VOICE RESPONSE SYSTEMS: TECHNOLOGIES AND APPLICATIONS

Julia B. Hirschberg, Stephen A. Riederer, James E. Rowley, and Ann K. Syrdal

Julia Hirschberg, ***Stephen A. Riederer***, ***James E. Rowley***, *and* ***Ann K. Syrdal*** *are members of technical staff at AT&T Bell Laboratories. Ms. Hirschberg and Mr. Rowley are in the Linguistics Research Department, Murray Hill, New Jersey; Mr. Riederer is in the Voice Transaction Systems Department, Columbus, Ohio; and Ms. Syrdal is in the Advanced Services Technology Department, Indian Hill Park, Naperville, Illinois. Ms. Hirschberg works on assigning intonational features in synthetic speech according to an analysis of the discourse characteristics of the text. She joined AT&T in 1985 with a Ph.D. in computer science from the University of Pennsylvania, Philadelphia. Mr. Riederer is responsible for systems engineering for current and future Conversant® speech processor products. He joined AT&T* (continued on page 51)

Voice response systems use recorded and synthesized speech for machine-to-human communication. This paper both describes current voice response systems, and presents our view of the ideal system. We discuss the current state of stored voice technology and text-to-speech synthesis in the context of their converging on the ideal voice response system. Such a system would be intelligible, simulate different voices, have an unrestricted vocabulary, be easy to use, and be compatible with various types of hardware. Although this ideal system does not exist today, it represents a goal that can be achieved building on current technologies in the voice processing field.

Introduction

The ideal automated voice response system should deliver messages that are as effective as those spoken by a human attendant, and do so at a fraction of the cost. Although the ideal voice response system is unattainable with current technology, this paper will identify dimensions that can be used to evaluate current systems and the goals that systems designers are pursuing.

The ideal system will have the following major features:

- It will deliver highly intelligible speech in a pleasant, natural voice.
- It will be able to simulate a variety of different voices, and match the voice quality of a given human speaker.
- It will have an unrestricted vocabulary, and will even "speak" foreign languages.
- System messages will be easy to create and change.
- It will be small, inexpensive, and easy to use without prior training.
- It will be *portable*, i.e., compatible with different types of hardware, and will have multichannel capabilities.

Although such a system does not exist today, it offers a goal for current technologies. This article describes current voice response systems, including those that use both stored voice technologies and text-to-speech synthesis. These technologies, though still short of the ideal,

are being used in many diverse applications, some of which are described below.

Panel 1. Terms and Acronyms in This Paper

APCM	adaptive PCM
ADPCM	adaptive differential PCM
ASCII	American Standard Code for Information Interchange
AUDIX	Audio Information Exchange (AT&T)
CCITT	International Telegraph and Telephone Consultative Committee
CELP	code-excited linear prediction
DSP	digital signal processor
LPC	linear predictive coding
MPLPC	multipulse LPC
PCM	pulse code modulation
SAM	speech-activated manipulator
SBC	subband coding
TDD	telecommunications device for the deaf
TTS	text-to-speech

Stored Voice Technology

Stored voice response systems rely on coding, storing, and reconstructing speech. In coding, the analog voice signal is sampled and digitized, typically at 64 kb/s (kilobits per second). Later computations may be used to compress the digital signal. The digitized speech can be stored until needed, then reconstructed through decompression and digital-to-analog conversion.

The basic challenge for stored voice systems is to lower the speech coding rate (i.e., the number of kb/s), while keeping the quality of the synthesized speech high. Another challenge is to design and build practical, cost-effective systems that meet the real-time demands of many simultaneous users. Central to both efforts is the need for high voice quality. In commercial voice response systems, this means the stored speech sounds intelligible, natural, and almost noise-free. In some applications, the listener must be able to recognize the speaker.

The 64-kb/s pulse code modulation (PCM) and 32-kb/s adaptive differential PCM (ADPCM) standards—both widely used in digital telecommunications, are the current International Telegraph and Telephone Consultative Committee (CCITT) standards for high-bit-rate, high-quality coded speech.[1–3] These are waveform coding techniques, which aim to capture and reproduce the original speech waveform as faithfully as possible to preserve the intelligibility, naturalness, and recognizability of the original speech, and provide good noise control. At lower rates, quality is inadequate for general telecommunications use. However, high bit rates translate into demands for memory stores too large—and bandwidth too high—for many applications.

