VOICE TECHNOLOGY AND BBN

JARED J. WOLF

BOLT BERANEK AND NEWMAN INC.
CAMBRIDGE, MASSACHUSETTS

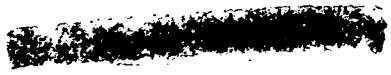
1. PREVIOUS WORK IN VOICE TECHNOLOGY

Bolt Beranek and Newman Inc. has engaged in research, development, and consulting on a broad spectrum of speech-related problems for over two decades. We have done work in at least the following areas:

- speech signal processing
- automatic speech recognition
- continuous speech understanding
- speaker recognition
- speech compression
- subjective and objective evaluation of speech communication systems
- measurement of the intelligibility and quality of speech when degraded by noise or other masking stimuli
- speech synthesis
- instructional aids for second-language learning and for training of the deaf
- investigation of speech correlates of psychological stress

In addition to these speech-related areas, we also work in experimental psychology, control systems, and human factors engineering, which are often relevant to the proper design and operation of speech systems.

The review of BBN's past and present speech-related projects presented below should not be regarded as delimiting our expertise or research interests. Given our role as an R&D and consulting firm,



they represent only specific places where our expertise and interests have intersected with needs of our clients.

1.1 Speech Understanding

BBN was a principal participant in the recent five-year Speech Understanding Research (SUR) project, sponsored by the Advanced Research Projects Agency (ARPA) of the Department of Defense. The objective of the SUR project research was to discover, evaluate, and to incorporate into a total system, techniques for using higher level linguistic constraints and advanced signal processing and acoustic-phonetic analysis to determine the best possible interpretation of an unknown speech utterance. These speech understanding systems were to:

> "... accept continuous speech from many cooperative speakers of the General American dialect, in a quiet room over a good quality microphone, allowing slight tuning of the system per speaker, but requiring only natural adaptation by the user, permitting a slightly selected vocabulary of 1ØØØ words, with a highly artificial syntax and a (well defined) task..... tolerating less than 1Ø% semantic error, in a few times real time (on a 1ØØ Mips machine), and be demonstrable in 1976 with a moderate chance of success."

BBN's speech understanding system, called HWIM (for Hear What I Mean), is a powerful research system for exploring alternative control strategies and the effects of different system features. We have used this system to develop some powerful speech understanding algorithms. System components include:

a) A linear predictive coding signal analysis component, which derives smooth spectral parameters, formant and pitch tracks, and other parametric information from the input speech waveform,

b) An acoustic-phonetic recognition component, which segments the acoustic input into a lattice of alternative possible phonetic labelings of the input,

c) An off-line dictionary generation component, which uses within-word and between-word phonological rules to produce word pronunciations expected to be encountered in fluent continuous speech,

d) A fast lexical retrieval component, which can efficiently find words in the vocabulary that match well acoustically with the speech input and which accounts for context-dependent across-word phonological effects,

e) An analysis-by-synthesis word verification component, which can synthesize the expected parametric representation of a hypothesized word (and its context) and compare it with the input parameters,

f) A grammar for interactions with a travel budget management system in natural English using a vocabulary of over 1ØØØ words,

g) A bi-directional parser for ATN grammars, which can parse a sentence from left-to-right, right-to-left, or middle-out,

h) A semantic network knowledge base, which contains general knowledge about trips and places, as well as specific information about planned trips, estimated costs, budgets, expenditures, etc., and

i) A flexible control component, which uses the other components to formulate, evaluate, and extend hypotheses into a complete interpretation of the sentence.

HWIM's speech understanding is set in the context of a travel budget manager's automated assistant, which keeps track of trips taken and planned and the budgets to which trip costs are charged, and it also allows the user to plan new trips. Users may interact with HWIM by speaking sentences from a rather general grammar (over 1ØØØ words, with a high average branching ratio and rejoining paths) forming a subset of natural English. Typical sentences from this task are:

> How much is left in the speech understanding budget?
> List all trips to California this year.
> What is the round-trip fare to Chicago?
> Cancel Jerry's trip to the ASA meeting.

