Speech Group, Media Laboratory

The Speech Research Group of the M.I.T. Media Laboratory is concerned with understanding human speech communication and building systems capable of emulating conversational behavior. We are concerned with voice both as data as well as a control channel. We have a strong interest in telecommunication applications as we are convinced that the telephone network can provide enhanced access to online information through voice interfaces. We see voice integrated with advanced workstations as providing an interface to capture and enhance many aspects of office and personal communication.

Current focuses include:

- identifying functional components of human dialog and applying these constraints to graceful conversational systems
- improving understanding of the human interface requirements to enhance the utility of speech, an otherwise relatively intractable medium
- building an integrated desktop audio and telephone environment in Unix workstations under the X window system, including voice messaging, editing, filing, and multi-media documents
- personalized telecommunications management, such as speed dialing from personal databases, intelligent call routing, and remote voice access to computer data
- applying semantic and pragmatic information to language understanding using speech recognition
- employing prosodic (rhythm and duration) cues for speech understanding and generation

Early Projects

PUT THAT THERE [1980][14] was an early conversational system which employed voice and gesture as input and voice and graphics as output. The user sat before a wall-sized display and manipulated a database using speech recognition and a magnetic two dimensional pointing device. To detect speech recognition errors it included a (very ad hoc) syntactic analyzer and parser and re-entrant dialog generator to query the user with an early speech synthesizer. Although in many ways primitive, the multi-media input was innovative and it was one of the first systems which used both speech input as well as generation for dialog.

Some exploratory work LIP SYNC [1982] [11] investigated the possibility of reconstructing a video image of a remote participant in a teleconference. The incoming audio signal was analyzed for features such as formant positions and energy to change lip position on the local display to convey added realism at no additional bandwidth.

VOICED MAIL [1983] allowed lab members to read electronic mail over the telephone and to generate voice replies as well as limited text replies. Parts of this system survived in PHONE SLAVE [1984] [8, 9, 7], an intelligent answering machine with multi-media capabilities. The PHONE SLAVE's conversational prompting to segment the caller's responses into smaller components. This facilitated the recording of complete messages and enhanced access to these messages by the owner. It used speech recognition technology to identify callers by voice in order to have an interaction specific to that person. PHONE SLAVE was accessed locally on a touch screen color display or remotely by telephone. The former included a variety of telephone management tools such as a Rolodex.

The CONVERSATIONAL DESKTOP [1985][10] explored issues in voice as a control and data channel in a more highly networked voice/data environment with a focus on office tasks. It included scheduling, telephone management, voice memos, and external database access employing a variety of input/output channels. A more sophisticated parser[13] attempted to detect a variety of speech recognition errors and the dialog generator used echoing as well as questioning to offer more user feedback. A pair of microphones was used to determine the direction in which one was speaking; this was used as a switch to enable recognition by turning to talk to the workstation.

DIRECTION ASSISTANCE [ongoing][6] is a program which gives driving directions in the Boston area using speech synthesis. A user calls in, specifies starting and ending address, and is given driving directions to the destination. It serves as an environment for exploring issues in user interface design[3], especially as versions of it have been running in public displays at local museums. It also illustrates the utility of prosodic aspects of speech generation for intelligible presentation of complex information; pause duration, pitch range, and the placement of pitch accents give clues to the semantic or intentional structure of the discourse.

GRUNT [1988][12] explored the use of "back channels" as flow control mechanisms in an interaction such as that of Direction Assistance. Back channels are responses that a listener makes (e.g. "uh-huh", "OK", "What was that?") which facilitate the talker's task of presenting the main topic of conversation. GRUNT attempted to employ only prosodic cues (utterance duration, pitch contour, and energy) to determine the discourse function intended by the listener; it deliberately avoided using any word recognition.

Several projects focus on the role of prosody in speech, attempting to generate or quantify it. PITCHTOOL [1987] provides a Sun window system based speech analysis tool for speech recorded by Linear Predictive Coding (LPC) on a dedicated PC server. A series of experiments [1988] attempted to quantify the magnitude of pitch accents as well as other pitch events (vowel intrinsic pitch, consonantal effects, and pitch range) at the perceptual level to determine the outstanding issues in detecting pitch accents.

GENERATIVE INTONATION [1987] translated pitch accents based on the model of Pierrehumbert into pitch contours for speech synthesis. This project generated realistic values for pitch, but also showed the inadequacy of current theories of prosodic rhythm.

Current work

The AFFECT EDITOR [1, 2] attempts to add a measure of emotional expressivity to synthetic speech. In this program, an "affect" is defined as a set of 19 perceptual parameters specifying pitch effects, timing changes, voice quality, and articulation. These parameters in turn are reduced to a smaller set of intonational controls and speech synthesizer settings. When tested with a set of six emotions, all six were recognized by listeners, though there were some confusions.

