the Architecture Machine Group for dense display of near-print-quality text on a conventional NTSC color television [Schmandt 83]. Characters are displayed as two-bit images, spatially low pass filtered, producing what is perceived as a high quality resolution image on a low resolution display. A standard resolution television monitor is advantageous in that it has the ability to mix computer generated graphics with other video images such as a video disc [Arons 84, Gano 83].




**Editor used in designing the grayscale fonts.**

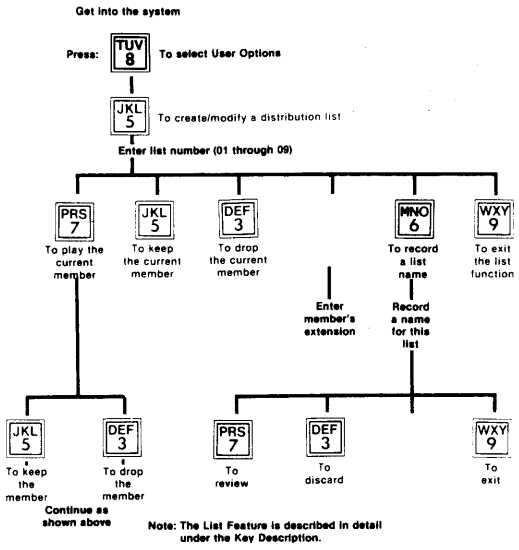
Color is used to convey information as well as to be pleasing to the eye. Text and graphics are displayed in low saturation colors on a neutral background. When an element is touched it is highlighted to visually confirm that the action has been registered, when the command is completed the colors fade into background.

## 2.2 Voice Reading of Electronic Mail

Electronic mail (Email), conventionally read on a CRT terminal connected to a centralized computer, is a common and growing mode of communication in many commercial and academic environments. A user can quickly summarize his correspondence, disregarding cryptically encoded headers, while mentally noting which messages are long or require immediate attention. He has the ability to selectively read, delete, or respond to individual letters. While at home or traveling, access to a terminal and a modem may be limited, and a regular user of Email may be out of touch. Inexpensive portable computers and terminals with integral modems are becoming more prevalent, but remote access to electronic mail without special hardware is desirable.

A voiced electronic mail facility developed at the Architecture Machine Group provides simple efficient access to an electronic mailbox using a Touch-Tone phone as a remote terminal [Baker 83]. A text-to-speech synthesizer is used for reading the text of the letters and providing instructions to the user. This facility is currently operating as a general utility at the Architecture Machine with approximately 20 people actively using it. When a call is placed to read one's mail, the caller must identify himself with Touch-Tones. We currently use the caller's seven digit home phone number as a unique identifier, with an optional four digit password.

There are many approaches and possible mappings of the functions of a DTMF driven system onto the twelve keys of a Touch-Tone telephone. A common method is to relate the letters on the key with a mnemonic that corresponds to some system function. For example in IBM's Audio Distribution System (see section 1.5.3) uses the PRS key to RECORD a message. One drawback to this technique is that the assignments of letters to keys is predefined, and it may be difficult to find suitable mnemonics that can be used on all the keys. Another approach, often used in conjunction with mnemonics, is the use of a hierarchical menu structure. This technique permits a virtually unlimited number of commands using only 12 keys, however the command tree is often complex and may require a manual describing all possible options.




**A commercial hierarchical command tree.**

The design philosophy for the Phone Slave system at the Architecture Machine Group was a little different. The goal was to distill the desired Email functions down to a minimum repertoire so that a tree structured command set would not be needed. A geographical rather than mnemonic approach was used for the assignment of commands to keys, for example, NEXT MESSAGE is adjacent to the PREVIOUS MESSAGE key, both of which neighbor the NEXT SENDER and PREVIOUS SENDER keys.

| | | |
|---|---|---|
| 1 Next Message | 2 Previous Message | 3 Repeat |
| 4 Next Sender | 5 Previous Sender | 6 More Info |
| 7 Yes | 8 No | 9 Reply |
| * Cancel | 0 Pause/ Continue | # Quit |

**Keypad command layout for the Phone Slave.**

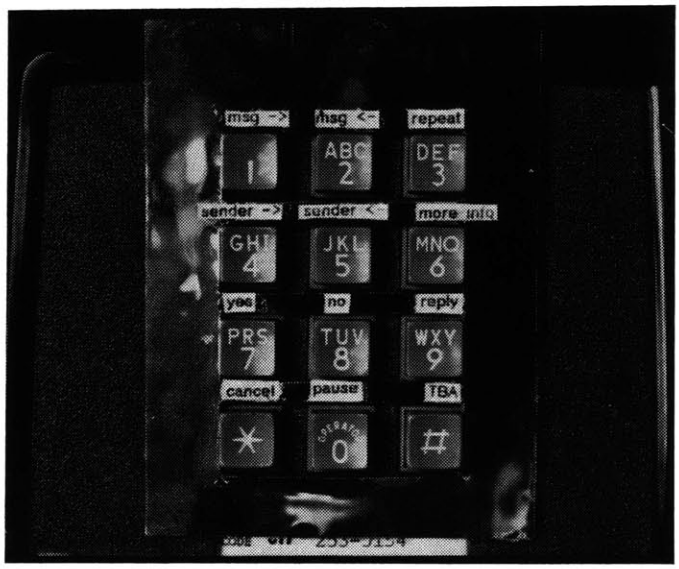
Another important design consideration is that the system is always interruptible and responsive. The choice of appropriate default actions is significant in this type of interactive system. The default condition, in this case, is to lead a naive user through all possible options and play all the messages. An experienced user can anticipate the action, interrupt the synthesizer, and re-direct the flow of control. This feature makes the system truly interactive, allowing the user to do what he wants when he wants it.

As speech is relatively slow compared to reading, it is important to present spoken information in a coherent format, providing the user with as much information as possible, without overloading him with useless details. When viewing text messages on a terminal it is possible to scan them and move amongst them in a random access fashion. It is easier to understand messages and their inter-relationships when the information is presented visually by the screenful than when read orally in a sequential fashion due to short-term memory limitations [Miller 56, Luce 83].

This situation is analogous to reading a newspaper versus watching a television news broadcast; in both cases the information is presented to the

viewer, but it is possible to browse with the newspaper. The reader can jump between stories that are of interest, rather than hearing about stories that the newscaster chooses. Messages, as in the case of news, should be presented in order of importance rather than chronologically. In this voiced mail system they are grouped by the the originator of the message and sorted in order of the number of messages from each sender. This feature exploits the fact that multiple messages from the same person are usually indicative of an ongoing coherent set of correspondence.

The mail is preprocessed, stripping off the body of text and removing the bulk of the header information. The header usually contains the time the message was sent and received, the sender's network mail address, etc. This information, while certainly important, becomes time consuming and burdensome to to hear at speech rates. By default it is not presented, but available on a per message basis through the use of the the MORE INFO key.




**An early Phone Slave keypad template.**

Mail is presented in such a fashion as to provide the user with as much information as possible without forcing him to hear the entire contents of a message. A short introduction such as "this is a rather long message, it's

about..." is spoken to give an indication of the content and length of the text. A short message is simply read without any introduction. The synthesizer's pitch is altered when it is speaking about a message as compared to when it is actually reading the text. This differentiation was explored as a means to increase the ability to comprehend the transitions between messages.

While the intelligibility of the Prose 2000 is significantly greater than its predecessors, it can at times be very difficult even for a trained user to understand [Nusbaum 83]. Mispronunciations commonly occur with acronyms and proper names, particularly those of foreign origin. In order for names and abbreviations that are used frequently in the environment of the Architecture Machine to be pronounced properly, the text is preprocessed through an external software dictionary before being sent to the synthesizer. If a word cannot be understood, the REPEAT key permits the current sentence to be played at a slower rate. If the REPEAT key is pushed a second time within the same phrase, the synthesizer is put into spell mode, thus enunciating each letter.

Here is a sample dialogue with the voiced mail system:

Please enter your I.D., use the star key to cancel the transaction and restart.

720 0232

Hello Barry, nice to hear from you.
You have five new messages.
Do you want to read your mail? Press 7 if yes, 8 if no.

YES-7

4 from Chris Schmandt.
Message 1, it's about "C.E.S. Paper."
Do you care to hear this one?

(no response) (after a short pause the default is to play the message)

We should start working on a longer and more complete version of the ICCE paper for the IEEE Transactions on Consumer Electronics. Can we talk about it later today?

REPLY-9

What kind of reply would you like? 7 for affirm...

YES-7 (interrupts to send affirmative reply)

Mail was sent.
Message 2, this is a very long message, it begins "New synthesizer specifications."
Do you care...

NEXT SENDER-4 (interrupts)

Ok, next sender.
1 from Louie Weitzman.
I'm glad to hear you're getting a new bike.
Have you seen a Bianci?

REPEAT-3

I'll repeat it for you.
Hhaavvee yyoouu sseeeenn aa BBiiaannccii?

REPEAT-3

OK, I'll spell it for you.
H-A-V-E Y-O-U S-E-E-N A B-I-A-N-C-I?