At the end of the SUR project in October 1976, HWIM correctly understood about half of its test utterances, spoken by three speakers. (1,4,7-13,16,18,19,23-29)

Continuous speech understanding systems with the capabilities of HWIM and the other ARPA SUR project systems are not yet ready for immediate application, but that was not the goal of the ARPA SUR project. That goal was the development of an advanced technology of speech recognition and understanding. The technology developed during the ARPA SUR project has clear utility in speech recognition and understanding applications that should be practical in the immediate future.

1.2 Speech Bandwidth Compression

BBN has been doing research in the speech compression area since 1972, with support from ARPA, and more recently from other sponsors also. BBN has been and is currently involved in developing speech compression systems with a wide range of transmission bit rates, ranging from 75 to 16ØØØ bits/sec, and with different operating conditions such as noisy or high-quality input speech, noisy or noise-free transmission channel, and fixed-rate (synchronous) or variable-rate (asynchronous) transmission. (2,9-13,16,21,22)

The overall goal of the ARPA speech compression research has been to develop linear predictive speech compression (LPC) systems that transmit good quality speech at low data rates. Speech compression techniques developed in this project have been designed for their use in the ARPA Network environment of packet-switched data communications, though they are easily extendible to other communications environments.

Recently developed techniques in linear prediction are used for the analysis and synthesis. We have developed several methods for reducing the redundancy in the speech signal without sacrificing speech quality. Included among these methods are preemphasis of the incoming speech signal, adaptive optimal selection of predictor order, optimal selection and quantization of transmission parameters, variable frame rate transmission, optimal encoding, and improved synthesis methodology. When we incorporated all of these in a floating point simulation of a pitch-excited linear predictive vocoder, we obtained synthesized speech with high quality at average transmission rates as low as 15ØØ bits/sec (21,22). Our more recent results include: development of a new class of stable linear predictive speech analysis methods (12); specifications for an asynchronous or variable data rate linear predictive speech compression system to be implemented by the various ARPA-sponsored sites for real-time speech transmission over the ARPA Network; application of nonlinear spectral warping techniques to either improve speech quality at a given bit rate, or to lower the transmission bit rate at a given speech quality.

One of the major results of the ARPA speech compression project has been to demonstrate real-time speech transmission on the packet-switched ARPA network. BBN participated in the implementation of the SPS-41-based initial system. More recently, a real-time system specified by BBN, transmitting at an average rate of 22ØØ bits/sec, has been implemented on a Floating Point Systems AP-12ØB at Information Sciences Institute. The system will be implemented at BBN on the AP-12ØB we are about to receive.

Our work on speech compression also includes the development of objective procedures for testing the quality of vocoded (or compressed)

speech (15,2Ø). Since the objective procedures must be validated against results from subjective listening tests, we also have a program for the subjective evaluation of speech quality. We have explored the perceptual dimensions of speech quality by multidimensional scaling methods (2).

1.3 Very-Low Rate Vocoder

An interesting outgrowth of our work in speech understanding, speech compression, and speech synthesis was a project combining phonetic speech transmission system operating at 75 bits per second (14). Based on this pilot project, we have proposed a real time implementation for such a system.

1.4 Speech Synthesis by Rule

Our experience in speech synthesis is derived mainly from the research in synthesis-by-rule being carried out by Dennis Klatt at MIT and at BBN (6,7). In our speech understanding system, synthesis played two roles, as a voice response component and as a component of an acoustic-phonetic word verifier, in which a hypothesized word (plus context, if any) was synthesized into an idealized time-varying spectral representation that was then compared against the analyzed utterance itself. In this way, generative acoustic-phonetic knowledge was used to evaluate how well a hypothesized word matched a portion of the utterance (1,4,5). In the phonetic speech transmission system, the receiver used a modification of the synthesis-by-rule program to resynthesize speech from the transmitted values of phoneme identify, duration, and fundamental frequency (14).

