SPEAKING TO, FROM, AND THROUGH COMPUTERS: SPEECH TECHNOLOGIES AND USER-INTERFACE DESIGN

Raymond W. Bennett, Steven L. Greenspan, Ann K. Syrdal, Judith E. Tschirgi, and John J. Wisowaty

Raymond W. Bennett, Steven L. Greenspan, Ann K. Syrdal, Judith E. Tschirgi, *and* ***John J. Wisowaty*** *are members of the Advanced Services Technology Department at AT&T Bell Laboratories' Indian Hill Park facility, Naperville, Illinois. Raymond W. Bennett, who joined the company in 1978, is a distinguished member of technical staff responsible for exploratory research on advanced telephony services. He has a Ph.D. in psychology from the University of Michigan, Ann Arbor. Steven L. Greenspan, who has a Ph.D. in psychology from the State University of New York at Buffalo, is a consultant to the technical staff currently working on the design and evaluation of user interfaces for application-oriented programming languages. Ann K. Syrdal, a member of technical staff, is responsible for improvements in text-to-speech technology,* (continued on page 30)

Speech is the prototypical, usually preferred, communication mode for humans. Recent progress in digitally processing and transporting human speech, in synthesizing speech from text, and in automatic recognition of human speech, promises improvements in both human-to-human and human-to-machine communication. We examine the roles of the human factors specialist, applied psychologist, and linguist, in developing and deploying these new technologies, and in constructing human-to-machine interfaces particularly suitable for speech input/output.

Introduction

Communication—the exchange of ideas—is the foundation for all social transactions. It is most commonly and efficiently conducted using a natural language, such as English, as a signaling system. The language is primarily a spoken communications medium. Thus, a speaker, listener, and transmission medium—such as air—are required to complete the transaction.

The telephone network provides yet another medium for spoken communication. While it alters communication by spatially—and sometimes temporally—separating partners, it nevertheless maintains many of the conventions of face-to-face conversation. Revolutionary advances in computer and speech technology, however, have made possible a new generation of telecommunication services. New signal processing techniques allow conversations to be transmitted more efficiently and securely. More radical advances allow computers to automate one side of a telephone conversation. The result is that text-to-speech synthesis can produce human-like speech, and automatic speech recognition will recognize human speech.

The continuing goal of these telecommunications services will be to enable people to use their familiar signaling system—conversational speech—for efficient and cost-effective communication. The challenge for speech technologists and behavioral scientists is to develop computer systems that approximate the human ability to speak and perceive language. Currently, no theory of speech perception or

Panel 1. Terms and Acronyms in This Paper

ADPCM	adaptive differential pulse code modulation coder
ASCII	American Standard Code for Information Exchange, a binary code for data that is widely used in communications.
ASR	automatic speech recognition
CCITT	International Telegraph and Telephone Consultative Committee
CELP	4.8 kb/s code-excited linear prediction coder
DAM	Diagnostic Acceptability Measurement
DRT	Diagnostic Rhyme Test
DSI	Digital speech interpolation
IEEE	Institute of Electrical and Electronics Engineers
Intonation Contour	underlying pitch pattern of fundamental frequency ($F0$) that occurs over time in speech phrases
kHz	kilohertz
sustention errors	errors referring to the distinction between interrupted and uninterrupted airflow, the two modes of consonant production
TTS	text-to-speech
WPT	wideband packet technology

production exists to account fully for the complexity of human language. The syntax, phonology, and semantics of human languages allow speakers to construct any number of entirely novel, unexpected, and potentially ambiguous statements. Furthermore, listeners can perceive—and even comprehend—these utterances practically instantaneously, often under imperfect conditions (e.g., whispering, during cocktail parties, or while standing on busy street corners).

Behavioral scientists have contributed theoretical constructs and empirical findings to develop the key technologies needed to speak and comprehend speech. Although computerized systems have yet to match these remarkable human abilities, they are developed enough to begin to be deployed in telecommunication products and services. When speech processing technologies fail to provide as transparent an interface as the existing telephone system, behavioral scientists also are active in designing structured transactions to help humans adapt to the limitations of the technologies.

This paper will discuss some work of behavioral scientists in three categories of voice communication systems: *human-to-human*, *computer-to-human*, and *human-to-computer*. The services discussed illustrate:

1. How voice quality is measured to ensure the continued high performance of the voice network as transmission technology advances
2. How linguistic and psychological research are incorporated into the development of text-to-speech synthesis
3. How people's behavior must be shaped to interact successfully with automatic speech recognition devices.