The 32-kb/s ADPCM rate is the upper limit of bit rates for practical stored voice, and provides a good tradeoff between quality and system resources. However, many applications cannot afford even a 32-kb/s rate, especially those applications that demand large amounts of storage (e.g., voice mail), many channels (e.g., transaction processing), or both (e.g., audiotext). Researchers have developed techniques for coding at 24 and 16 kb/s that produce quality high enough for some applications.[2] For example, both rates are available in AT&T's voice mail systems, AUDIX (audio information exchange) and AUDIX Voice Power.[4,5] Though no national or international standards exist for coding at these rates, AT&T has developed and adopted its own internal standards.[6]

Voice coding (*vocoding*) techniques, used to obtain lower bit rates, are based on the observation that human hearing is not equally sensitive to noise and distortion at all frequencies. Thus, some information in human speech can be removed without losing perceived quality. Vocoding techniques are considerably more complex and computationally intensive than waveform coding. Nevertheless, they have been implemented on single digital signal processors (DSPs), and provide real-time coding. Subband coding (SBC) techniques use a filter bank to separate the speech signal into subbands

that are coded separately using adaptive PCM (APCM). The bands least sensitive to quantizing distortions have fewer bits.

Linear predictive coding (LPC) techniques use a model of the human vocal tract as a linear filter excited by pitch pulses (which occur periodically to act like vocal cord vibrations) or white noise (i.e., noise with all frequencies represented equally and randomly intermixed). A recent advance—multipulse LPC (MPLPC)—substitutes a sequence of pulses whose amplitudes and locations are computed to minimize the perceptual difference between the original and coded speech signals. Another recent advance is code-excited linear prediction (CELP), where the excitation pulses are chosen from a "codebook" of white noise sequences, again to minimize the perceptual distance between original and result.

At present, 16 kb/s is the lowest coding rate widely used for commercial stored voice applications. Improved voice coding techniques, discussed in the "Coding of Speech and Wideband Audio" paper in this issue,[2] hold great promise for producing quality high enough for commercial applications, even at rates as low as 4.8 kb/s. Thus far, however, voice quality is insufficiently high to satisfy service providers, and these coding methods are slower and more computationally expensive than those now used for higher coding rates. Because these coding rates remain at 16 kb/s and above, designers and developers of stored voice systems face significant challenges in designing circuit boards and networks that can reliably handle simultaneous access from many channels to huge volumes of stored speech. This involves designing elements at every level with the highest bandwidth that is economically feasible, devising schemes for intelligently blocking stored speech files, caching speech fragments that may be used again, and sharing resources among channels. Other papers in this issue—Fischell et al., Verma et al., and Berkley and Flanagan—discuss some commercially feasible applications, their system requirements, and architecture choices made in designing and developing them.

Administrative and operational concerns—and technical issues involved in reducing coding rates—also affect the practicality of stored voice systems. For example, in applications that deal with variable numeric amounts—e.g., account, catalog or order numbers; dates and times; account balances; or other dollar amounts—it is clearly impossible to record all possible messages in advance. In such cases, messages may be constructed as needed by concatenating smaller words and phrases. For example, to report a checking account balance of $137.61, the following sequence could be concatenated from recorded segments: "checking" "account balance is" "one" "hundred" "thirty" "seven" "dollars" "and" "sixty" "one" "cents." This strategy is further complicated by having to keep several versions of some phrases, particularly numbers, to achieve a natural sounding message.

In the account balance example, the two instances of "one" should have different inflections: the first rising, the second falling. Without this inflectional difference, the message would not sound natural and would be difficult to comprehend. Not only must more phrases be recorded and stored to achieve natural sound, but messages also must be parsed to determine the variant that should be selected for a given context.

In addition, voluminous or frequently changing information often is difficult to manage with stored voice. The costs of eliciting high-quality and error-free recordings can be prohibitive, and the storage and retrieval dynamics are highly complex. Finding the "right voice" for an application requires work to ensure that the voice conveys the desired overtones, such as confidence, reliability, and trust. Some coded voices sound better than others, and it can take much effort to get a good "take" for a particular phrase. And with stored voice, the way a phrase is recorded determines how it will sound on playback. There is little dynamic control over factors like the length of pauses and speaking rate. All these factors limit the flexibility and choices available with stored voice response systems. At one level, they are natural con-

straints to be engineered. But at another level, the hope of removing these limits provides added energy to the drive to improve text-to-speech synthesis.