1.5 Instructional Aids Systems

The instructional aids systems are self-contained computer-based systems for real-time speech analysis and display. A minicomputer receives information about speech-related waveforms via microphones and accelerometers connected to analog and digital preprocessing circuits. Algorithms for analysis and display operate on the data, sometimes under the control of the user, in such a way as to provide concurrent visual and auditory representations of speech sound that may be useful to the user in the modification of his articulation.

The second-language training system is designed to supplement the standard language laboratory. It allows a student to visually compare his efforts with pre-recorded teacher's versions. This system has been evaluated in the context of two language pairs: English speakers learning Chinese and Spanish speakers learning English (3).

The deaf-training system involves a trained teacher working with the student, with the system operating as a tool to enhance their interaction. In this case, attempts have been made to develop displays that are appropriate for use with very young children with severe language limitations as well as profound hearing losses. The prototype system is now being tested at two schools for the deaf (17).

1.6 Other Projects

Other projects dealing with voice technology include:

- adapting our variable frame rate speech compression approach to fixed rate transmission operating at 24ØØ bits/sec over a noisy transmission channel,

- ultra-high quality analysis/synthesis of telephone quality speech at 16ØØØ bits/second or less, where the resynthesized speech must be equal in quality to the original input, and

- an investigation of how the psychological state of the user may be reflected in his speech characteristics.

2. PRESENT PROJECTS IN VOICE TECHNOLOGY

With one exception, our current research projects in speech processing are continuations of some of the projects described above.

Our work in low rate speech compression continues in the direction of improving the quality of vocoded speech without sacrificing low data transmission rates. Presently under advanced testing is an improved voice source model incorporating both periodic and noise components, which largely eliminates the "buzziness" often associated with vocoded speech. We will also be bringing into real-time vocoder implementation many of the quality improvement techniques already demonstrated in our floating point vocoder simulations. We also expect to be starting work on high-quality speech synthesis of the type required for a very-low-rate phonetic vocoder system.

Also continuing are the projects on:

- variable-to-fixed rate transmission over a noisy channel
- ultra-high quality analysis/synthesis at a 16 kbit rate
- vocal indicators of the speaker's psychological state.

One new project, not mentioned above, is to develop a processing system to improve the intelligibility of speech that has been corrupted by wideband noise.

3. ANTICIPATED CAPABILITIES IN VOICE TECHNOLOGY

3.1 Staff

With its experience in a wide variety of projects dealing with voice processing, BBN numbers among its staff many with training and experience in the field. In 1977, 11 full-time scientists and 3 regular consultants are engaged in voice technology research and development, almost all of these with advanced degrees. We expect to maintain at least this level of staffing in the foreseeable future. BBN's Information Sciences Division, within which our speech projects are based, numbers over 1ØØ scientists from a broad variety of fields, particularly computer science, artificial intelligence, computational linguistics, electrical engineering, and the behavioral sciences.

3.2 Facilities

The BBN Research Computer Center (RCC) has four DEC PDP-1Ø's and one DECsystem-20. Three of the PDP-1Ø's run TENEX, a virtual memory time sharing system developed by BBN. The other PDP-1Ø and the DECsystem-2Ø run TOPS-2Ø, a DEC supported time sharing system based on TENEX. Much of the speech processing work not requiring real-time processing is carried out on the KL 1Ø/9ØT system which runs TOPS-2Ø. All of the program libraries used in the speech and signal processing are runnable on both TENEX and TOPS-2Ø.

BBN's Speech Processing Laboratory contains equipment for speech signal acquisition, display, editing, storage, and playback, and it provides a facility for advanced real-time speech processing systems research and development. It currently includes a DEC PDP-11/4Ø, a Signal Processing Systems Inc. SPS-41 signal processor (including dual A/D and D/A converters), and an Imlac PDS-1 graphics display processor. Delivery of a Floating Point Systems Inc. AP-12ØB array processor is scheduled for the beginning of calendar 1978; this addition will substantially enhance our real-time processing capabilities. The PDP-11/system is connected to the ARPANET, which is used for data and program transfers to and from the RCC or any other site on the ARPANET, and for packet speech experiments for our continuing speech compression projects. The Laboratory also contains audio equipment for producing, manipulating, and recording audio signals.