MORE INFO-6

You want more info: Louie Weitzman, logged in as weitzman@NPRDC, yesterday at 5:25 PM.

.
.
.

## 2.3 The Interactive Answering Machine

A parallel research effort was an interactive answering machine, an intelligent message taking system which asks questions and records responses. The machine uses a limited amount of speech recognition as well as some simple heuristic measures to steer the conversation in the right direction. The conversation is robust enough to handle naive callers as well as those familiar with the system. The stereotypical consumer answering machine's "I can't get to the phone right now, please leave your message at the tone... BEEP", does not provide the caller nor the machine's owner with much information. Rather than being rigidly tied to a single informational announcement and fixed length recording, the conversation can take many paths depending upon the caller's identity (see diagram on page 67).

At this point in time it is not feasible for state-of-the-art speech recognizers and computers to monitor and analyze a free form human reply to a question. It may be possible to use a speech recognizer to do keyword spotting, but the syntactic and semantic analysis of these fragmented sentences is beyond the scope of current software and hardware. Rather than attempting to tackle these difficult problems in this manner, a much simpler scheme was used.

The computer seizes control of the conversation and leads the caller through it by asking questions and listening to the responses. In order to format the responses for later retrieval, the questions are stated in such a way that they elicit a very specific response. Five questions are asked (see section 3.1 for sample dialogues), the responses to which can be categorized as: caller's name, topic of the call, telephone number, time to call back, and a full message.

When someone calls they are greeted by the machine and queried for their name. The machine's voice in this case is a human voice that has been previously recorded on the sound system. For our prototype system a female

voice was chosen to contrast it with the owner's voices (currently both male), not because of cultural stereotypes [Leveen 83].

The greeting is of the form *"Hello, Barry's telephone speaking, who's calling please?"*. The phrasing of this sentence 1) greets the caller according to the American custom, 2) informs him of who's telephone he has reached, 3) states in a nonobtrusive manner that he is talking to a machine, and 4) asks him for his name. The soundbox is now put into record mode and the DP-100 is connected to the audio line to monitor the caller's voice.

The response is recorded until the caller has stopped talking, a point which is determined by an adaptive pause and background noise detector. The data from the soundbox A/D are analyzed in real-time to determine the background noise level. There are several timeouts associated with the recording of a message: 1) how long to wait for the person to start talking, 2) after he has started, the length of silence necessary to determine that he has stopped talking, and 3) the maximum message length. The first timeout is fixed while the second can dynamically increase if the person is speaking slowly with pauses between words or phrases.

If the caller exceeds the maximum message length, it is assumed that he is not answering the question properly. The machine interrupts the conversation in a louder voice, informs the caller that he is speaking to a machine and asks him to simply answer the question. After a message has been recorded it is power normalized so that all messages, regardless of the level of the speaker's voice or the telephone connection, can be played back at the same volume. Although not currently implemented, the machine could ask the caller to speak up if it sensed he was speaking to softly.

The DP-100 is designed to be neither a speaker identification nor a speaker verification system [Rosenberg 76]. Within the context of this project, however,

we did use the speech recognizer as a means of identifying callers. The DP-100 was trained to recognize the owner and his most frequent callers identifying themselves over phone. People generally identify themselves in the same manner when speaking over the phone (e.g. "Hi, this is Barry" or "It's Chris"), and this *identification signature* is used to train the DP-100. After a conversation with an unknown caller the speech recognizer is trained on their recorded identification signature. This permits the answering machine's owner to leave personal recordings for all callers, even those who are not currently entered into the owner's telephone directory.

If the caller's identification signature is recognized, the conversation takes a different branch and carries on a more personalized dialogue. A frequent caller is familiar with the machine and vice versa, so it is not necessary to ask the default questions. Information, such as the caller's telephone numbers, is known to the machine so these items are not requested unless they need to be updated. If a known caller is inadvertently not identified by voice, he may enter his ID[4] with Touch-Tones once he realizes that he was not greeted by name.

A known caller receives a personal recording from the owner and an acknowledgment as to the status of any messages previous left by the caller (see section 3.1). The message that a caller receives can be of three of different types: 1) a personal message from the owner to the caller, 2) a message for a class or distribution list of people or 3) a general *message of the day*. It is possible to leave a message for everyone who belongs to the class *Architecture Machine Group* that says "I will be home tonight working on my thesis," while a message could simultaneously exist for the class *DKE fraternity brothers* that says "Let's get together tonight for a few beers."

The message of the day can be created at anytime or it may be chosen from

---

[4]As with the previously described Email system, his home number.

a selection of previously recorded standard messages, such as "I'm out to lunch and will be back at 1:00." At present, these recordings must be explicitly selected by the owner. The selection of outgoing recordings could be automatically invoked by the machine based upon the owner's weekly schedule. For example, the recording that states "I am in Lippman's Digital Video class today, try again after 3:30" could be activated on Tuesday and Thursday during the appropriate hours.

# Chapter Three

## The Personal Integrated Telecommunications System

> *This one might be too far fetched. Wouldn't it be nice to be able to walk into your office and say out loud "Call Harry" and the phone would automatically dial Harry without having to touch any buttons. Computers can easily be taught to recognize simple voice commands. Why not phones?*
>
> Holiday Wish List, Teleconnect Magazine, December 1983

Though involved in the design and implementation of the previously described subsystems, my work focused on the integration of these tools with other telephony functions, such as auto-dialing and directory management, to create a unified user interface[5] to a personalized telecommunications system. Interaction occurs in three forms; locally via a color raster display terminal with a touch sensitive screen, or remotely over the phone using either voice commands or Touch-Tones [Schmandt 84a, Schmandt 84b]. This multi-mode access strategy provides convenient, efficient, and flexible message retrieval through a consistent user interface.

Messages gathered from two sources, electronic mail via a local area computer network and voice messages from the interactive answering machine, can be readily accessed by the owner. These messages are aggregated so that the owner can view them from a single source. He no longer has to deal with a computer and terminal to view his electronic mail messages and a different computer and terminal (probably a telephone with a Touch-Tone keypad) to hear his voice messages.

---

[5]The system owner's interface.

There are three graphics screens which allow touch access to all the PITS' functions. A *message summary screen* shows the status and permits access to all incoming messages. The *card-file screen* enables the creation of outgoing recordings through the use of an on-line directory. The *keypad screen* allows dialing by name and the placing of other outgoing calls. Transitions between screens are made through the small ideograms at the bottom of the screen.

Voice and text communications are represented on the summary screen by bars indicating the length and current status of each message. Mail messages can be viewed locally by displaying a text window on the screen. Remotely electronic mail is read using the voiced mail facility previously described.

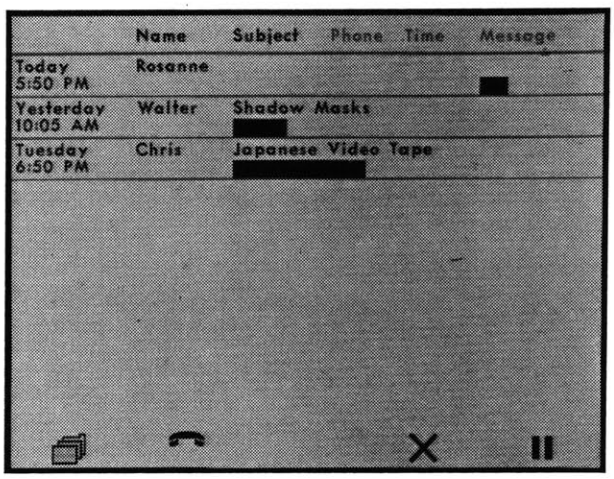
## 3.1 Demonstration

The next few pages give a brief introduction to the PITS as it is currently implemented. The same dialogue is shown in parallel for the three interaction techniques to highlight their similarities and differences.

The following notation is used to differentiate the speech sources:

- The system owner and callers are in the standard text font.
- *System prompts played from the soundbox are in italics.*
- RECORDED MESSAGES AND OUTGOING RECORDINGS ARE IN SMALL CAPITALS.
- `The text-to-speech synthesizer is in typewriter font.`
- **KEYS PRESSED ON THE TOUCH-TONE KEYPAD ARE IN SMALL BOLD CAPITALS.**

---




The telecommunications system at its initial state displaying three unviewed messages. The first message is a voice message from a known caller. The second and third messages display a subject field extracted from the electronic mail message header.

