Talking with computers

Vishwa Gupta, Matthew Lennig, Paul Mermelstein, Douglas O'Shaughnessy

Since the computer was first introduced in the 1950s, much effort has gone into easing human communication with it. On the input side, typewriter-like keyboards have generally replaced the old punched cards, and, on the output side, video display units form an alternative to the more conventional paper listings. Today, we stand on the threshold of using the most convenient communication method of all — that of communicating with the computer through ordinary speech, both for input and for output. But just as the VDU did not eliminate our need for paper listings, speech communication will not so much replace as add to the variety of existing input/ouput techniques.

Speech communication technology holds the promise that, one day, the telephone will become a ubiquitous computer input/output device. Information could be entered in response to spoken queries from the computer and retrieved through a structured dialogue that allows the identification of that information. By dialling the database directly, the user of such a service could not only determine, for example, the departure of the next flight from one city to another and whether seats were available, but could reserve a seat and at the same time charge the cost to his credit-card account. Although direct voice access to such information services has not yet been realized, current technology could support simple forms of such systems within the next three years.

Speech recognition and generation technology could also enhance existing telecommunication services. Although text messaging has been very successful where terminals are readily available and people frequently use the computer, voice messaging promises similar benefits without the need for the terminals. The fact that a large percentage of business calls cannot be completed because the called party is unavailable suggests the need for a voice messaging service. A service that would automatically connect the caller to an automated message system when the called number was busy or unanswered could greatly increase customers' satisfaction.

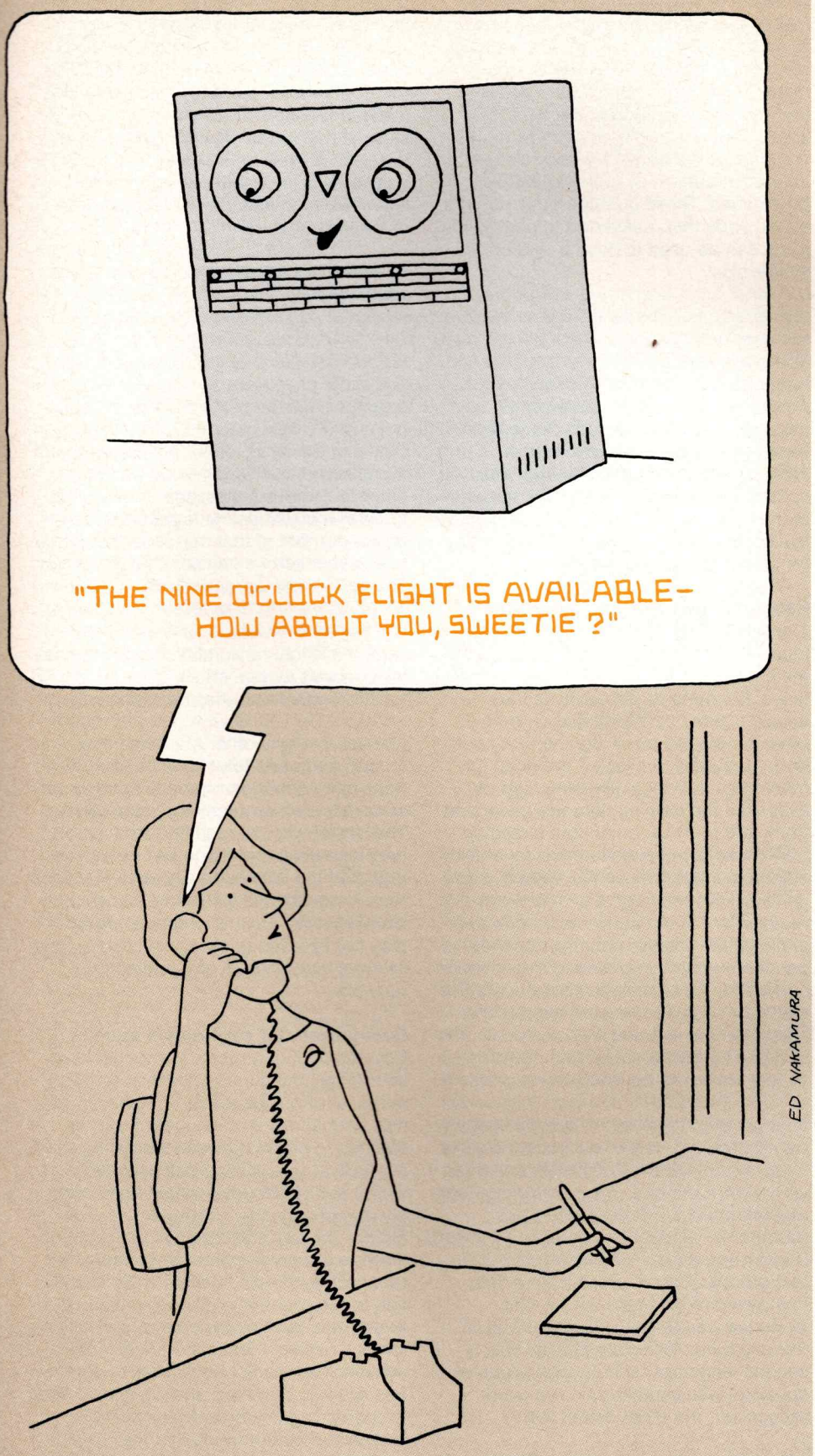
In 1977 Bell-Northern Research formed its own speech communication systems research group. In association with the speech research group of INRS-Telecommunication (University of Quebec), the Bell-Northern Research group examines and reports on the fit between the existing technology and potential applications and identifies technical problems that require further research in order to realize those applications.