This article discusses behavioral science's contributions to developing both basic voice-user interface technologies and products that rely on them.

Human-to-Human Communication: Speech Transmission

Human-to-human communication is the core business of the telecommunication industry, and the global telephone voice network is the most ambitious, economically important application of speech technology. The overall usefulness of a telecommunication network for an end user depends on everything from how long it takes to install a new telephone, to the likelihood that a call cannot be completed because of network congestion. The fidelity of the speech carried by the network is important to users. As the network evolves by incorporating more reliable and cost-efficient technology, speech quality must be maintained at acceptable levels.

Network Voice Quality Evaluations. How does one decide if an improvement in speech quality is worth the extra cost, or whether a cost savings is worth a loss in

quality? Such decisions depend on having a *quantitative index* of speech quality. At present, *human* judgment is the only feasible method to assess speech quality, because defined or measured acoustic distortions do not necessarily correspond to user perceptions.

Consider, for example, measuring a straightforward attribute of quality: the "loudness" of a circuit. *Loudness* means the perceived intensity of sound. Circuit loss (or gain) must be controlled so speech loudness is at an acceptable level. For pure tones, loudness is determined by the frequency and sound pressure level of the tone. However, for more complicated signals (e.g., any impulse noise such as typing or hammering, or any spectrally complicated signal such as speech), there is no satisfactory way to predict the loudness of a sound source from its physical attributes. Since other perceptually salient characteristics of the speech signal are even harder to predict from acoustic measures, human judgment provides the only reliable assessment of speech quality.

The most straightforward approach to measuring speech quality is to ask for overall quality ratings on a numeric or verbally labeled scale. In a typical experiment using a numerical system, participants listen to one or two sentences over a real or simulated circuit and rate them on a 5-point scale. In an hour, each judge can rate several hundred samples, typically making a few dozen ratings per circuit. The "mean opinion score"—i.e., the average rating assigned a circuit by judges—is commonly used to summarize overall ratings. This procedure can be used with any set of speech samples, without concern for the particular impairments introduced. The judges need no training and only brief instructions. Various versions of this procedure are used throughout the telecommunications industry, and have been incorporated into IEEE and CCITT practices for measuring speech quality. (IEEE is the Institute of Electrical and Electronics Engineers, and CCITT is the International Telegraph and Telephone Consultative Committee.)

An alternative procedure is to itemize the possible types of perceptual impairments, then have participants judge the severity of those impairments in speech samples. A global quality measure can be devised by weighting the magnitude and importance of the various impairments. Voiers' Diagnostic Acceptability Measure (DAM)[1] is the best known of these procedures. Participants listen to samples of simple sentences and rate each sample on several attributes. The ratings include judgments of both the characteristics of the speech signal (e.g., *fluttering*, *thin*, *nasal*, or *muffled*) and the background (e.g., *hissing*, *buzzing*, or *rumbling*). The DAM provides scores with diagnostic value, not only measuring the fidelity of speech reproduction, but also identifying what needs improvement. However, the cost of developing such scales may be high. To produce a quality metric, they require both an inventory of impairments and a nonlinear rule for combining ratings of impairments. Also, it is often necessary to use trained judges.

To assess a technology's acceptability for network use, measures of speech quality should predict user satisfaction with comparable service during day-to-day use of the telephone network. Field studies, because they are made in the physical and social context in which we are most interested in predicting behavior, provide the most ecologically valid observations. In some field studies, customers use experimental circuits to make routine telephone calls, without notification that the circuit is atypical. They are then asked to judge the circuit's quality.

For example, the network planning organization regularly evaluates equipment, designed either by AT&T or outside vendors, to gauge its suitability for use in the network. In a typical experiment, two central offices (e.g., in Cleveland and Phoenix) are selected, and experimental circuits are installed. There might be one baseline condition where AT&T standard equipment is used, and two or more experimental conditions that use the real or simulated circuits to be evaluated. Customers make routine calls, some of which are routed over either the baseline or the experimental circuits. After the call is completed, either the called or the calling party is interviewed, using a standard questionnaire such as that described by DiBiaso.[2] The respondent is asked if there were any problems with the call, and about several