Text-to-Speech Synthesis

Text-to-speech systems take arbitrary text as input (together with optional user-specified commands to vary system parameters such as phrasing, rate, and prominences) and produce real-time synthetic speech. The basic steps in the conversion of text to speech are discussed below. They include:

- *Text normalization*, where input text is filtered to identify phrase and sentence boundaries, expand conventional abbreviations, and translate nonalphabetic characters (e.g., $5 to "five dollars")
- *Syntactic analysis*, where each word is labeled for part-of-speech (i.e., noun, verb, preposition, etc.)
- *Letter-to-phoneme conversion* and *lexical stress assignment*, where the input string is mapped to a suitable string of *phonemes* (minimal sound units of a language) with appropriate stress markings
- *Prosodic assignment*, where timing and pitch for the utterance are determined and associated with the phoneme string
- *Synthesis*, where the analyzed input text is realized in speech.

Text Normalization. Many abbreviations, acronyms, nonalphabetic characters, punctuation, and digit strings are encountered in text. Periods in the text may show either an abbreviation or end of a sentence, and so must be made unambiguous.

Many common abbreviations potentially have multiple meanings. For example, *St.* is an abbreviation for both *Saint* and *Street*. Common punctuation characters and numbers may be expanded differently, depending on context: *1/2*, for example, may be expanded as *one-half* in some contexts, and as *January second* in others. All departures from full-word equivalents are problematic for text-to-speech systems. In addition, ASCII (American Standard Code for Information Interchange) text frequently contains embedded escape sequences and other nonalphabetic characters that are not intended to be part of the textual message.

Existing text-to-speech systems have dealt with the problems of text normalization in several ways. Some systems simply spell out all words that have nonalphabetic characters. Other systems rely on the host computer to preprocess the material used in its application, so text normalization decisions are handled outside the text-to-speech process. Still other systems provide specialized normalization modes for different types of user specified text. With this last approach, taken by the AT&T text-to-speech (TTS) system, the user may specify the type of text (e.g., address, date), and the TTS system will adjust its normalization procedures accordingly.

Word Pronunciation. Correct pronunciation of words is critical to a text-to-speech system's intelligibility and acceptability. To pronounce a word correctly, the system must map a string of letters from text to a string of phonemes, symbols that represent the smallest contrastive units of sound in a language. The English language presents particular problems in this respect because pronunciation has evolved considerably over time, while spelling has been more stable, and because there are many borrowings from other languages, such as French, Latin, and Greek. As a result, the way words are spelled in English is frequently related in a less-than-straightforward way to pronunciation. These difficulties are exacerbated for the pronunciation of American surnames, which represent an even wider variety of languages of origin.

Many English words exhibit regular patterns of pronunciation and stress assignment. However, dictionaries are needed for words that do not. But it is impossible to list all the words that can occur in English text, particularly when proper nouns are considered. So, most text-to-speech systems use both a dictionary and letter-to-sound rules to map from spellings to pronunciations. There have been several attempts to increase coverage with minimal dictionary entries.

Morphemic decomposition strategies involve breaking down words into smaller meaning units, such as decomposing *unhelpful* into a stem, *help*, prefix *un*, and suffix *ful*. Such decomposition permits words to be stored as stems, and allows unfamiliar words to be pronounced in terms of already known components. Pronunciation by analogy can also increase the coverage of a limited number of dictionary entries. In this approach, the pronunciation of a novel word is predicted by one or more similar words whose pronunciations are known. Coker and Church[7] describe how morphological and rhyming analogy approaches to dictionary-based word pronunciation in the AT&T TTS system have blurred the distinction between traditional letter-to-sound rules and dictionary-based methods. The rhyme analogy method, for example, uses dictionary look-up to determine the pronunciation of the rhyming portion of the word and letter-to-sound rules to pronounce the novel beginning of the word.

This hybrid approach to word pronunciation provides greater coverage than traditional methods, and is far more reliable than traditional letter-to-sound rules. For example, pronouncing surnames is a difficult but important task for any text-to-speech system. While a vocabulary of under 150 words will cover about half the ordinary words in text, a vocabulary of more than 23,000 entries is needed to cover the same percentage of surnames. For this task, the hybrid approach used by the AT&T TTS system provides acceptable pronunciations for over 98% of American surnames.