The 10-member Bell-Northern Research/INRS-Telecom team is looking at problems in two main areas. The first is the efficient coding of speech for digital transmission. The second is the design of effective man/machine communication systems. The principal problem in the first area is to encode speech in a way that minimizes the required transmission capacity without degrading the speech quality. Problems in the second area include the development of techniques for computer recognition of spoken information, computer generation of speech for human listeners, and the design of communication protocols that integrate these two techniques for specific applications. To bring the appropriate expertise to such a broad variety of subjects, the joint group includes communications theorists, computer scientists, linguists, and psychologists.

In what follows, we will discuss some of the current Bell-Northern Research/INRS-Telecom research into speech synthesis and recognition. Naturally, only the highlights are covered and the reader is referred to the technical literature for more details.

Speech generation and synthesis

The main factor determining how spoken output is generated by machine is the variety of messages that is needed. Where only a few messages are needed, each message can be individually recorded, stored in digital or analog form, and replayed on demand. However, where machine memory (storage) must be less than that needed to store each message as a whole, synthesis of the spoken output from smaller units is required. Uses for spoken output range from maintenance systems with voice warning signals, which only require a few different messages, to the conversion to speech of unlimited text, where the storage of each individual message is impracticable. Where there are



intermediate requirements in message variety, the cost of coding the speech into a compressed form and decoding it on demand must be weighed against the reduced storage costs. A demand that messages be not only intelligible but also natural-sounding can further limit the extent to which the speech signal can be compressed.

One of our current investigations is trying to assess the quality of messages generated using highly compressed forms of individual words. Naturally-spoken words are processed to separate the information concerning the pitch or vibration frequency of the vocal cords from the information specifying individual speech sounds. When playback is required these words are reproduced with pitch and duration adjustments to aid comprehension of the message.

When a person speaks, he modifies the acoustic form of most of the sounds to fit the structure of the particular sentence. The syllables of a multisyllabic word receive varying degrees of stress, which is controlled primarily by the duration of the syllable and its pitch. The pitch pattern of the sentence is further modified according to which word or phrase requires stress to indicate a particular significance. Thus in response to the two queries:
''Is the nine o'clock flight available?''
''Is there an available flight after eight o'clock?''
the same response text can be used, namely:
''The nine o'clock flight is available.''
However, the response to the first query could be made more intelligible by stressing the phrase ''is available'', whereas the response to the second query could be made more intelligible by stressing ''The nine o'clock flight'' — and perhaps especially the word ''nine''. When a word is stressed, not only is its pitch marked by a significant change but its duration increases, especially the duration of the vowel of its stressed syllable. In addition, words that precede a pause, such as the last word of a phrase or sentence, are generally lengthened. Without these modifications synthetic speech appears disconnected and can be difficult to understand.

When words are spoken in rapid succession the first and last sounds of each word may be modified, the modification depending on the sounds of the words on either side. This influence is particularly marked on short function-words such as prepositions, as illustrated by the difference in the pronounced forms of ''to'' in ''to Ottawa'' or ''to Toronto''. Special rules, known as coarticulation rules, can be used to predict the effects of adjacent sounds on each other and to modify the stored forms of such words according to the actual context.