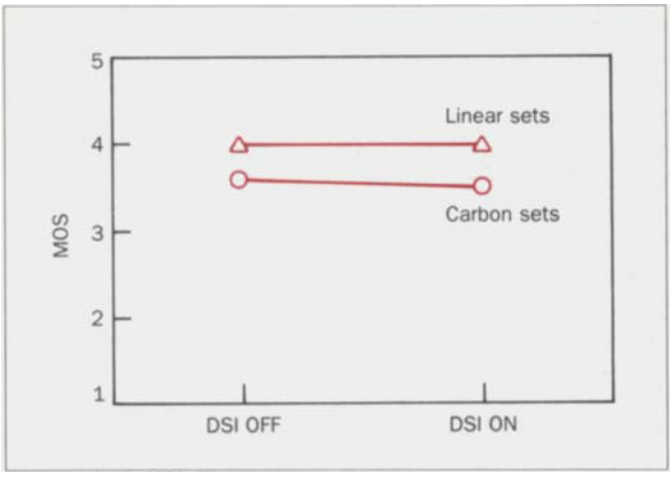

Figure 1. Effects of DSI on perceived speech quality as a function of microphone.

possible difficulties such as low volume, crosstalk, noise, echo, or delay. The respondent then rates the overall quality of the connection on a four-point scale (excellent, good, fair, or poor). Both global quality ratings and detailed descriptions of impairments are obtained without using trained judges. However, interviews can take 10 minutes or more.

Data from such field studies are most useful in predicting user reactions to speech technology embedded in a real service. However, such studies are expensive and time-consuming. Collecting sufficient data on even a few circuits can require months. Early in a product's life, developers and researchers frequently need measures of how well several variants of the product or prototype are performing. Because field studies are usually impractical, some variety of laboratory experiment must be substituted.

Laboratory experiments use either *listening* or *conversational* tests. In listening experiments, participants judge the quality of selected speech samples heard over real or simulated circuits. These experiments are the least expensive, but cannot be used to evaluate impairments such as delay or talker echo. Conversational experiments mimic a more natural telephone interaction, because pairs of judges converse over experimental circuits for several minutes before rating speech quality. In both types of experiments, judges either can make overall quality ratings, or can evaluate specific impairments.

D. O. Bowker and C. A. Dvorak of AT&T Bell Laboratories have recently completed an extensive series of both conversational and listening tests[3] that evaluated the quality of speech generated using wideband packet technology (WPT). One advantage of WPT is that bandwidth can be compressed when a customer application requires it. In particular, embedded digital speech coders can be used to decrease bit rate from the standard 64 kilobits per second (kb/s) to 32 or even 16 kb/s during periods of packet buffer overflow. It also is possible to turn off the packet stream when no speech is being produced (usually referred to as digital speech interpolation, or DSI). Bowker and Dvorak used conversational tests and global quality ratings to show that DSI had no detectable effects on perceived speech quality. Yet their experiment was sensitive enough to show a difference between electret and carbon microphones (see Figure 1). They further demonstrated the significant but small effects of decreasing effective bit rate from 4.0 bits per sample to 3.7 bits per sample using an embedded adaptive differential pulse code modulation (ADPCM) coder.

An attractive application of WPT is digital circuit multiplication. In normal use, a T1 digital facility operating at 1.544 megabits per second (Mb/s) can carry 24 simultaneous voice channels. Listening tests were used to evaluate the effect on voice quality of using WPT to extend the capacity of the T1 facility to up to 160 conversations. Loads of up to 80 channels could be carried with quality equivalent to that provided by 32 kb/s ADPCM, which is just slightly inferior to 64 kb/s PCM (see Figure 2). There were decreases in voice quality with higher loads that might be acceptable depending on the user's application.

Low-Bit-Rate Coded Speech Evaluations. While telephone-quality speech—a bandpass of roughly 300 to 3,000 hertz (Hz) and a signal-to-noise ratio of at least 25 decibels (dB)—is comfortable for most uses, both higher and lower quality channels are used regularly. End-to-end digital transmission of speech will make it possible to use high-bit-rate coders (56 or 64 kb/s) that will reproduce speech about as well as commercial AM radio stations using a bandpass of 15 Hz to 7.5 kilohertz (kHz). Coding speech at much lower bit rates (1.2 kb/s to 9.6 kb/s) is desirable when communications channels are of limited bandwidth, or when quantities of digitized speech must be stored.

When speech is coded at low-bit rates, assessment difficulties occur that are not present in toll-quality speech:

- First, if telephone speech is processed by low-bit-rate coders, the normal intelligibility of that speech cannot be assumed. For most applications of low-bit-rate speech, then, intelligibility is the most important criterion.
- Second, even if a coder produces understandable speech, other information affecting communication may be missing from the speech signal. For example, it may be difficult to identify even a familiar speaker. Judgments about the speaker's physical and emotional state may be unreliable.
- Finally, listening to low-bit-rate coded speech may take extra effort. Having to concentrate on understanding the less intelligible coded speech signal may fatigue the listener and limit his or her ability to pay attention to other tasks—such as following the conversation or flying an airplane—and is likely to add to fatigue.