Syntactic Analysis. Syntactic analysis provides information necessary for word pronunciation and higher-level *prosody*, i.e., the intonational phrasing and variation in prominence that human speakers can use, for example, to make a declarative sentence sound like a question, or to make one word in a sentence sound more important than another. Many words in English are pronounced differently depending on their part of speech. For example, words like *object* and *desert* are stressed differently depending on whether they function as nouns or verbs. However, determining a word's part of speech in a given sentence can be difficult to do automatically, particularly when several words in the sentence may have different part-of-speech assignments, as in *Time flies*.

Part-of-speech information is also used in most text-to-speech systems to make phrasing and accent decisions. In the AT&T TTS system, like most others, words are divided into *function words* (e.g., prepositions and articles) and *content words* (e.g., nouns and verbs) to determine pitch accent and local phrase boundaries. In general, content words are accented, or made intonationally prominent, while function words are deaccented. In a sentence like *the CAT LIKES to EAT MICE*, for example, the content words (upper case) will be synthesized with greater prominence than the function words (lower case). The distinction between function and content words is also generally used to determine local intonational phrasing, thus grouping words in a sentence into prosodic units to make the phrases sound more natural.

Structural information about larger units of the text can also be used to improve synthesis. For example, identifying nominal compounds in a text (e.g., *city hall parking lot*) permits acceptable pronunciation of these phrases.[8] While fuller syntactic information has been used in experimental interfaces to text-to-speech systems,[9] and in message-to-speech systems that produce synthetic speech from an abstract representation[10,11] to produce more natural-sounding phrasing, a broad coverage parser operating in real time has yet to be incorporated into any functioning text-to-speech system.

Prosodic Assignments. Determining appropriate prosody for a synthetic utterance involves, at the least, assigning appropriate timing for the words to be synthesized, and aligning them with an appropriate intonational contour. So, for example, words at the end of phrases should be longer than those at the beginning. And sentences analyzed as questions should *sound* like questions. The duration of the phonemes that make up each word is assigned by rule, taking into account factors such as the identity and context of the phoneme to be

synthesized; the stress pattern of the word to be spoken; and the location of the phoneme within the word, phrase, and sentence. The improvement of duration rules is an active area of research in text-to-speech systems[12-16] that depends heavily on collecting and analyzing large speech databases.

Current text-to-speech systems generally use only the simple syntactic analysis of text described above, in addition to punctuation that is present in the input text, to guide prosodic assignments such as phrasing and pitch prominence. However, higher-level semantic and discourse information clearly will also have to be brought to bear, especially as applications that require synthesizing longer texts (as opposed to single words or sentences) become important. For example, in human speech, previously mentioned words and phrases in text tend to be deemphasized, but items new to the discourse are more prominent. And items contrasted with other items (e.g., *GEORGE likes BUTTER, but BILL likes MARGARINE*) are often uttered with special emphasis. The topic structure of a discourse also influences how phrases are related to each other intonationally. Speakers often raise their voices to suggest they are beginning a new topic, and lower them to suggest they are ending one. The difficulty in inferring such information from unrestricted text has generally limited its use to experimental systems[17] or message-to-speech systems[18] where what will be conveyed is "known" to the system in advance. However, current work in text analysis, combined with examining natural speech databases, is providing some of the discourse-level information used to construct rules to vary features such as intonational phrasing and prominence.[19]

Synthesis. Synthesis is sometimes called *phoneme-to-speech conversion*. At least two different techniques are used in the synthesis process in different approaches to text-to-speech systems:

- *Acoustic domain rule-based synthesis* describes the method of using rules to generate each phoneme, usually by specifying acoustic parameters that characterize a digital filter. The rules also describe how the parameters should change when the acoustic signal changes from one phoneme to the next. By stringing these parameters together and using them to control a digital filter, speech is produced. Several synthesizers currently in production use this synthesis method.
- *Diphone synthesis*, another method to convert phonemes into speech, does not use rules to specify vocal-tract parameters. A diphone—i.e., a stored transition between two phonemes—is constructed by cutting an appropriate small piece of natural speech from a longer utterance. Storage of diphones is done in terms of their vocal-tract parameters, derived by mathematical analysis of the natural utterance. About 1,000 diphones are enough to produce arbitrary English speech. For synthesis, the appropriate diphones are retrieved from a table and concatenated, and phoneme durations are adjusted by stretching or compressing the diphones to the desired length. As with the acoustic domain rule-based synthesis, this method produces parameters that are fed to a digital filter to generate speech.

