Voice Interaction

in an

Integrated Office and Telecommunications Environment

Christopher Schmandt, Barry Arons, and Charles Simmons
Media Laboratory, Massachusetts Institute of Technology

Proceedings, American Voice Input/Output Society Conference 1985, AVIOS, Palo Alto, CA, 1985, pp. 51-56.

Introduction

The *Conversational Desktop* explores the use of speech input/output technologies for machine mediated voice communication in an office and telecommunications environment. Central to this work is an interface design which models several aspects of human conversational behavior. These include the ability to carry on a dialog to resolve ambiguous input, the ability to apply syntactic and acoustical context to the progress of the conversation, and sensitivity on the part of the machine to when *it* is being addressed by human voice. The latter is particularly relevant in an environment in which speech is being used for a variety of purposes, such as audio memos, speaking over a telephone, and alarm functions, in addition to the command channel to control the machine.

Earlier work had demonstrated the utility of dialog based on syntactic analysis as an approach to coping with recognition errors [Schmandt 82], although the parser used was hard coded for the particular application and extensible only in design. The *Phone Slave* [Schmandt 84, Schmandt 85] successfully exploited people's willingness to participate in a computer driven conversation, but it was a fairly passive system which made little use of knowledge of other activities its owner was engaged in. This project attempts to synthesize both of these approaches.

The Environment

This project is based on the concept of an integrated office workstation which combines the functions of a powerful personal computer and an intelligent telecommunications system. In addition to conventional personal computer applications, this workstation is actually an active node on a digital network. It handles its owner's schedule, travel plans, telephone management and message taking, and event-activated audio memoranda or reminders. As will become clear, the more the workstation is cognizant of its owner's activities, the greater its ability to make correct inferences about its own proper behavior in response to stimuli from the outside world.

As a telecommunications node, we base this work on a vision of point-to-point communication including simultaneous voice and data links: the latter need not be high speed. Thus, nodes are able to engage in joint activity requiring localized databases, such as scheduling meetings between the owners of separate workstations or the intelligent handling of the control signals of a voice telephone connection. When one's workstation is instructed to "*place a call to X,*" it first makes a digital connection to X's workstation to determine whether X will take the call, and, if so, at what location (*telephone number*) to connect the voice link. Similarly, a digital connection to a process on another node is used to negotiate meeting times between remote schedulers.

Our implementation of the Conversational Desktop has been on Sun Microsystems workstations using Internet protocol on Ethernet hardware as the data link, and ordinary analog telephone connections as the audio link. With the evolution of digital telephone exchanges and the provision for service protocols such as ISDN, it is reasonable to assume that the voice/data channels will be available in an integrated telephone system in the not too distant future.

Each workstation is equipped with a variety of speech peripherals, including recognition, synthesis, and digital record/playback hardware. The devices actually used are configurable at run-time and the system will run, with reduced capability, with only a subset of them. The main focus of this work, however, is the synergy of these speech technologies, particularly in a context which exploits voice in a range of interrelated and interconnected tasks.

The workstation is designed to be driven entirely by voice, engaging its owner in a conversation interleaved with transactions with remote nodes. The repertoire of available operations at the time of this writing includes: scheduling meetings with individuals or groups, placing outgoing calls, taking incoming voice messages, and recording voice memos related to the above activities. The incoming message taking is based on a conversational *answering machine* described previously as Phone Slave.

Conversational Aspects

The Desktop is inherently conversational; dialog is used both as a steady flow of feedback as well as a means of resolving ambiguities or errors on the part of the speech recognizer.

The output from a speech recognizer tends to be very noisy, characterized by a combination of insertion, substitution, and undetected word errors. It is necessary to build a *robust* parser to scan the output from the recognizer and build up a data structure describing knowledge of the input which can be used to generate dialog. Standard parsing techniques from the natural language processing community [Winograd 83] are generally inadequate, as they are based on the assumption of correct input (usually typed) in the first place.

The solution employed is a context free grammar and parser based on the Unix YACC (Yet Another Compiler-Compiler) parser generator. Each token is an instance of a syntactic class such as 'command-which-requires-a-date-and-time.' The parser takes output from the recognizer and runs all ordered substrings through YACC, calculating a score at each node of the grammar, skipping those which can be pruned in comparison with the current high score. For example, the string ABC would be parsed for ABC, AB-, -BC, A-C, A--, -B-, --C.