4. REFERENCES

(1) Cook, C., and R. Schwartz (1977), "Advanced Acoustic Techniques in Automatic Speech Understanding," IEEE International Conference on Acoustics, Speech and Signal Processing, Hartford, CT, 1977, pp. 663-666.

(2) Huggins, A.W.F., R. Viswanathan, and J. Makhoul (1977) "Quality Ratings of LPC Vocoders: Effects of Number of Poles, Quantization, and Frame Rate," Proc. 1977 IEEE International Conf. on Acoustics, Speech and Signal Processing, Hartford, CT, 1977, pp. 413-416.

(3) Kalikow, D.N., and J.A. Swets (1972) "Experiments with Computer-Controlled Displays in Second-Language Learning," IEEE Trans. Audio Electroacoust., AU-2Ø, 23-27.

(4) Klatt, D.H. (1975) "Word Verification in a Speech Understanding System," in D.R. Reddy (ed.), Speech Recognition: Invited Papers Presented at the IEEE Symposium, Academic Press, 321-341 (see also Bolt Beranek and Newman Inc., Report No. 3Ø82, Cambridge, MA).

(5) Klatt, D.H. (1975) "The Design of a Machine for Speech Understanding," in Speech Communication, Vol. 3, G. Fant (ed.), Halsted Press, pp. 277-289.

(6) Klatt, D.H. (1976) "Structure of a Phonological Rule Component for a Synthesis-by-Rule Program," IEEE Trans. Acoustics, Speech, and Signal Processing ASSP-24, pp. 391-398.

(7) Klatt, D.H., C.C. Cook and W.A. Woods (1975) "PCOMPILER -- A Language for Stating Phonological and Phonetic Rules," Bolt Beranek and Newman Inc., Report No. 3Ø8Ø, Cambridge, MA, pp. 18-23.

(8) Klatt, D.H. and K.N. Stevens (1973) "On the Automatic Recognition of Continuous Speech: Implications of a Spectrogram-Reading Experiment," IEEE Tran. on Audio and Electroacoustics AU-21, 21Ø-217.

(9) Makhoul, J. (1973) "Spectral Analysis of Speech by Linear Prediction," IEEE Trans. Audio and Electroacoustics, AU-21, 3, pp. 14Ø-148, June 1973.

(10) Makhoul, J. (1974) "Linear Prediction vs. Analysis-by-Synthesis," Speech Communication Seminar, Stockholm, Sweden, Vol. 1, pp. 35-43, Aug. 1974.

(11) Makhoul, J. (1975) "Linear Prediction in Automatic Speech Recognition," in Speech Recognition: invited papers presented at the IEEE Symposium, D.R. Reddy (ed.), New York: Academic Press, pp. 183-22Ø, 1975.

(12) Makhoul, J. (1975) "Linear Prediction: A Tutorial Review," invited paper, IEEE Proceedings, special issure on Digital Signal Processing Vol. 63, No. 4, pp. 561-58Ø, April 1975.

(13) Makhoul, J. (1975) "Spectral Linear Prediction: Properties and Applications," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, No. 3, pp. 283-296, June 1975.

(14) Makhoul, J., R. Schwartz, C. Cook, and D. Klatt (1977) "A Feasibility Study of Very Low Rate Speech Compression Systems," BBN Report No. 35Ø8, Bolt Beranek and Newman Inc., Cambridge, Mass., February 1977.

(15) Makhoul, J., R. Viswanathan and W. Russell (1976) "A Framework for the Objective Evaluation of Vocoder Speech Quality," IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, pp. 1Ø3-1Ø6, April 1976.

(16) Makhoul, J., and J. Wolf (1972) "Linear Prediction and the Spectral Analysis of Speech," Report No. 23Ø4, Bolt Beranek and Newman Inc., Cambridge, Mass., Aug. 1972.