---

An example dialogue with an unknown caller:

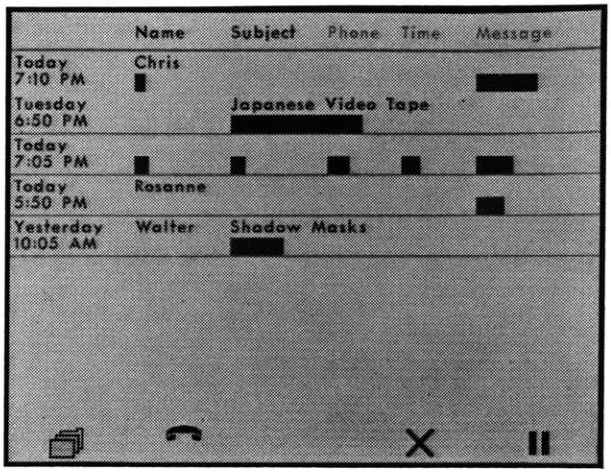
**machine:** *Hello, Barry's telephone speaking, who's calling please?*
**caller:** Charlie Evans
**m:** *What is this in reference to?*
**c:** My visit to your lab.
**m:** *He's not available at the moment, but he left this message:*
<Barry's voice>I'M OUT FOR A LONG BICYCLE RIDE, BUT SHOULD BE BACK AT THE LAB AROUND 8:00.
*At what number can he reach you?*
**c:** 494-8683
**m:** *When will you be there?*
**c:** I should be around here the rest of the evening.
**m:** *Can I take a longer message?*
**c:** I just wanted to confirm tomorrow's visit to the Architecture Machine.
**m:** *I'll be sure he gets that. Thanks for calling. Goodbye.*
**c:** Bye.

---

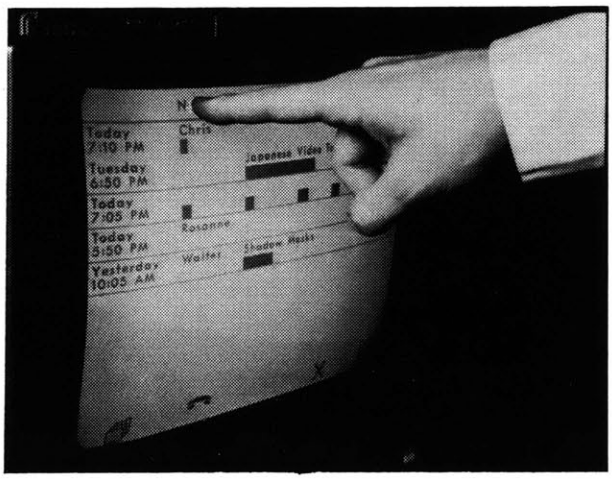
An example dialogue with an known caller:

**m:** *Hello, Barry's telephone speaking, who's calling please?*
**c:** It's Chris.
**m:** *Hello Chris. He got your last message.*
*<Barry's voice>* CHRIS, THANKS FOR THE INTONATION DETECTION ARTICLES, BUT I DON'T THINK I'LL HAVE TIME TO PUT ANY OF IT IN MY THESIS.
*If you'd like to leave another message I'll record it now, otherwise, just hang up and I'll tell him you called again.*
**c:** Now that I've got the Linear Predictive Coding software for the IBM PC running we should try doing speech recognition using the LPC coefficients.
**m:** *If you can't be reached at your usual number, please tell me where to have him call.*
**c:** (could have spoken number or used keypad to enter 256-9562)
**m:** *Thanks, I'll give him your message. Bye*

---



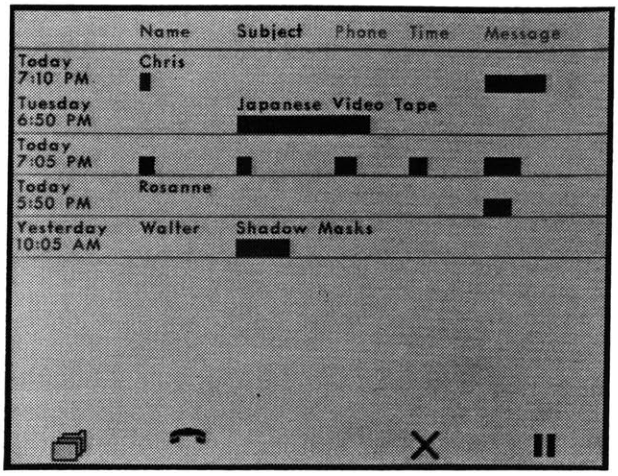

Message screen after arrival of the two new voice messages. Note that the two messages from Chris are grouped together, and the five message segments from the unidentified caller (Charlie).



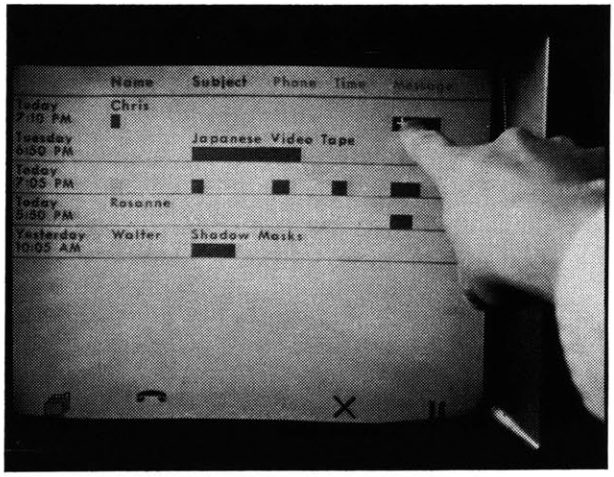
Upon returning to his personal computer, the owner touches the top of name column to see who called.

The owner calls his machine and identifies himself:
**machine:** *Hello, Barry's telephone speaking, who's calling please?*
**owner:** Hi this is Barry.
**m:** *Hi Barry,* you have five new messages. Two from Chris.
**o:** Who left messages?
**m:** CHRIS, CHARLIE EVANS, Rosanne, Walter.

**o:** 720-0232 (enters his ID with the keypad)
**m:** *Hi Barry,* you have five new messages. Two from Chris.




As CHARLIE EVANS is played, the red bar changes color in sync with the recording to give a a visual indication of the proportion of the message that has been played.



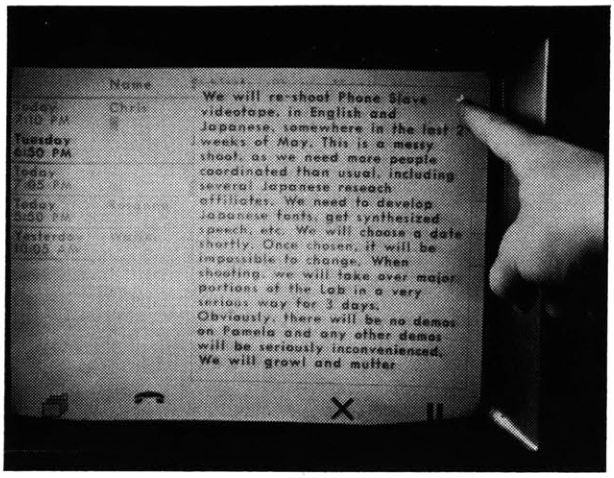

Any individual message segment may be viewed by directly touching it. In this case Chris's voice message about LPC is heard.

**o:** What did he say?

**m:** NOW THAT I'VE GOT THE LINEAR PREDICTIVE CODING SOFTWARE...

**m:** Message one.

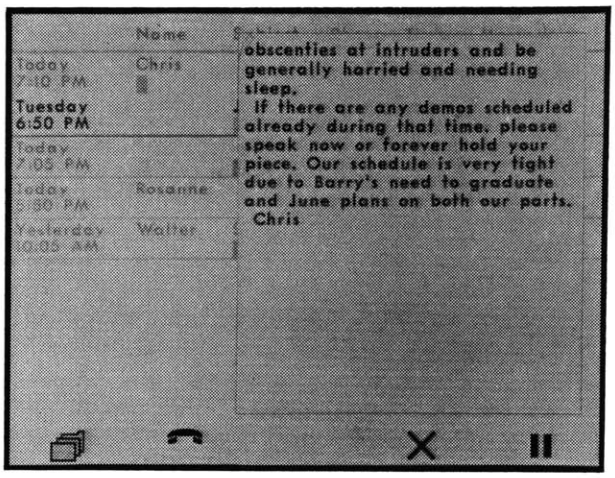
NOW THAT I'VE GOT THE LINEAR PREDICTIVE CODING SOFTWARE...



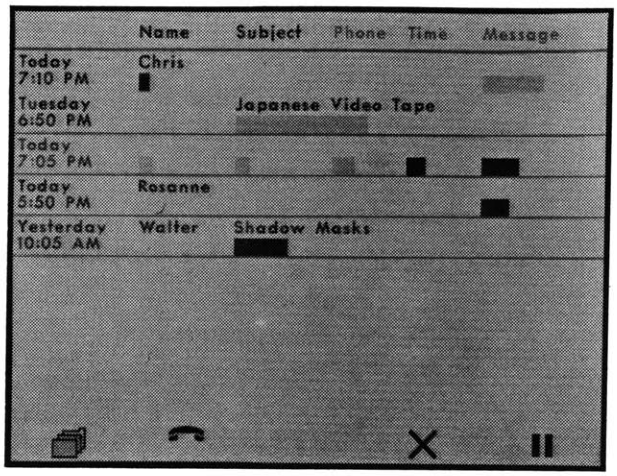

When a text window is displayed, the remainder of the display is faded. A long text message can be paged through by touching the *page turn* symbols in the upper corners.

**o:** Next Message.
**m:** `Message two, it's about "Japanese Video Tape."`
**o:** What'd he say?
**m:** `We will re-shoot Phone Slave videotape, in...`

**m:** `Message two, its about "Japanese Video Tape."`
`Do you care...` (interrupts)
**o:** YES-7
**m:** `We will re-shoot Phone Slave videotape, in...`




The second page of text.