Speech Intelligibility. There are two major laboratory-based assessment techniques for measuring speech intelligibility. A *comprehension task* requires participants to listen to a recorded passage, then answer content questions about it. In measures of *segmental intelligibility*, listeners hear a list of words or syllables without context, and identify the basic speech sounds. Segmental intelligibility studies assume that, for larger units of speech—words, phrases, or sentences—to be understood, the basic speech units must necessarily be recognized.

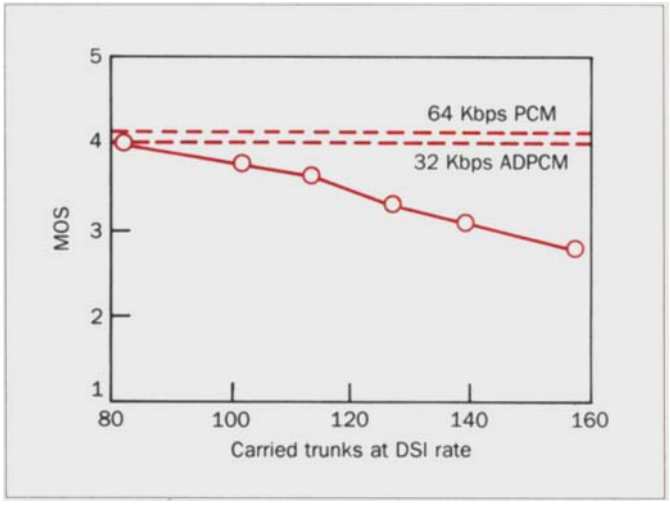

Figure 2. Effect of number of channels on perceived speech quality carried at DSI rate.

A popular method of measuring segmental intelligibility is the Diagnostic Rhyme Test (DRT), described by Voiers. Participants listen to monosyllabic words, and then select the word they heard from two alternatives that differ only by minimal distinctions in their initial consonants. For example, the initial consonant of one alternative may be voiced, as the */b/* in *bond*; the other alternative is unvoiced, as the */p/* in *pond*. The pattern of errors made by listeners in the DRT shows how well the speech transmission system under study preserves the six types of distinctions tested. Though this information is restricted to a small subset of all possible intelligibility confusions, it can be used effectively by developers to adjust the hardware and the coding algorithm to improve performance.

Testing During Product Development. The AT&T *Security-Plus* phone is the company's response to a