AT&T's current advances in diphone synthesis include improving the analysis techniques that provide vocal-tract parameters, leading to more accurate representations of these parameters.[20] Experiments with new glottal sources also show promise of improving voice quality and naturalness, as well as making it possible to change voice timbre easily when desired.

Another advance in diphone synthesis is the development of more sophisticated schemes for diphone concatenation. Programs have recently been developed to concatenate longer units of stored speech so some units may contain three or more phonemes.[21] This work allows concatenative synthesis of stored units to approach more closely the naturalness of stored speech, while maintaining the flexibility to pronounce arbitrary text and impose different prosody on the speech. For example, an often-used carrier phrase, such as *Your account balance is . . . .* could be cut from speech and stored as one unit in the table of sounds. The dollar

amount that follows could be synthesized from shorter sound units. This approaches the method used in stored voice response, where the same phrase is stored in one piece. The advantage of a TTS system in this example lies in its ability to synthesize the monetary amount with better prosody and less choppiness than stored voice techniques can deliver.

Voice Response and Stored Voice

Most current voice response applications use stored speech—not text-to-speech synthesis—because the former gives better control over intelligibility and voice quality. Intelligibility—including the naturalness of the voice response—is important to service providers because many of their callers use the service infrequently; they cannot be expected to learn the nuances of an artificial voice. Voice quality often assumes great importance in transactions; most service providers want a voice that will be perceived as pleasant, warm, friendly, and trustworthy.

Most stored speech applications involve simple announcements played out to callers; no speech from callers is recorded for future use, though speech input may be recognized to guide the transaction.[22] Perhaps the simplest such application is *call routing* (i.e., automated attendant), where stored announcements prompt the caller to select (by Touch-Tone or voice response) one of several destinations, or to enter a particular extension. A more complex transaction is *order tracking*, where announcements prompt the caller to enter an identification or security code, and then an order number, to get information about the status of an order. More complex still is *order entry*, which prompts the caller to enter identification, items desired, method of payment, and delivery address.

One complex transaction that has generated considerable interest is college registration by telephone. Another is banking by phone, where callers identify bills and merchants to be paid, payment amounts, and dates. These applications are widely used today because they use the strengths of stored voice technology: a modest number of announcements can be called on as needed to guide callers through a limited and well-structured transaction.

Voice store-and-forward applications not only use prerecorded announcements, but also capture speech from the caller for later use. The best known of these is voice mail, where the caller can record a message to be delivered to one or more "voice mailboxes." Such messages usually are targeted for specific persons or groups, much like paper mail. However, broadcast messages are also possible, such as when a system administrator sends a message to all mailboxes.

Another type of voice store-and-forward application now gaining in popularity is voice capture and transcription. Instead of leaving messages for particular receivers, callers leave messages for any agent who is capable of responding. A typical example of this application is automated order entry, which is used to handle agent overload during peak call times or permit unattended order entry. In such applications, voice quality can generally be lower, because the listeners are trained agents who can replay messages as required.

The realm of applications for stored speech is growing, spurred by practical needs to streamline business operations. Nevertheless, there are applications that are feasible only with text-to-speech technology. These applications would involve having a speaker record huge volumes of material, or material that changes rapidly, at considerable expense and inconvenience for the provider of the service.

Applications of Text-to-Speech Systems

Current applications of text-to-speech technology include "talking" terminals and training devices, proofreaders, warning and alarm systems, talking aids for the vocally handicapped, and reading aids for the blind. Audiotex services allow customers to use telephones as terminals to retrieve information from public or private databases. The information may include names and addresses from a telephone directory, financial accounts, stock quotations, weather reports, reservations, sales

orders and inventory information, or locations of commercial dealers. While some of this information could be provided using stored human speech, text-to-speech systems are appropriate when services access a large or frequently changing database that would require recording large amounts of new material. Text-to-speech technology reduces storage needs from 32 kb/s for stored speech to a few hundred bits for an equivalent text sentence. Maintaining the database is also simplified, because only the text data has to be updated.