(17) Nickerson, R.S., D.N. Kalikow, and K.N. Stevens (1974) "A Computer-based System of Speech-Training Aids for the Deaf," BBN Report No. 29Ø1. Abbreviated version published in AFIPS Conference Proceedings, 43, pp. 125-126.

(18) Schwartz, R., and J. Makhoul (1974) "Where the Phonemes Are: Dealing with Ambiguity in Acoustic-Phonetic Recognition," IEEE Symposium on Speech Recognition, Contributed Papers, Carnegie-Mellon Univ., Pittsburgh, PA, pp. 85-88, April 1974. Also in IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, No. 1, pp. 5Ø-53, Feb. 1975.

(19) Schwartz, R. and V. Zue (1976) "Acoustic-Phonetic Recognition in BBN SPEECHLIS," IEEE International Conference on Acoustics, Speech and Signal Processing, April 12-14, 1976, Philadelphia, 1976, pp. 21-24.

(2Ø) Viswanathan, R., J. Makhoul and W. Russell (1976) "Towards Perceptually Consistent Measures of Spectral Distance," IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, pp. 485-488, April 1976.

(21) Viswanathan, R., J. Makhoul, and R. Wicke (1977) "The Application of a Functional Perceptual Model of Speech to Variable-rate LPC Systems," Proc. 1977 International Conference on Acoustics, Speech and Signal Processing, Hartford, CT, 1977.

(22) Viswanathan, R., and J. Makhoul (1975) "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans.

Acoustics, Speech and Signal Processing, Vol. ASSP-23, No. 3, pp. 3Ø9-321, June 1975.

(23) Wolf, J.J. (1976) "Knowledge, Hypotheses, and Control in the HWIM Speech Understanding System," Conf. Record, 1976 Joint Workshop on Pattern Recognition and Artificial Intelligence (IEEE Catalog No. 76CH1169-2C), Hyannis, MA, June 1-3, pp. 113-125.

(24) Wolf, J.J. (1976) "Speech Recognition and Understanding," in K.S. Fu (ed.), Digital Pattern Recognition, Springer-Verlag, Berlin, Heidelberg, New York.

(25) Wolf, J.J. (1977) "HWIM, A Natural Language Speech Understander," Conference Record, 1977 IEEE Conference on Decision and Control, New Orleans, La., 7-9 December 1977.

(26) Wolf, J.J. and W.A. Woods (1977) "The HWIM Speech Understanding System," Conference Record, IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, Ct, pp. 784-787, May 1977.

(27) Woods, W.A. (1977) "Shortfall and Density Scoring Strategies for Speech Understanding Control," 1977 Int'l Joint Conference on Artificial Intelligence, MIT, Cambridge, Aug. 22-25, 1977.

(28) Woods, W.A., et al. (1974) "Natural Communication with Computers: Speech Understanding Research BBN," BBN Report No. 2976, Vol. I, Bolt Beranek and Newman Inc., Cambridge, Mass., Dec. 1974.

(29) Woods, W.A., et al. (1976) "Speech Understanding Systems: Final Report," BBN Report No. 3438, Vols. I-V, Bolt Beranek and Newman Inc., Cambridge, Mass., December 1976.

BIOGRAPHICAL SKETCH

Jared J. Wolf

B.E.E. (summa cum laude), Union College, Schenectady, N.Y., 1965; S.M., Ph.D. (Electrical Engineering), Massachusetts Institute of Technology, 1967, 1969.

As a graduate student, staff member, and Research Associate in the Speech Communications group at M.I.T., Dr. Wolf did research on techniques of speech analysis and speaker recognition. He was a Leverhulme Postdoctoral Fellow at the Department of Psychology, University of Edinburgh, Scotland, from 1970 to 1971. Since 1971, he has

been a Senior Scientist at Bolt Beranek and Newman Inc., Cambridge, MA, where he has primarily been concerned with the development of speech understanding systems, particularly in the areas of signal processing, phonological rules, and control strategy.

Dr. Wolf is the author of many papers and reports dealing with signal processing, speech analysis, speaker recognition, and speech recognition and understanding.