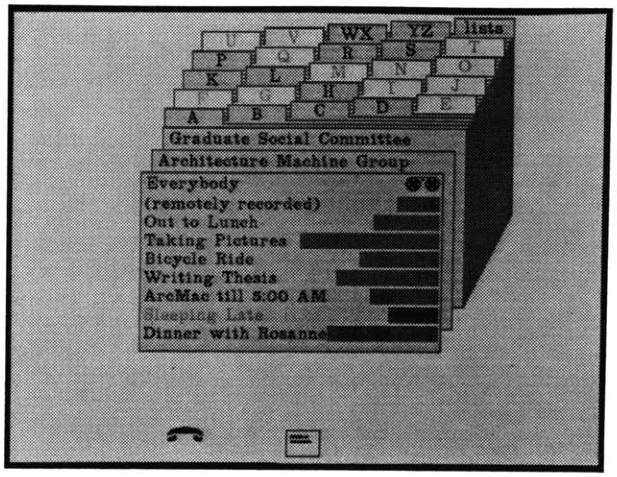



The entirety of a message, in this case the unidentified caller's, can be heard by touching the date box on the left side of the screen. The message segments which have been viewed are displayed in a gray similar to the background, the time segment is currently being played.

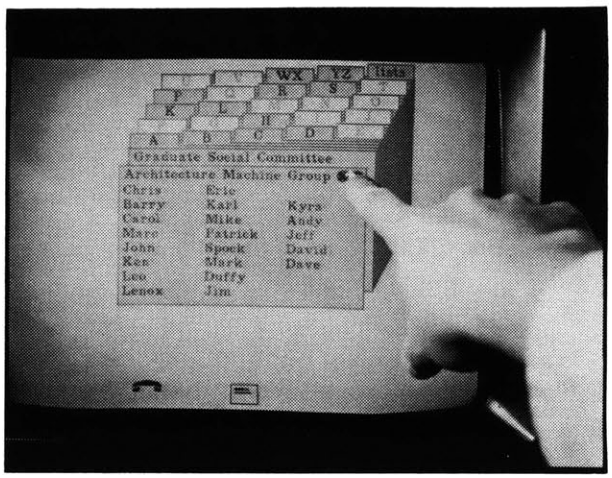
**o:** Next person.
**m:** One from CHARLIE EVANS.
**o:** What's it about?
**m:** MY VISIT TO YOUR LAB.
**o:** What's the message?
**m:** I JUST WANTED TO CONFIRM TOMORROW'S VISIT TO THE ARCHITECTURE MACHINE.
**o:** What's his number?
**m:** 494-8683
**o:** When should I call?
**m:** I SHOULD BE AROUND HERE ALL EVENING.
**o:** When was it?
**m:** Message received this evening at 7:00.

**o:** NEXT SENDER-4
**m:** CHARLIE EVANS... MY VISIT TO YOUR LAB... I JUST WANTED TO CONFIRM TOMORROW'S VISIT TO THE ARCHITECTURE MACHINE... 494-8683... I SHOULD BE AROUND HERE ALL EVENING.

**o:** MORE INFO-6
**m:** Message received this evening at 7:00.




Touching the small card-file image in the lower left hand corner causes the directory screen to be displayed. The top *everybody* card is used to create and update the current default message of the day. The currently selected outgoing recording is *sleeping late*. Touching any card causes that card to be brought to the front.

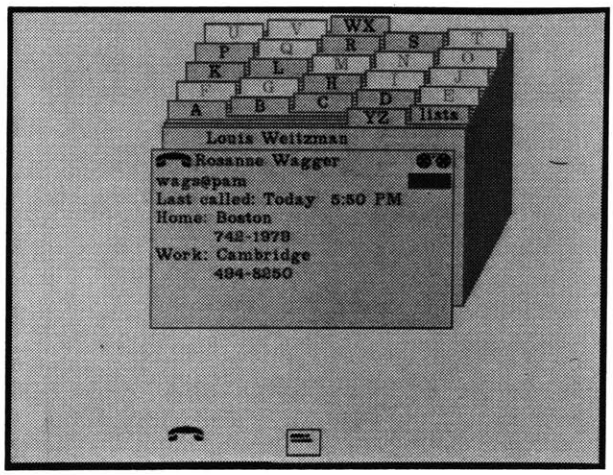



The small record symbol permits an outgoing recording to be recorded for the *Architecture Machine Group* mailing list.

**o:** Take a message for the Architecture Machine Group.
**m:** *Ready to record.*
**o:** Charlie Evans is up from Florida and will be visiting the lab at 4:00, the Media Room and Pamela will not be available until 5:00 pm.
**m:** *Stopped recording.*

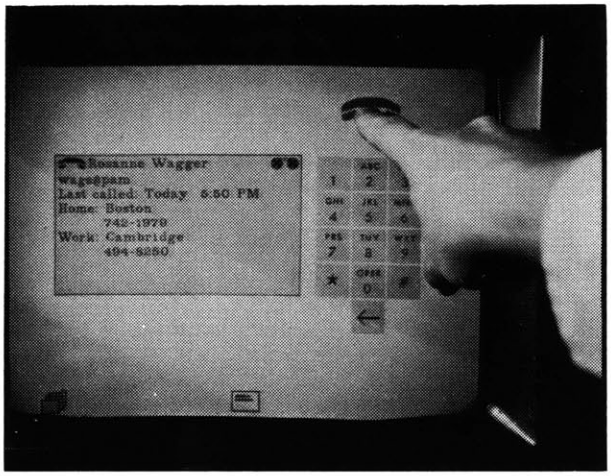
**o:** REPLY-9
**m:** `Please keypress the name of the person or group for whom you wish to leave a message.`

**o:** ABC-2 PRS-7 ABC-2 (enters name using the the keypad)
**m:** `Recording a message for the Architecture Machine Group.`
*Ready to record.*
**o:** Charlie Evans is up from Florida and will be visiting the lab at 4:00, the Media Room and Pamela will not be available until 5:00 pm.
**m:** *Stopped recording.*

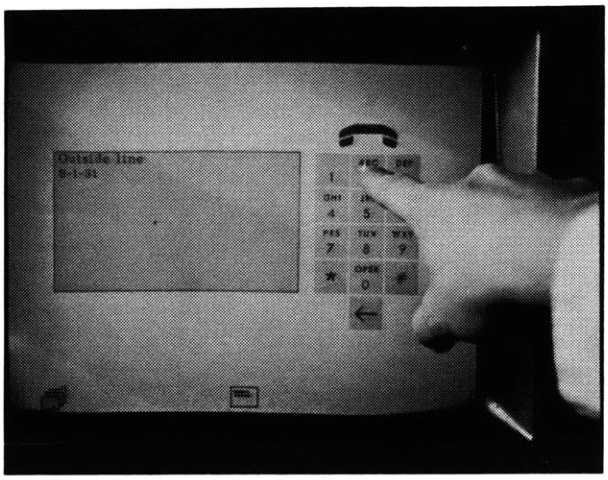



Touching the WX tab brings the appropriate directory entries to to front. The bar under the record symbol indicates there is a personal recording pending for Rosanne. A phone call is placed by touching the handset ideogram in the upper left corner of the card.

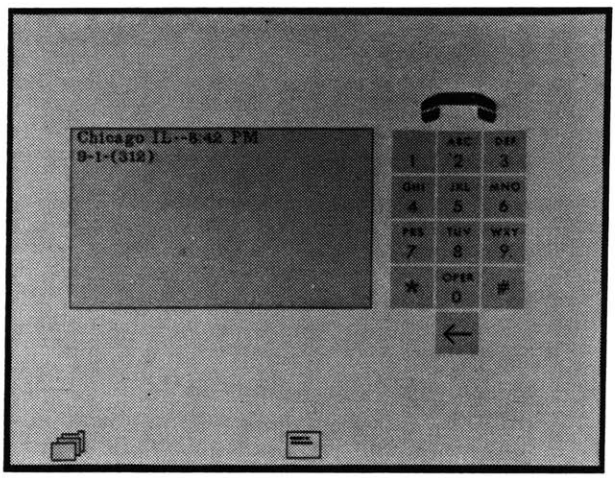
**o:** Tell me about Rosanne.
**m:** `Rosanne Wagger, her home number is 742-1797, work number is 494-8250`




The home or work telephone number is selected based on the time day. A call can be disconnected by touching the phone, causing it to drop to is normal position.




Telephone numbers may be entered directly using the keypad, the backspace key can be used to correct mistakes.




The locations for area codes and local exchanges are displayed after three digits have been entered. The local time is displayed for calls placed to other time zones.

## 3.2 Design Considerations

The owner's interface makes no distinction between text messages and voice messages, there is only a simple generic *message*. Interaction in this system is actively directed by the owner[6] who queries and converses with the machine.