Several telephone-based services using AT&T TTS technology have been offered on a trial basis in recent years, and results from these trials have been positive. The trials have involved sales orders and inventory information, the synthesis of names and addresses taken from a database, voice prompts, and even the synthesis of user-entered unrestricted text, as in the dual-party relay trial for hearing- or vocally-impaired customers. In a normal dual-party relay service, a customer uses a telecommunications device for the deaf (TDD) to type messages that are read by a human communications assistant to the hearing recipient of the call. A communications assistant is typically assigned for the duration of each dual-party relay call to read the TDD text and to type the recipient's spoken response and transmit it back to the TDD customer, allowing two-way communication. Applying text-to-speech synthesis to this service in a trial being conducted during the fall of 1990, typed input from a customer using a TDD, such as:

Input	Meaning
HD YES HD SORRY	Hold. Yes. Hold. Sorry.
I HAVE BUSY	I'm busy.
I LL CALL U BACK LATER	I'll call you back later.
OK WITH U	Okay with you?
Q GA	Go ahead.

is translated into synthetic speech for transmission to the hearing recipient of the call. Use of text-to-speech synthesis results in greater privacy for the TDD user and savings in human operator effort. The text translation involves some interpretation of the message, as well as breaking up generally unpunctuated sentences into manageable speech units. Such applications illustrate the increasing importance of text analysis techniques in speech synthesis, to improve the naturalness and intelligibility of synthetic speech.

There are several research applications in AT&T involving text-to-speech synthesis. These include an Associated Press (AP) News Reader, a telephone news service designed for the handicapped, and weather and mail readers. Text-to-speech synthesis is also used as a front end for a robot, the speech-activated manipulator (SAM),[23] to provide prompts for a speaker verification system, and for applications in the HuMaNet project (see Berkley and Flanagan's paper in this issue). Through TelSpeak, a telephone interface, text-to-speech programs and other speech software can be used in general telephone applications. A number of speech-to-speech systems currently being researched, including the AT&T Airline Reservations System,[24] use text-to-speech. The most ambitious speech-to-speech project to date is the VEST project, a joint project between Bell Laboratories and Telefónica, Spain's telephone company. VEST is a Spanish-English translation project that takes spoken input in one language and produces synthesized responses in the other. This project has involved building a Spanish-language version of the synthesizer.

Conclusion

In this article, we have briefly surveyed the current speech technology available for voice response systems. Stored voice, based on well-developed waveform coding and voice coding techniques, is practical and effective for many applications. To extend the range of these applications will require improvements that keep voice quality high while reducing both coding rates and computation to more economical levels. Even with such progress, however, many applications will be beyond the capabilities of stored voice, and will require the greater versatility of text-to-speech synthesis. Improvements in text analysis, syntactic processing,

prosodic assignment, synthesis methods, and voice quality should also extend the range of applications for text-to-speech technology. Such advances will bring the ideal voice response system closer to reality.

In the past, the overriding issue involving both stored voice and text-to-speech systems has been the question of which was more appropriate for a particular application, given limitations of storage, speed, quality, and task domain. Today, as speech synthesizer output comes closer to natural speech, these technologies are coming closer together as well. So, not only do applications that integrate synthetic and stored speech become ever more feasible and attractive, but shared problems may also have common solutions.