Techniques that generate messages by modifying the stored forms of naturally spoken words currently result in better quality speech than those that generate messages by assembling the constituent speech sounds. Although only about 40 basic sounds or phonemes are used in English, their specific form when spoken is highly varied. Perhaps as many as 200 sound units (allophones) are needed to represent all the important sound variants. In addition, these discrete sounds must be joined to each other by a complex method of overlapping and interpolating that mimics the changes in the human vocal tract as it moves from the shape of one sound to the shape of the next. Although these rules are well enough known to allow intelligible speech to be generated, the naturalness of synthetic speech is limited. Perhaps because of this unnatural quality the understanding of synthetic speech requires much more attention from the listener than does natural speech.

The integrated circuit chips for speech synthesis that are currently available, such as those in the Texas Instruments Speak and Spell device, do not modify the spoken word but generate the speech signal from its compressed form in a way that is inexpensive to implement. The same chips can be used to generate phrases and sentences regardless of whether they are stored as entire units or assembled on demand from smaller units. Such devices generate fairly good-quality speech providing the underlying parameters have been properly specified.

To reveal aspects of the assembled signal that may lead to unnatural results, we compare the form of a naturally spoken message with the same form assembled from discrete units. Because there is no established underlying theory on the naturalness of speech sounds, much of our research is by trial and error guided by insight. We postulate rules for the transformation of individual sounds or groups of sounds into coherent units and evaluate them in a large number of contexts. Problems are identified and used to modify the rules until acceptable results are obtained.

Although the generation of unlimited text of a quality acceptable for use in the public telephone network is not yet feasible, the generation of such text for use in a reading machine for the blind has enjoyed a large degree of success. One machine is already available for English and others will be in production soon; however, no such machine exists for French.

To meet this need, we are designing pronunciation rules for French that will eventually be applied to a machine that reads French text.[1] An optical scanner is used to read a typed or printed page into the machine, and a character recognition program transforms the input into a sequence of coded characters with punctuation marks. A list of pronunciation rules is invoked to convert the characters to speech sounds and lexical stress marks. A simple sentence parser controls the intonation pattern by marking phrasal boundaries. These procedures result in a sound code that, converted to parametric form, can be used to drive a speech synthesizer.

Conversion of the sound code to parametric form requires the use of rules for mimicking the physical constraints of the human vocal system. In human speech production, sounds are constrained by the physical characteristics of the vocal tract — the speed with which it can change shape, for example. The final speech signal is obtained as the output of the synthesizer. Simulation of the entire synthesis system is expected to be ready for demonstration this year.

Recognizing words and sentences

The simplest word recognition systems must be trained by having the user recite the entire system vocabulary one or more times. Recognition capability is thus speaker-dependent and speech from other people will be recognized with much lower accuracy. However, the need for training by individual speakers can be avoided if the training data are generated from, and represent the input expected from, a variety of speakers typical of the population that will use the system. Since universal accessibility is an important requirement in the public telephone system, speech recognizers must be independent of both speaker and transmission conditions — recognizers must locate the start and end of each word despite the presence of line noise and cross-talk. The reduced bandwidth of speech transmitted by telephone is a far less serious problem.

In an attempt to solve the speaker-dependence problem, one of our recent studies examined the effect of different accents on the performance of a word recognition algorithm used in a simulated word recognizer.[2] Spoken words from a 20-word English vocabulary were collected from two groups of speakers; Group A was composed of speakers whose first language was English, Group B was of francophones for whom English was a second language. Half of each group of speakers was used to train the word recognizer, the other half to test it.

Although the word recognizer recognized 97% of the words spoken by Group A speakers, its recognition accuracy dropped to 93% for the same words spoken by Group B. The wider variation among the francophone accents appears to account for the decline in recognition accuracy: the same number of stored exemplars of each word is less able to capture this greater variation among the francophone speakers and consequently leads to higher error rates. Additional study is needed to determine the best way to select the reference patterns for a linguistically nonhomogeneous group of speakers such as the general population of telephone users.

To illustrate how word recognition can be used on the public network, we have designed and simulated the storage and retrieval of voice messages.[3] Any telephone user can dial the message facility and enter or retrieve a message through a series of verbal prompts from the computer and verbal responses from the user. The user is first asked for his identification number and then whether he wishes to enter or retrieve a message. If he wishes to enter a message, he is asked the destination number of that message, and once that is checked he can state the message. The sender has an opportunity to review his message and if necessary to change it. If the user wishes to retrieve a message, he is told how many messages have been stored for him and is given an opportunity to select the ones he wants.

The use of digits other than the initial access code and telephone number to transmit information on the telephone network is known as secondary signalling. The use of word recognition as a secondary signalling technique has the advantage that the telephone system will always remain compatible with voice transmission, whereas any other special signal may not be adequately transmitted by one or more components of the telephone network.