Access is multi-modal, with the same basic information accessible through any input technique. Upon receipt of a command immediate feedback is given, the style and content of which depend not only on the state of the messages, but upon the input technique as well. For example, the "Next sender." command entered by voice will cause the machine to say *"one from Chris"*, and wait for further directives before playing any piece of Chris' message. If the NEXT SENDER command is registered through Touch-Tones, the entirety of Chris' message will be played, as this is the limit of command detail available through the keypad.

As with the voiced Email subsystem, the messages are aggregated by the originator of the messages, thus making it easier to follow an ongoing dialogue when remotely accessing the messages. The relative weighting of voice and text messages is alterable. The owner may feel, for example, that a voice message has more urgency than a mail message and that it should have a higher priority. The owner can also tell the system "I'm expecting a message from Rand" which will cause any messages from Rand to be flagged and presented first.

## 3.3 Remote Access

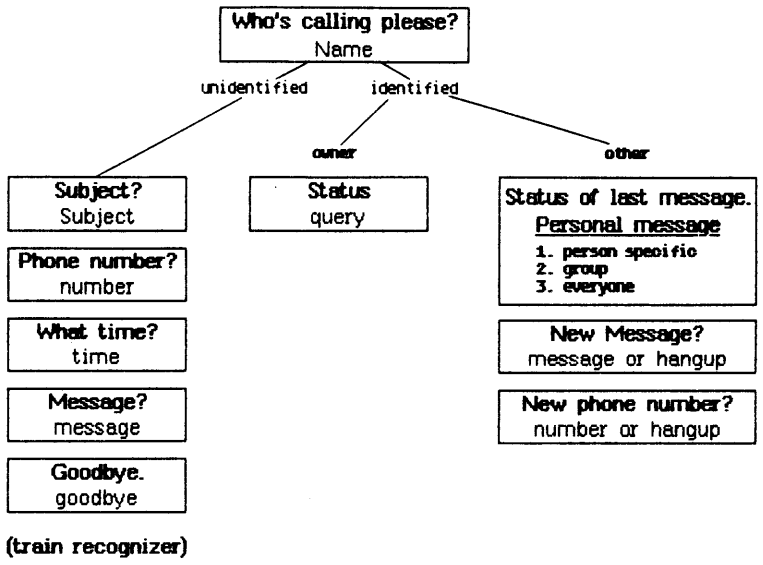
The primary mode of remote message access is by voice, although a limited number of commands are available through Touch-Tones[7]. The keypad provides

---

[6] In the voiced mail system all messages would be heard by default without intervention by the user (section 2.2).

[7] The keypad layout is consistent with the voiced mail facility.

a backup channel for voice commands over noisy phone connections and allows the synthesizer or soundbox to be interrupted. Voice access permits detailed bits of information to be extracted due to a large vocabulary. Upon calling his personal telecommunications manager, the owner identifies himself, is greeted by the machine, and given a brief summary of its status. From this point on the system is actively driven by the owner; commands are spoken, and aural feedback is immediate.




**Tree of possible conversations.**

There are three main command classes. Global commands allow queries as to the overall state of the system. For example, "Who left messages?" causes the names of the message originators to be listed, the overall state of the system is left unchanged. Relative commands such as "When was it?" or "Next person." are used to move among the messages or within the current message. The last class of commands, pertain to a specific person or group. "Tell me about Walter." would cause information about Walter, such as his phone number and whether he called in to receive his personal recording, to be spoken.

Personalized recordings can be created and left for any individual or distribution list. No voice storage is assumed to exist elsewhere, all messages are

stored locally. If the party has an electronic mail address, mail of the following form is sent informing them of their pending voice recordings:

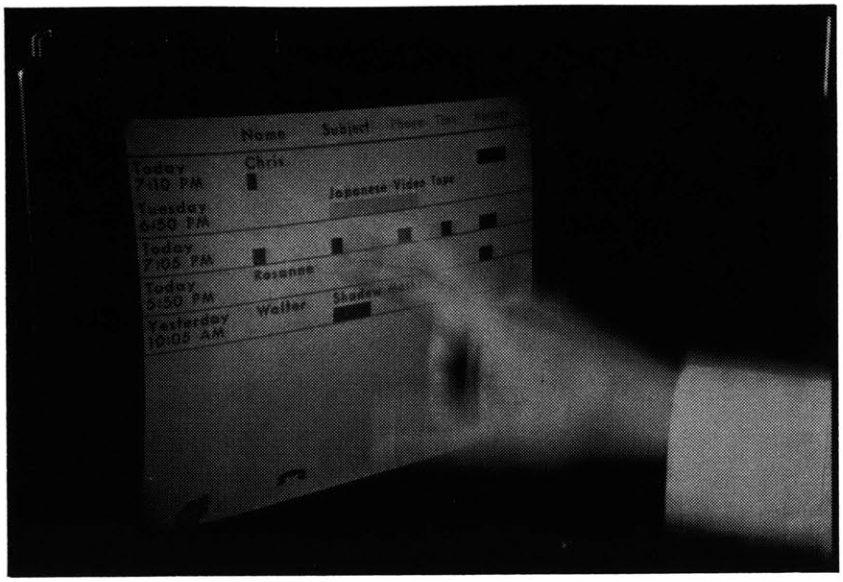
> Date: Wednesday 4 January 1984 15:34:21 EDT
> From: Barry Arons <barons@mit-pamela>
> Sender: The PITS <phone@mit-pamela>
> Subject: Reply to your message about Deep Dish Pizza.
> To: Allyson Haut <ally@ucolumbia.bitnet>
>
> I left a voice message for you on my telephone (617) 258-6681.
> You can hear it by identifying yourself or keypressing the
> identification number: 01057.

After receiving this electronic mail message, Allyson could call in to hear this recording in two different ways. She could enter the unique identification number (01057) to play that specific recording, or she could call and identify herself by voice. In this case she would be engaged in a limited conversation as described earlier.

## 3.4 Graphical Access

The bit-mapped display graphically represents all the incoming and outgoing communications, the owner can peruse messages with a simple touch. The screen organization is a two-dimensional analog of what is presented verbally. Visual cues, through color and graphical ideograms, are given regarding the length and status of messages, telephone numbers, etc.

The time and date of arrival of each message is displayed, not in the conventional month/day/year-hours:minutes:seconds format, but rather in times relative to the present. If the message arrived recently, it says "Today", "Yesterday" or "Friday". Times are rounded to the nearest five minutes and presented in AM/PM format. If a message becomes older, the precise time of arrival becomes less important and only the date is displayed.

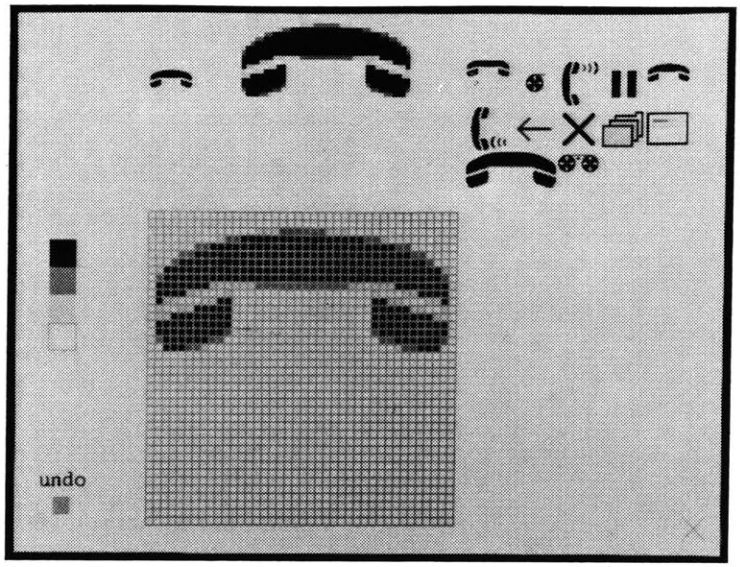



**Touch screen access to information.**

The name of the sender is displayed, if known, but only once per screen to keep the display uncluttered. The interactive answering machine breaks a message up into five distinct pieces; text messages generally consist of a topic and body of text. These consistent classifications permit the message segments to be displayed and organized by content.

Colored bars are used to represent each of the individual message segments. The length of the bar is proportional to the length of the message, its color is indicative of its status. An unviewed message is red, to draw attention. The bar changes to blue in sound sync with a message as it is played. After it is viewed it changes to gray, fading into the background.

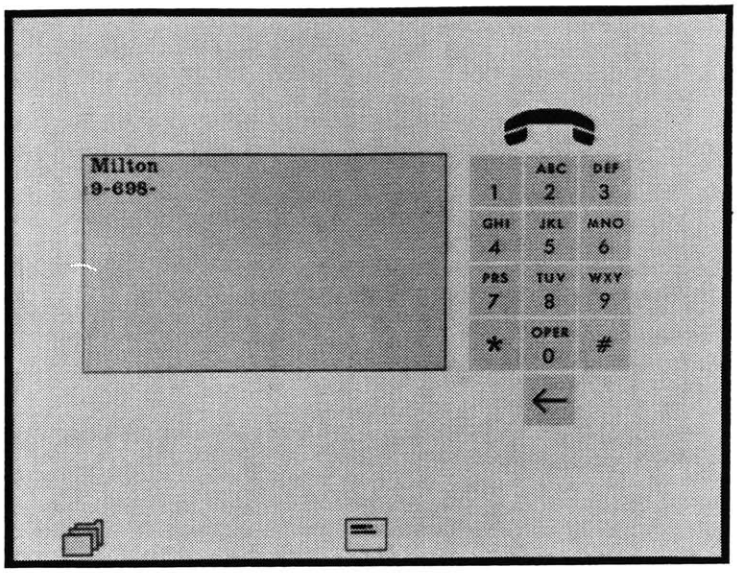
Touching any region immediately highlights it and activates it. An individual segment of a message may be viewed in a random access fashion, or an entire row or column can be played. A command can be interrupted at any time by touching a new region.