DESKTOP AUDIO explores voice as both data as well as a user interface medium for a desktop workstation. It is our belief that no single application of voice is powerful enough to make it ubiquitous in computer environments, but that the interaction between a number of voice and telephone utilities will result in powerful new workstation environment. This includes basic architectural issues of supporting voice in a non-real-time operating system (Unix) and providing multi-media selection mechanisms to move data between applications. Applications include a conversational telephone answering machine, speed dialer, calendar, rolodex, audio editor, and multi-media document tools.

XSPEAK is an evaluation of the utility of speech recognition as an input medium for navigation in an overlapping window system. There has been little research into the user interface implications of window systems. What little has been done indicates that overlapping windows may be superior for performing some tasks (those that do not cleanly fit within the geometry of a window) unless navigation between windows interferes with the task. There is strong evidence that performing multiple tasks across several media results in increased "performance" over use of a single medium. There is not much evidence suggesting that voice recognition is useful for replacing the keyboard for an experienced typist.

This suggests the use of speech recognition to augment the mouse for window system navigation. Windows may be "named", and speaking a window's name will cause that window to move to the top of the stacking order and the pointer to be warped within. Thus, a user need not remove his/her hands from the keyboard to interact with a number of clients.

Currently a pilot study is being run on student programmers. We wish to find out whether they will actually use voice while performing their normal daily tasks. If not, we will try to discover why not; are they uncomfortable talking to the computer, is the speech recognition too poor, or is the task we have outlined really not appropriate? If they do use recognition, we seek to establish what sorts of activities and under what situations it was most useful, and, perhaps more important, does it change the way they use their window system (number of windows, layout, degree of overlap).

NETWORK ACCESS TO VOICE SERVICES investigates use of voice for accessing computer databases by voice telephone lines. Some of these databases will be mixed voice and text. Some of the services will involve advanced telephony functions implemented in an ISDN (Integrated Services Digital Network) environment.

The BACK SEAT DRIVER [ongoing][4, 15, 5] uses synthetic speech to provide real-time driving directions in an automobile equipped with a position sensor. Work has concentrated on determining the required form and content of driving directions (what to say, when to say it) and the necessary contents of the underlying map database. The Back Seat Driver is currently running in our research vehicle, an Acura Legend.

References

[1] Janet E. Cahn. Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology, June 1989.

[2] Janet E. Cahn. Generation of affect in synthesized speech. In *Proceedings of 1989 Conference*, pages 251–256. American Voice I/O Society, Sept 1989.

[3] James R. Davis. A voice interface to a direction giving program. Technical Report 2, MIT Media Laboratory Speech Group, Apr 1988. Replaces the paper in the 1986 AVIOS proceedings titled "Giving Directions: A voice interface to an urban navigation program".

[4] James R. Davis. *Back Seat Driver: voice assisted automobile navigation.* PhD thesis, Massachusetts Institute of Technology, September 1989.

[5] James R. Davis and Chris Schmandt. The back seat driver: Real time spoken driving instructions. In *Vehicle Navigation and Information Systems*, pages 146–150, 1989.

[6] James R. Davis and Thomas F. Trobaugh. Direction assistance. Technical Report 1, MIT Media Laboratory Speech Group, Dec 1987.

[7] C. Schmandt. Speech synthesis gives voiced access to an electronic mail system. *Speech Technology*, pages 66–68, August/September 1984.

[8] C. Schmandt and B. Arons. A conversational telephone messaging system. *IEEE Trans. on Consumer Electr.*, CE-30(3):xxi–xxiv, 1984.

[9] C. Schmandt and B. Arons. Phone slave: A graphical telecommunications interface. *Proc. of the Soc. for Information Display*, 26(1):79–82, 1985.

[10] C. Schmandt, B. Arons, and C. Simmons. Voice interaction in an integrated office and telecommunications environment. In *Proceedings of the AVIOS '85 Voice Input/Output Systems Applications. Conf.*, 1985.

[11] C. Schmandt and W. Bender. A programmable virtual vocabulary speech processing peripheral. In *Proceedings Voice Data Entry Systems Applications Conference, American Voice Input/Output Society*, 1983.

[12] Chris Schmandt. Employing voice back channels to facilitate audio document retrieval. In *Proceedings*, pages 213–218. ACM Conference on Office Information Systems, 1988.

[13] Chris Schmandt and Barry Arons. A robust parser and dialog generator for a conversational office system. In *Proceedings of 1986 Conference*, pages 355–365. American Voice I/O Society, 1986.

[14] Christopher Schmandt and Eric Hulteen. The intelligent voice interactive interface. *Human Factors in Computer Systems*, pages 363–366, 1982.

[15] Christopher M. Schmandt and James R. Davis. Synthetic speech for real time direction giving. *IEEE Transactions on Consumer Electronics*, 35(3):649–653, August 1989.