The scoring metric reflects knowledge of the types of recognizer error; in connected speech, errors often come in bursts, as the result of incorrect segmentation decisions. It gives points for number of words recognized, with bonuses for complete sentences and adjacent correct words.

The dialog which ensues is an attempt by the machine to fill in gaps in the parse tree based on the parse with the highest score. Phrasing of the questions is critical for several reasons. The dialog employs echoing techniques [Hayes 83] to implicitly confirm prior communication. For example, "*Schedule a meeting with Walter at <mumble>*" would trigger a query of "*When do you wish to meet with Walter?*". These questions are geared toward eliciting single word responses wherever possible, as they are much more likely to be recognized.

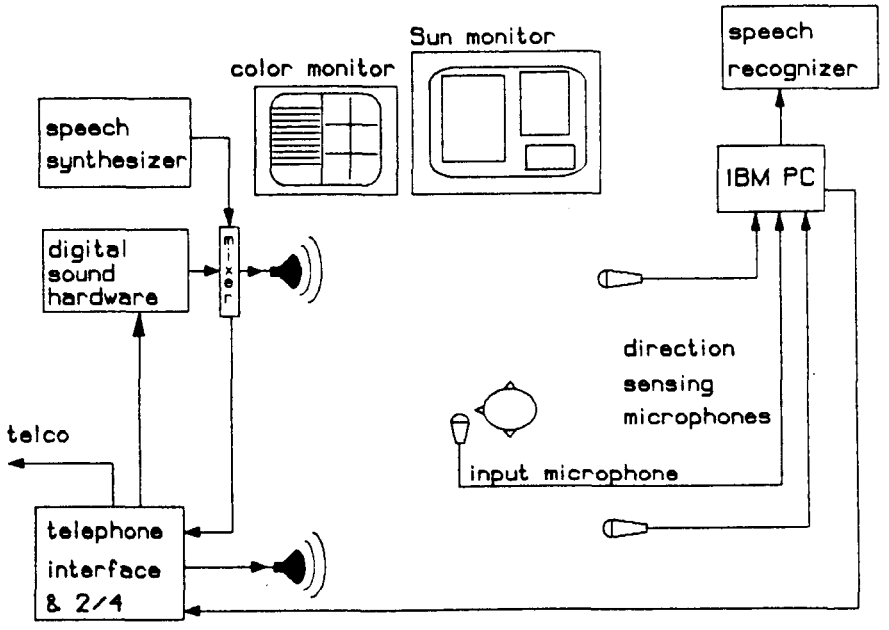
Another aspect of the conversational ability of the system is its method of taking incoming phone messages. Callers are greeted by a recorded voice which asks a series of questions, recording each one, while an adaptive pause detection algorithm triggers the next query. The answers to questions such as "*Who's calling please?*", "*What's this in reference to?*"

and *"At what number can you be reached"* are recorded into individual sound files. This sequence of recordings provides a context or handle on the content of the audio data. Even without recognizing any of the words in the answer to the *"Who's calling please?"* question, the machine knows it is appropriate to play this segment when its owner asks *"Who left messages?"*

Addressability

Through a variety of cues, especially eye contact, a person in a small group can determine when the speaker is addressing them in particular. We wish to apply similar techniques to speech recognition so that the computer can determine when *it* is being addressed, as opposed to the telephone or someone else in the office.

To facilitate this, we have assigned spatial orientation to the system: the computer is assigned the direction of the owner's right, and the telephone the left. The system display, which shows calendar entries and phone message status, along with the loud speakers through which the Desktop talks, are both situated on the right side of the office. The 'telephone' is a hands-free arrangement using the head mounted microphone for input and a speaker for output to the left side of the office.



Audio paths in the Conversational Desktop, showing the spatial orientation of devices.

A pair of microphones placed behind the user determines the direction towards which the person is speaking. The microphone receiving the *minimum* signal when speech is detected in the head-mounted (recognizer) mike is in the opposite direction of the voice addressing. Microphones were placed to the rear to take advantage of the greater direction sensitivity based on the radiational characteristics of the human head [Flanagan 60]. Hardware and software running on an IBM PC communicate this information to the Sun Workstation.

The same hardware also controls noise-free ramped switches which direct audio to various

devices in the room. When the computer speaks, either by playing a recording or through the text-to-speech synthesizer, input to the recognizer is disabled to prevent spurious recognition. Recognition output is parsed only when speech is being transmitted in the direction of the recognizer. While speaking on the phone, the owner may have a private conversation with his Desktop by turning to the right; as soon as speech is detected in this direction, audio input to the telephone connection is temporarily disabled.