**Designing the colored ideograms with the font editor.**

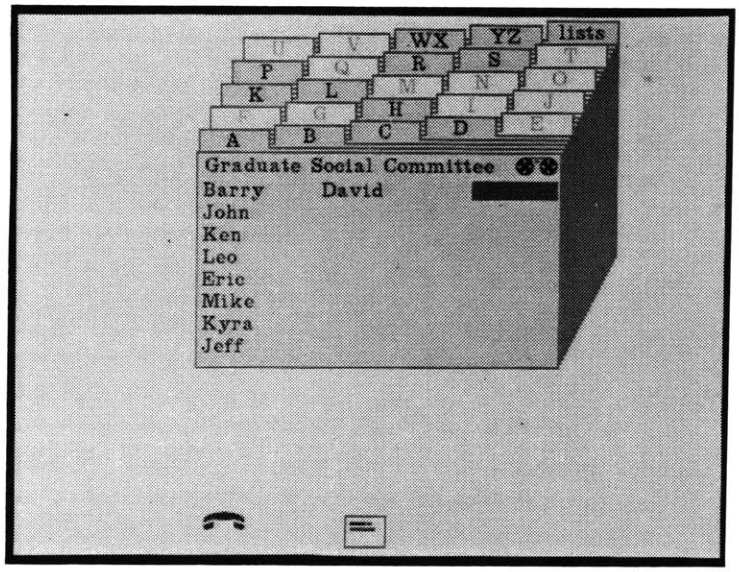
Touching the HANDSET ideogram at the bottom of the screen while viewing a message invokes a new screen which permits calls to be made. It is possible to place or return a telephone call in several ways. The home and work numbers of the individual are displayed, with the preferred number is highlighted based upon the time of day. Touching a number causes the telephone HANDSET ideogram to raise and the number to be dialed. The owner's voice is picked up by a microphone, the called party is heard over a speaker. The call will be disconnected if a busy tone is detected or by touching any other active region on the screen.

If someone is to be reached at a number that is not in the directory, it can be input via the keypad. The number is displayed character by character as it is entered: a backspace key is provided so the entire sequence need not be re-entered if an error is made. After three digits have been entered the location of the called number is displayed based upon the area code or local exchange. If the location is in a different time zone, the correct time for that locale is displayed.




**Example of local exchange lookup.**

A card-style directory permits dialing-by-name and the creation of several classes of outgoing recordings (see section 2.3). The file contains 25 labeled cards; A through Z, and *lists* which contains the various mailing lists. Individual cards with entries are displayed in light blue, empty cards are gray.




**A group mailing list with an outgoing recording.**

Recordings can be created for individuals or groups by touching the dual reel RECORD ideogram on the appropriate card. The owner is prompted by

changing the RECORD ideogram to red and by a verbal *"Ready to record"* cue. When the owner stops talking, the recording is stopped and the display is updated. As on the message viewing screen, the colored bar may be touched to review the outgoing recording.

## 3.5 Software Design

The PITS software system adopts some of the principles of object-oriented and data-directed programming, but is implemented within a procedure-oriented context. The availability and creation of software tools and high level interfaces to physical devices was crucial in developing a software system as complex as this. Most device dependent calls are isolated in a few selected low level routines so that only minor changes would be necessary to use an alternative frame buffer, speech synthesizer, or sound system which employs the same high level calling conventions.

Two design considerations played an important part in formulating the the overall structure of the software system; the system had to be interruptible at all times, and multi-modal input had to be handled in a consistent manner. The machine, regardless of its state, can always be asked to stop what it is doing and be redirected to begin another task.

Due to the timing dependent nature of the communication protocol with the speech recognizer, it is not possible to have the DP-100 generate hardware interrupts, so the device must be polled. The input devices produce data in a variety of forms: the DP-100 transmits slot numbers in the range 0-119, the TSD produces x and y coordinates between 0-639 and 0-439, and the DTMF decoder outputs the common 0-9, *, and #. These diverse inputs are all converted into a common form so that, when appropriately tempered by context, they all represent meaningful commands such as "Next message." or "Take a recording for Linda."

## 3.6 Multi-modal Input Interface

With a limited command set it is often necessary to allow a simple global command (e.g. YES or NO) to have multiple meanings dependent upon the environment in which it is invoked. This becomes most apparent when using the limited set of 12 Touch-Tone keys for input. The scheme used throughout the software system is relatively simple yet powerful. Devices are polled from the lowest level of software. Whenever a command is entered, the polling routine halts execution and checks to see if there is a local handler for the particular input received. If such a handler does not exist, the newly entered command is passed up to the next highest level of subroutine, where again there may a local command handler. This happens repeatedly until a local handler is found which can deal with the input. The top level of software can handle all commands.

Input from the various devices are converted into a common command language so that all routines can manage the data in a consistent format. The vernacular used consists of two parts in the form:

<type> <instance>

where <type> is the syntactic category and <instance> specifies the item to be operated on[8]. This scheme allows simple syntactical rules to be easily applied at the lowest level. For example, the command "Tell me about Walter." is broken up as:

<type = 7> a person specific command, <instance = 2> "Tell me about"
<type = 5> a person, <instance = 22> "Walter Bender"

A syntactic analysis routine checks that the types are consistent for a properly formed command. For example, if "Walter" was not recognized, the machine would ask "Which person?"

The system is data-directed in that a request to the dispatcher, or top-level command handler, may be an abstract *play a message* request in which the

---

[8]This technique was first implemented by Eric Hulteen in Put That There [Schmandt 84c].

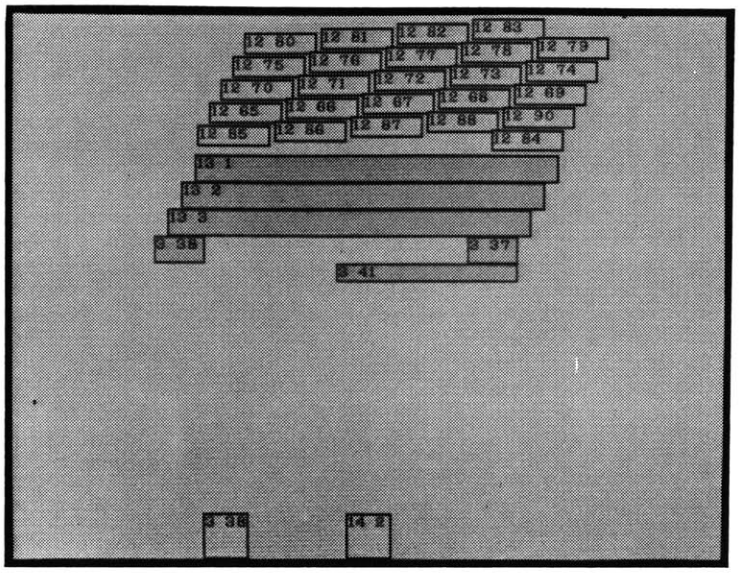
message type is not explicitly stated. The invoked routine is a generic operator, that is, a procedure that operates on a variety of data types, each with a different natural mode of presentation [Abelson 84]. The manner in which the message is viewed is a function of data type and the state of the telecommunications system. Data structures contain the current state of the machine, the current command and its mode of input, whether a message is being viewed, and which devices are currently configured to be in use.

An object-oriented graphical interaction system was developed which permits the types and instances to be properly assigned to the TSD input within the low level polling routine. Commands entered through the keypad or recognizer map directly to types and instances, however, coordinates from the TSD must be tempered by the current state of the display. The same x-y pair from the TSD may mean "What's Walter's phone number?" or "Take a recording for the Architecture Machine Group.", depending upon which screen is currently displayed.

During system initialization a database is created which contains the types and instances for active regions of all possible screen configurations. A storage reduction scheme is used which allows one screen and up to two overlays to be active simultaneously. For example, there are 25 possible positions for the lettered tabs of the card-file, each of which must be maintained as a separate item in the graphical interaction database. There can be up to five personal directory cards present on top of the card file, so rather than using 25*5 = 125 entries[9] for all possible combinations of cards, only 25 + 5 = 30 entries are used.

---

[9]Where each entry contains eight fixed point numbers for each of the 25 + active region on the screen.



**Active regions of the card file with three overlay cards.**

Whenever the display changes, such as when switching to the keypad screen or removing a card from the top of the card file, a call is made to the graphical interaction database to update its current state variables. When a gesture is registered on the TSD, the graphical database is called with the coordinates of the touch, and the type and instance for the activated region are returned.