References

1. R. E. Crochiere and J. L. Flanagan, "Speech Processing: An Evolving Technology," *AT&T Technical Journal*, Vol. 65, No. 5, September/October 1986, pp. 2-11.
2. N. S. Jayant, V. B. Lawrence, and D. P. Prezas, "Coding of Speech and Wideband Audio," *AT&T Technical Journal*, Vol. 69, No. 5, September/October 1990, pp. 25-41.
3. N. Benvenuto, G. Bertocci, W. R. Daumer and D. K. Sparell, "The 32-kb/s ADPCM Coding Standard," *AT&T Technical Journal*, Vol. 65, No. 5, September/October 1986, pp. 12-22.
4. M. A. Van Andel, "While You're Away, AUDIX Will Answer," *AT&T Technology*, Vol. 3, No. 3, 1988, pp. 34-41.
5. A. C. Gillon and J. W. Moffett, "Voice Power Gives You Voice Messaging—And Then Some," *AT&T Technology*, Vol. 4, No. 2, 1989, pp. 6-11.
6. J. G. Josenhans, J. F. Lynch, Jr., M. R. Rogers, R. P. Rosinski, and W. P. VanDame, "Speech Processing Applications Standards," *AT&T Technical Journal*, Vol. 65, No. 5, September/October 1986, pp. 23-33.
7. C. Coker, K. W. Church, and M. Liberman, "Morphology and Rhyming: Two Powerful Alternatives to Letter-To-Sound Rules for Speech Synthesis," *Proceedings of the European Speech Communications Association Workshop on Speech Synthesis*, Autrans, France, September 1990, pp. 83-86.
8. R. Sproat, "Stress Assignment in Complex Nominals for English Text-to-Speech," *Proceedings of the European Speech Communications Association Workshop on Speech Synthesis*, Autrans, France, September 1990, pp. 129-132.
9. Joan Bachenko, Eileen Fitzpatrick, and C. E. Wright, "The Contribution of Parsing to Prosodic Phrasing in an Experimental Text-to-Speech System," *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, New York, 1986, pp. 145-153.
10. S. J. Young and F. Fallside, "Speech Synthesis from Concept: A Method for Speech Output from Information Systems," *Journal of the Acoustical Society of America*, Vol. 66, No. 3, September 1979, pp. 685-695.
11. L. Danlos, E. LaPorte, and F. Emerard, "Synthesis of Spoken Messages from Semantic Representations," *Proceedings of the 11th International Conference on Computational Linguistics*, Bonn, West Germany, August 1986, pp. 599-604.
12. H. S. Gopal and A. K. Syrdal, "Interaction of Speaking Rate and Postvocalic Consonantal Voicing on Vowel Duration in American English," *Journal of the Acoustical Society of America*, Vol. 82, 1987, p. S16.
13. A. K. Syrdal, "Improved Duration Rules for Text-To-Speech Synthesis," *Journal of the Acoustical Society of America*, Vol. 85, 1989, p. S43.
14. J. P. H. van Santen and Joseph P. Olive, "The Analysis of Contextual Effects on Segmental Duration," *Computer Speech and Language*, Vol. 4, 1990, in press.
15. Jan P. H. van Santen and Joseph P. Olive, "Diagnostic Tests of Segmental Duration Models," *Journal of the Acoustical Society of America*, Vol. 85, 1989, p. S43.
16. Michael D. Riley, "Some Applications of Tree-based Modeling to Speech and Language," *Proceedings of the DARPA Speech and Natural Language Workshop*, Moragn Kaufman, Cape Cod, Massachusetts, October 1989, pp. 339-352.
17. K. Silverman, *The Structure and Processing of Fundamental Frequency Contours*, Ph.D thesis, Cambridge University, England, 1987.
18. James R. Davis and Julia Hirschberg, "Assigning Intonational Features in Synthesized Spoken Directions," *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, Buffalo, New York, 1988, pp. 187-193.
19. Julia Hirschberg, "Accent and Discourse Content: Assigning Pitch Accent in Synthetic Speech," *7th National Conference of the American Association for Artificial Intelligence*, Boston, Massachusetts, July 29–August 3, 1990, pp. 952-957.
20. D. Talkin and J. Rowley, "Pitch—Synchronous Analysis and Synthesis for TTS Systems," *Proceedings of the European Speech Communications Association Workshop on Speech Synthesis*, Autrans, France, September 1990, pp. 55-58.
21. J. P. Olive,"A New Algorithm for a Concatenative Speech Synthesis System Using an Augmented Acoustic Inventory of Speech Sounds," *Proceedings of the European Speech Communications Association Workshop on Speech Synthesis*, Autrans, France, September 1990, pp. 25-29.
22. D. R. Fischell, S. S. Kanwal, and D. S. Furman, "Interactive Voice

Technology Applications," *AT&T Technical Journal*, Vol. 69, No.5, September/October 1990, pp. 61-76.
23. M. K. Brown, B. M. Buntschuh, and J. G. Wilpon, "SAM: A Smart Speech Activated Manipulator," *Symposium on Flexible Automation*, Kyoto, Japan, July 9-11, 1990.
24. S. Levinson and L. Rabiner, "A Task-Oriented Conversational Mode Speech Understanding System," *Bibliotheca Phonetica*, Vol. 12, 1985, pp. 149-196.

Biographies (continued)

in 1978 with a Ph.D. in psychology from the University of Texas, Austin. Mr. Rowley works on analysis and synthesis algorithms used in the text-to-speech system, and also is responsible for a version of TTS being ported to Conversant hardware. He joined AT&T in 1970 with a Master of Electrical Engineering from Purdue University, West Lafayette, Indiana. Ms. Syrdal is responsible for improvements in text-to-speech technology, and is currently focusing on developing a female synthetic voice. She joined AT&T in 1986 with a Ph.D. in psychology from the University of Minnesota, Minneapolis.

(Manuscript received June 10, 1990)