Context

The direction-sensing microphones are also used to detect background noise (defined as signal present with no speech on the owner's headmounted microphone) which alerts the system to the presence of other humans in the office. This *background speech present* signal is used for a class of operations characterized by knowledge of the acoustical context of events occurring in the domain of the Desktop system.

When it is time to play an audio reminder, for example, the system first checks this signal and can *postpone* the reminder until a time when the owner is alone in his office. In general, the system follows the rule of not interrupting when the owner is engaged in some detectable activity; future work will attempt to prioritize current and alarm events. It will not, for example, interrupt the owner for a phone call during other activity; instead, it automatically takes a message.

The more the system knows about its owner's activities, the more it can take advantage of context to understand speech input and guide its activities. Several examples relate to telephone conversations. While engaged in a call, the command "*Schedule both of us a meeting*" refers to the owner plus the party on the other end of the connection; the system knows who this is, as it originated the call in the first place.

When the owner tells the Desktop "*I am going out to lunch*", it knows to automatically replace the current outgoing answering machine message with the "out to lunch" one.

In a similar vein, system activities may be triggered by external events. Audio *reminders* can be recorded by a sequence like "*When I talk to Barry remind me to ...*". Although the system then performs no recognition on the body of the reminder, it knows enough *about* it to automatically remind the owner, by playing the speech file, when asked to "*Place a call to Barry.*". The same reminder also gets played as part of the process of accepting an incoming call from Barry. Reminders may be triggered by a telephone connection, a meeting about to occur, or a directive such as "*I am going home.*"

Future Work

Currently we are engaged in expanding the capability of the Desktop in several ways.

The first is added functionality, e.g., placing airline reservations, checking the state of the weather before the owner sets off to bicycle home, etc. There is a wide range of electronic databases which the Desktop can access over digital telephone connections to automatically update local knowledge about the state of the world and potentially alert the owner.

The second is personalization. Several aspects of this are already in place. For example,

each node's scheduler is driven by a set of owner specific preferences: one person may set up preferences to avoid morning meetings, while another node may attempt to avoid any meetings after five in the evening.

A more difficult challenge is to modulate preferences by some sense of the importance of the calling party. For example, I would never meet a student before 10 AM, but maybe would come in early for an important visitor. Similarly, some incoming calls should interrupt many of my activities, but most calls should (in my own preference) never interrupt a conversation with another person in the office.

Acknowledgement

This work has been supported by NTT, the Nippon Telegraph and Telephone Corporation.

References

Flanagan 60 Flanagan, J. L.
Analog Measurements of Sound Radiation from the Mouth.
J. Acoust. Soc. Am. 32(12), 1960.

Hayes 83 Hayes, P. and Reddy, R.
Steps Toward Graceful Interaction in Spoken and Written Man-Machine Communications.
Int'l J. Man-Machine Studies 19:231-284, 1983.

Schmandt 82 Schmandt, C. and Hulteen, E.
The Intelligent Voice Interactive Interface.
In *Human Factors in Computer Systems.* NBS, ACM, 1982.

Schmandt 84 Schmandt, C. and Arons, B.
A Conversational Telephone Messaging System.
IEEE Trans. on Consumer Electr. CE-30(3):xxi-xxiv, 1984.

Schmandt 85 Schmandt, C. and Arons, B.
Phone Slave: A Graphical Telecommunications Interface.
Proc. of the Soc. for Information Display 26(1), 1985.
In Publication.

Winograd 83 Winograd, T.
Language as a Cognitive Process - Syntax.
Addison-Wesley, 1983.

Christopher Schmandt
Principal Research Scientist
Mr. Schmandt received his B.S. in Computer Science and his M.S. in computer graphics from MIT. He has continued his work as a Principal Research Scientist at the Architecture Machine Group, a component of the Media Laboratory. His research interests there are focused on interactive systems and human-interface issues, with emphasis on voice interaction and telecommunications.

Barry Arons
Research Associate
Mr. Arons received his B.S.C.E. and M.S. in computer graphics and interactive systems from MIT. His research interests include speech input output, raster graphics, and interactive video.

Charles Simmons
Undergaduate Researcher
Mr. Simmons is completing his senior year at MIT. The topic of his Bachelor's thesis is the design and implementation of the speech direction sensing hardware and software.

The authors can be contacted at:
Media Laboratory
Massachusetts Institute of Technology
20 Ames Street, Room E15-327
Cambridge, MA 02139