Another object-oriented database manager serves as a personal telephone directory and maintains all outgoing recordings. Each directory entry contains the time the person last called, a usage count, a linked list of pending personal recordings, phone numbers, and addresses. Requests must be made to the database manager to create, update, delete, or obtain information from the database. This provision insures that the database will always remain internally consistent.

# Chapter Four

## Discussion

*Bit by bit, the telephones are becoming digital dinosaurs. Just as the Touch-Tone surpassed the rotary dial, new technology will make the Touch-Tone beeps obsolete...In fact, if your phone isn't also a personal computer, the chances are your personal computer will also be a phone.*

Michael Schrage, The Washington Post, December 13 1983

The personal integrated telecommunications system is self-disclosing and intuitive, and can be used without any prior training[10]. Commands and interaction techniques are consistent throughout the entire system. The colored *message* bars, for example, are used to both represent outgoing recordings in the card-file as well as pending incoming mail on the message summary screen. The same techniques are used to create recordings for mailing lists or individuals.

Many of the PITS design considerations were meant as an *exploration of the interface*, not as a statement of the definitive way to do things. The date and time format developed, for example, is not desirable in some systems, but has been found useful when this information must be communicated by voice. The card-file is organized so that each labeled card tab always appears in the same relative position (e.g. the "A" card is always on the left). This style of card-file was chosen to assist the owner in locating the card that he uses often, and follows the same physical constraints of an actual Rolodex.

---

[10] A demonstration of the system can be seen by viewing the short videotape which accompanies this thesis. The five minute tape shows the conversationality and intuitiveness of the PITS, something that is extremely difficult to describe in words and still photographs.

Within the Architecture Machine Group, users of the Phone Slave tended to listen to their mail in the default order, and features such as MORE INFO were not used very often. The change in pitch of the synthesizer was not found to be particularly helpful in distinguishing informational announcements from the actual text of the Email messages. Distinctly different voices (e.g. male vs. female) would perhaps be more effective.

The conversational answering machine was found to be surprisingly effective at gathering coherent messages. *On-the-fly* training of unknown callers, while not always successful, is enhanced by the use of a connected-speech recognizer. If someone calls and initially identifies themself as "Beth Carlson", then on a subsequent call identifies herself as "This is Beth Carlson", there is a fair chance that *Beth Carlson* will be identified in the second instance. This would not be possible if an isolated word recognizer were used.

In recent years the use of iconic imagery has been methodically employed in many computer systems and is found to be a significant improvement over traditional text-oriented interfaces [Smith 82]. While the use of ideograms is perhaps undesirable [Lippman 84], they do provide a visual way of presenting abstract information. Conversing with a computer should be accomplished in a natural fashion without the use pictograms, and this is certainly possible with voice interaction. To interact with a graphical display, however, there must be an image of some sort which can be seen and selected.

The underlying software structure of the system will permit other input/output devices and alternative forms of multi-media mail, such as a written or voice annotation over a video message, to be easily integrated into the PITS. Only minor modifications would be necessary to increase the capabilities of this system because device polling and user feedback is controlled from a few low level routines.

Outgoing calls would enable the distribution of messages to the system owner, his acquaintances, or even to other personal telecommunications systems. The telephone network of the future will be entirely digital, the most likely channel being high-bandwidth optical fiber. When this time arrives, the transfer of speech and images between PITS-like personal computers seems highly likely.

Single boards containing a digital signal processing chip are now available for personal computers. These devices are capable of performing most of the audio processing associated with the PITS such as DTMF decoding and generation, and speech coding for speech recognition and reduced storage. A software configurable system based on hardware such as this permits maximum flexibility of resources, and is potentially more useful than a collection of individual devices (e.g. speech coders and recognizers) which hang off a telephone line.

The PITS is more than an intelligent answering machine; it combines multi-modal communications and merges them into a single form so that voice and data messages are not considered separate entities. The interface is responsive, allowing messages to be accessed by voice, keypad, or touch screen with simultaneous verbal and graphical feedback. A personal computer forms a new type of telecommunications environment around its owner, gathering and disseminating information from several information networks.

The personal integrated telecommunications system explores and successfully demonstrates mechanisms for unified voice and gesture interaction through a diverse range of communication tasks. The interface design philosophy provides a cogent framework for the future influx of conversational computers into man-machine interaction.

# Afterword

Chris Schmandt has been my partner in research, office-mate, and friend since the inception of this project. We have spent many long hours brainstorming and writing code, this time has always been interesting, educational, and a lot of fun.

Walter Bender was invaluable in providing programming tools and encouragement throughout my two years at the Architecture Machine Group. His fine sense of form and color helped shape and improve the design and layout of the PITS graphical interface.

Andy Lippman, director of the lab, supported me on this and other projects. He provided inspiration and many ideas, but always encouraged me to think and further develop concepts on my own.

Marc Spehlmann designed, built, and continuously improved the telephone interface hardware used in this project.

Dave Chen took pictures when I couldn't be behind the camera and provided advice concerning the layout of this document.

The Architecture Machine is a unique and very special place, it has become more of a home than my apartment in Boston. I thank everyone in the group for all their help, friendship, and great ideas. Special thanks to my fellow graduate students, particularly Ken Carson, Steve Gano, and Eric Brown for their sense of quality, thoroughness, and ability to understand the big picture in designing intelligent systems.

# References


[Abelson 84]
Harold Abelson and Gerald Sussman.
*Structure and Interpretation of Computer Programs*.
MIT Press and McGraw-Hill, 1984.
Draft

[Allen 76]
Jonathon Allen.
Synthesis of Speech from Unrestricted Text.
*Transactions of the IEEE* 4:433-442, 1976.

[Allen 81]
Jonathon Allen.
Linguistic-based Algorithms Offer Practical Text-to-speech Systems.
*Speech Technology* 1(1), Fall, 1981.

[Arons 84]
Barry Arons.
Discursions.
*Educational and Industrial Television* 16(6), June, 1984.

[Baker 81]
Janet M. Baker.
How to Achieve Recognition: A Tutorial/Status Report on Automatic Speech Recognition.
*Speech Technology* 1(1):30-43, Fall, 1981.

[Baker 83]
Caren Hope Baker.
Voice Access to Electronic Mail.
Bachelor's thesis, MIT, June, 1983.

[Bayer 83]
D. L. Bayer and R. A. Thompson.
An Experimental Teleterminal - The Software Strategy.
*The Bell System Technical Journal* 62(1):121-144, January, 1983.

[Bergland 82a]
G. D. Bergland.
Experiments in Telecommunications Technology.
*IEEE Communications Magazine* 20(6):4-14, December, 1982.

[Bergland 82b]
G. D. Bergland, W. T. Hartwell, and G. W. Smith.
1A Voice Storage System: Prologue.
*The Bell System Technical Journal* 61(5):815-819, May-June, 1982.

[Boyle 81]
Thomas Boyle.
The Design and Implementation of a Crash Proof File System.
Bachelor's thesis, MIT, June, 1981.

[Bruckert 84]
Ed Bruckert.
A New Text-to-speech Product Produces Dynamic Human Quality Voice.
*Speech Technology* 2(2):114-119, Jan/Feb, 1984.

[Cagle 71]
W.B. Cagle, R.R. Stokes, and B.A. Wright.
The Picturephone System: 2C Video Telephone Station.
*The Bell System Technical Journal* 50(2):271-312, February, 1971.
Special Picturephone Issue of BSTJ

[Cornell 82]
R. G. Cornell and J. V. Smith.
1A Voice Storage System: Architecture and Physical Design.
*The Bell System Technical Journal* 61(5):841-861, May-June, 1982.

[Dautrich 83]
B. A. Dautrich, L. R. Rabiner, and T. B. Martin.
The Effects of Selected Signal Processing Techniques on the
Performance of a Filter-Bank-Based Isolated Word Recognizer.
*The Bell System Technical Journal* 62(5):1311-1336, May-June, 1983.

[Elographics 80]
*Graphical Digitizer Operating Manual.*
Elographics, Inc., P.O. Box 388, Oak Ridge TN, 37830, 1980.

[Flanagan 79]
J. L. Flanagan, M. R. Schroder, B. S. Atal, R. E. Crochiere, N. S. Jayant,
J. M. Tribolet.
Speech Coding.
*IEEE Transactions on Communications* Com-27(4):710-737, April, 1979.

[Flanagan 81]
J. L. Flanagan.
Synthesis and Recognition of Speech: How Computers Talk.
*Bell Laboratories Record* 59(4):123-130, April, 1981.

[Gano 83]
Steve Gano.
Forms for Electronic Books.
Master's thesis, MIT, June, 1983.

[Gates 82]
G. W. Gates, R. F. Kranzmann, and L. D. Whitehead.
1A Voice Storage System: Software.
*The Bell System Technical Journal* 61(5):863-883, May-June, 1982.

[Gould 82]
John D. Gould and Stephen J. Boies.
*Speech Filing - An Office System for Principals.*
Research Report RC 9769, IBM, December 14, 1982.

[Gould 83]
John D. Gould and Stephen J. Boies.
Human Factors Challenges In Creating a Principal Support Office System - The Speech Filing System Approach.
*ACM Transactions on Office Information Systems* 1(4):273-298, October, 1983.