Recognizers for continuous speech

Development of a recognizer capable of accepting continuous speech is an important goal of our research.[4] Such a recognizer would make it unnecessary for a speaker to pause between words. A major problem in recognizing both isolated words and continuous speech is acceptably recognizing the constituent sounds despite the large differences in the way individual speakers pronounce them. In recognizing isolated words, a partial solution to the problem of pronunciation differences is to store multiple reference forms to reflect these differences. The same solution is not applicable to continuous speech, however, and we are currently addressing this problem as part of a research contract funded by the Canadian Department of National Defence. The

approach taken in this project is to determine whether it is possible to estimate the general speech characteristics of an individual speaker and then to use these to improve the accuracy of the recognizer. The recognizer adapts to a particular speaker by modifying its reference data in accordance with its estimates of the new speaker's characteristics. Early results with this technique indicate that the error rate for digits can be reduced by up to a factor of three using estimates derived from as few as three words.[5] We expect this research to lead to substantial improvements in the performance of systems, such as recognizers of continuously spoken digit strings, that recognize continuous speech from a limited vocabulary.

Applications abound for speech communication between man and the computer largely because of the rapid evolution of large scale integrated circuit chips. Techniques for speech synthesis and recognition, which only a few years ago appeared to be of no more than laboratory interest, now represent cost-effective solutions to real communication problems. These developments have in turn fuelled a new thrust to achieve better technical performance on more complex tasks. High quality microprocessor-based continuous speech synthesizers and recognizers are likely to be realized within the next few years. The current challenge is to ensure that these new capabilities are used effectively to enhance man's communication with computers. As yet we have very limited experience in how to exploit this new speech-communication channel. Designing systems to meet that requirement represents a significant research objective for the coming years.

References

1. D. O'Shaughnessy, M. Lennig, P. Mermelstein et al: Simulation d'un lecteur automatique du français. *12 ième Journées d'Etude sur la Parole*, Groupement des Acousticiens de Langue Française, Montréal, Canada, 1981. p 315

2. V. Gupta, P. Mermelstein: Effects of accent on the performance of an isolated word recognizer. *J. Acoust. Soc. Am.* Vol. 68, 1980. p S86

3. P. Mermelstein, V. Gupta: An experimental voice messaging system controlled by word recognition. *International Symposium on Computer Messaging Systems*, April 1981, Ottawa, Canada.

4. M.J. Hunt, M. Lennig, P. Mermelstein: Experiments in syllable-based recognition of continuous speech. *International Conference on Acoustics, Speech and Signal Processing*, Denver, CO, 1980. p 880

5. M.J. Hunt: Speaker adaptation for word-based speech recognition systems. *J. Acoust. Soc. Am.* Vol. 69, 1981. p S41

Vishwa Gupta, a native of India, obtained his doctorate in electrical engineering in 1977 from Clemson University, where he then served as a visiting instructor for a year. In the speech communication research department at Bell-Northern Research, which he joined in 1978, he has worked on the message storage and retrieval system, on isolated word recognition, and on the recognition of digit strings.

Matthew Lennig, a member of scientific staff at Bell-Northern Research, Montreal, also serves as visiting professor at INRS-Telecommunication, University of Quebec, and as a faculty lecturer at McGill University, Montreal. He obtained his undergraduate education at Princeton University and his graduate education at the University of Pennsylvania, from which he received a PhD in linguistics in 1978. He then joined Bell-Northern Research. His current research interests include the study of linguistic variation in colloquial speech, speech recognition, speech synthesis, and natural language understanding.

Paul Mermelstein is manager of speech communications research at Bell-Northern Research, Montreal. He also serves as visiting professor at INRS-Telecommunication, University of Quebec, and as an auxiliary professor at McGill University. A native of Czechoslovakia, he obtained his undergraduate education at McGill University and his graduate education at the Massachusetts Institute of Technology, graduating with a DSc degree in electrical engineering. Between 1964 and 1977 he held research positions in speech communication at Bell Laboratories and Haskins Laboratories. His current research interests focus on applying models of human speech communication to enhance speech communication with machines.

Douglas O'Shaughnessy is a professor at INRS-Telecommunication, University of Quebec, and an auxiliary professor in the electrical engineering department at McGill University, Montreal. A native of New York, he obtained his undergraduate and post-graduate education at the Massachusetts Institute of Technology, graduating with a PhD in electrical engineering and computer science in 1976. After a one-year post-doctoral position at MIT, he joined INRS-Telecommunication. His current research interests include man-machine speech communication and the digital coding of speech signals.