[Hagelbarger 83]
D. W. Hagelbarger, R. V. Anderson, and P. S. Kubik.
Experimental Teleterminals - Hardware.
*The Bell System Technical Journal* 62(1):145-152, January, 1983.

[Hayes 82]
*Smartmodem 1200 Owner's Manual.*
Hayes Microcomputer Products, Inc., 1982.

[Henning 72]
H. H. Henning and J. W. Pan.
D2 Channel Bank: System Aspects.
*The Bell System Technical Journal* 51(8):1641-1657, October, 1972.

[Holmgren 83]
J. E. Holmgren.
Toward Bell system Applications of Automatic Speech Recognition.
*The Bell System Technical Journal* 62(6):1865-1880, July-August, 1983.

[IBM 82]
*IBM Audio Distribution System Subscriber's Guide.*
International Business Machines Corporation, 1982.

[Ishii 82]
Naoki Ishii, Yuhji Imai, Ryohei Nakatsu, and Makato Ando.
Speaker-Independent Speech Recognition Unit Development for Telephone Line Use.
*Japan Telecommunications Review* :267-273, July, 1982.

[Itakura 75]
F. Itakura.
Minimum Prediction Residual Principle Applied to Speech Recognition.
*IEEE Transactions on Acoustics Speech and Signal Processing* ASSP-23:67-72, 1975.

[Kazar 78]
Michael Kazar.
Dynamic Linking in a Small Address Space.
Bachelor's thesis, MIT, June, 1978.

[Kazar 80]
Michael Kazar.
*The Only MagicSix Survival Guide.*
1980.

[Klapman 82]
Richard Klapman, Robert Lauver, and Kenneth Welton.
An EPIC Journey Toward the Information Age.
*Bell Laboratories Record* 60(9):240-245, November, 1982.

[Klatt 77]
D. H. Klatt.
The ARPA Speech Understanding Project.
*Journal of the Acoustical Society of America* 62(6):1345-1366, 1977.

[Klemmer 71]
E. T. Klemmer and F. W. Snyder.
Measurement of Time Spent Communicating.
*Journal of Communications* 22:142-158, 1971.

[Leveen 83]
Steven Leveen.
Technosexism.
*The New York Times* :23, November 12, 1983.

[Levinson 78a]
S. E. Levinson.
The Effects of Syntax Analysis on Word Recognition Accuracy.
*The Bell System Technical Journal* 57(5):1627-1644, May-June, 1978.

[Levinson 78b]
S. E. Levinson, A. E. Rosenberg, J. L. Flanagan.
Evaluation of a Word Recognition System Using Syntax Analysis.
*The Bell System Technical Journal* 57(5):1619-1626, May-June, 1978.

[Levinson 80]
S. E. Levinson and K. L. Shipley.
A Conversational-Mode Airline Information and Reservation System Using Speech Input and Output.
*The Bell System Technical Journal* 59(1):119-137, January, 1980.

[Lippman 84]
Andrew Lippman.
conversations.

[Luce 83]
P. A. Luce, T. C. Feustel, and D. B. Pisoni.
Capacity Demands on Short-Term Memory for Synthetic and Natural Word Lists.
*Human Factors* 25(1):17-32, 1983.

[Lundin 83]
Fred Lundin, Mats Blomberg, and Kjell Elenius.
Voice-Controlled Dialing in an Intercom System.
In *Proceedings of the Voice Data Entry Systems Application Conference*. American Voice Input/Output Society, 1983.

[McPeters 84]
David L. McPeters and Alan L. Tharp.
The Influence of Rule Generated Stress on Computer-Synthesized Speech.
*International Journal of Man-Machine Studies* 20(2):215-226, February, 1984.

[Miller 56]
George Miller.
The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity to Process Information.
*Psychology Review* 63(2):81-97, 1956.

[Mulla 84]
Hoshang Mulla.
A PABX that Talks and Listens.
*Speech Technology* 2(2):74-79, Jan/Feb, 1984.

[NEC 82]
Nippon Electric Company.
*Voice Recognition and Voice Response System By Telephone Use.*
Technical Report, NEC-Computer Application System, 1982.

[Nusbaum 83]
Howard C. Nusbaum, Eileen C. Schwab, and David B. Pisoni.
Perceptual Evaluation of Synthetic Speech: Some Constraints on the use of Voice Response Systems.
In *Proceedings of the Voice Data Entry Systems Application Conference.* American Voice Input/Output Society, 1983.

[Nussbaum 82]
E. Nussbaum.
1A Voice Storage System: Voice Storage in the Network - Perspective and History.
*The Bell System Technical Journal* 61(5):811-813, May-June, 1982.

[Parks 79]
Lee Parks.
The Design and Implementation of a Multi-Programming Virtual Memory Operating System for a Minicomputer.
Bachelor's thesis, MIT, June, 1979.

[Pathe 83]
Peter Pathe.
A Virtual Vocabulary Speech Recognizer.
Master's thesis, MIT, June, 1983.

[Rabiner 76]
L. R. Rabiner & R. W. Schafer.
Digital Techniques for Computer Voice Response: Implementations and Applications.
*Proceedings of the IEEE* 64:410-576, 1976.

[Rabiner 78]
L. R. Rabiner & R. W. Schafer.
*Digital Processing of Speech Signals.*
Prentice-Hall, 1978.

[Rabiner 80]
L. R. Rabiner, J. G. Wilpon, and A. E. Rosenberg.
A Voice-Controlled, Repertory-Dialer System.
*The Bell System Technical Journal* 59(7):1153-1163, September, 1980.

[Reddy 76]
D. R. Reddy.
Speech Recognition by Machine: A Review.
*Proceedings of the IEEE* 64(4), April, 1976.

[Rosenberg 76]
A. E. Rosenberg.
Evaluation of an Automatic Speaker-Verification System Over Telephone Lines.
*The Bell System Technical Journal* 55(6):723-7444, August, 1976.

[Rosenberg 79]
A. E. Rosenberg and C. E. Schmidt.
Automatic Recognition of Spoken Spelled Names for Obtaining Directory Listings.
*The Bell System Technical Journal* 58(8):1797-1823, October, 1979.

[Rosenberg 80]
A. E. Rosenberg, L.R. Rabiner, and J.G. Wilpon.
Recognition of Spoken Spelled Names for Directory Assistance Using Speaker-Independent Templates.
*The Bell System Technical Journal* 59(4):571-592, April, 1980.

[Schmandt 82a]
Christopher Schmandt and Eric Hulteen.
The Intelligent Voice-Interactive Interface.
In *Proceedings, Human Factors in Computer Systems*. Association of Computing Machinery, 1982.

[Schmandt 82b]
Christopher Schmandt.
Speech Communications, A Systems Approach.
In *Proceedings of the Voice Data Entry Systems Applications Conference*. American Voice Input/Output Society, 1982.

[Schmandt 83]
Christopher Schmandt.
Fuzzy Fonts: Analog Models Improve Digital Text Quality.
In *Conference Proceedings*. National Computer Graphics Association, 1983.

[Schmandt 84a]
Christopher Schmandt and Barry Arons.
Phone Slave: A Graphical Telecommunication Interface.
In *Digest of Technical Papers*. SID International Symposium, 1984.

[Schmandt 84b]
Christopher Schmandt and Barry Arons.
A Conversational Telephone Messaging System.
In *Digest of Technical Papers*. IEEE International Conference on Consumer Electronics, 1984.

[Schmandt 84c]
Christopher Schmandt.
*Speech Communication with Computers.*
Ablex, 1984, .
In publication

[Smith 82]
David C. Smith, Charles Irby, and Ralph Kimball.
The Star User Interface: an Overview.
In *National Computer Conference Proceedings*. AFIPS, 1982.

[Steinberg 74]
Seth Steinberg.
A PL/1 Subset for a Minicomputer.
Bachelor's thesis, MIT, June, 1974.

[Swinehart 83]
D. Swinehart, L. Stewart, and S. Ornstein.
Adding voice to an Office Computer Network.
In *Proceedings of GlobeComm*. IEEE Communications Society, 1983.

[Telesensory 82]
*PROSE 2000 Text-to-SpeechConverter User's Manual.*
Telesensory Systems, Inc., 3408 Hillview Ave., P.O. Box 10099, Palo Alto CA, 94304, 1982.

[Tsuruta 79]
Shichiro Tsuruta, Hiroaki Sakoe, and Seibi Chiba.
DP-100 Connected Speech Recognition System.
In *Proceedings of the Third International Telecommunications Exposition*, pages 48-52. Horizon House International, 1979.

[Vershel 80]
Mark Aaron Vershel.
The Contribution of 3-D Sound to the Human-Computer Interface.
Master's thesis, MIT, June, 1980.

[Worral 82]
D. P. Worral.
1A Voice Storage System: New Custom Calling Services.
*The Bell System Technical Journal* 61(5):821-839, May-June, 1982.

[Zientara 84]
Peggy Zientara.
DECtalk Lets Micros Read Messages Over the Phone.
*Infoworld* :21-23, January 16, 1984.