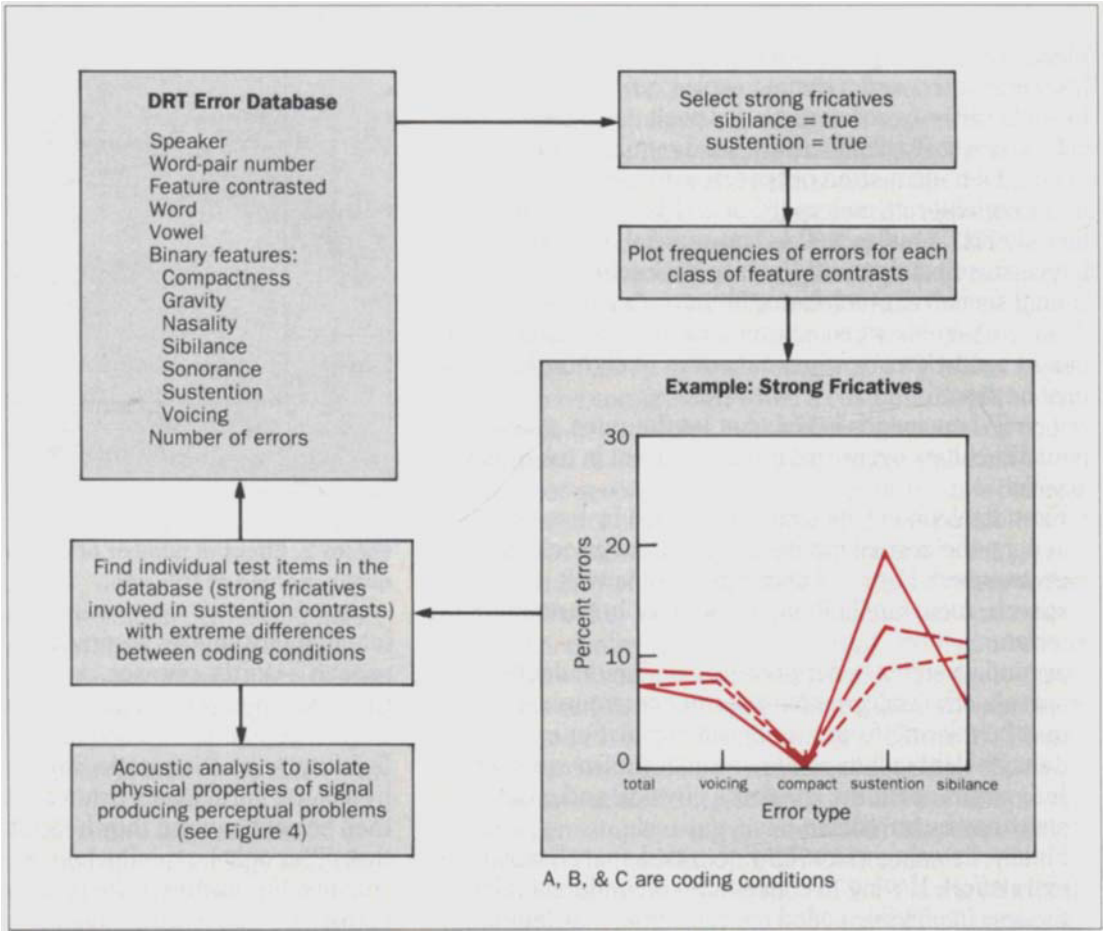

Figure 3. Example of the use of DRT error base in improving a low bit-rate voice coder.

competitive federal program to develop the next generation of the secure terminal unit (the STU-III). This station set provides secure communication by coding speech at 2.4 kb/s or 4.8 kb/s, encrypting it, and then transmitting the encrypted speech via a modem through the existing analog network to another similarly equipped station set. The Security-Plus phone was designed to meet rigorous U. S. government specifications and acceptability measures, including speech intelligibility and quality.

Government-supervised speech evaluations were conducted by an independent testing laboratory. The DRT was used to measure speech intelligibility, and the DAM to measure speech quality. AT&T Bell Laboratories developed voice coding algorithms for implementation in the secure telephone. Human factors input during the development process provided feedback about aspects of the coded speech signal that needed improvement, and often provided information about the direction improve-

ments should take.

During product development, DRTs and DAMs were regularly conducted by the outside laboratory. AT&T's human factors specialists took responsibility for statistical analyses of the results, and for devising procedures that allowed developers to track their own progress. Creation of a DRT error database provided detailed analyses of perceptual errors made in intelligibility tests.[4] The analysis involved storing, classifying, sorting, and counting listeners' perceptual errors, using a database containing other fields of descriptive phonetic information. Errors could thus be broken down in a variety of ways for further examination. The major advantages of detailed DRT results was that they could be related more directly to the acoustics of the signal being evaluated, and allowed great flexibility in the types of analyses that could be performed. The detailed DRT analyses proved useful in diagnosing specific problems in coding algorithms and filtering procedures.

Figure 3 illustrates how the error database was used to improve a speech coder. First, filters specifying search conditions were used to sort stimuli into six broad phonetic classes tested by the DRT:

- Weak fricatives (*f, v, th, dh*)
- Strong fricatives (*s, z, sh, zh*)
- Stops (*b, d, g, p, t, k*)
- Nasals (*m, n*)
- Liquids and glides (*w, r, l, y*)
- Affricates (*ch, j*).

Within these classes, additional search conditions were used to identify particularly problematic errors, such as voicing errors for strong fricatives. Then, further search conditions were used to isolate extreme examples of problematic error types (i.e., a specific word spoken by a specific speaker) to examine the acoustic properties associated with the perceptual errors.

For example, a phonetic classification of perceptual errors led to the observation that a much higher proportion of sustention errors for strong fricatives occurred for words coded by a hardware prototype than by its software simulation. (Sustention errors refer to the distinction between interrupted and uninterrupted airflow, two modes of consonant production.) The error database was used to locate test words spoken by specific speakers for which the number of errors differed substantially among the systems being compared. These coded test words were then analyzed acoustically to determine the acoustic characteristics of the coded signals responsible for the large perceptual differences observed.

Figure 4 shows a specific example of the acoustic characteristics of a test word—*shoes*—spoken by a specific male speaker, and processed by different systems. The filtered but uncoded word (see waveform in top panel) was correctly identified by 100 percent of the listeners. The same word, filtered and coded by software simulation (middle panel), was identified correctly by 88 percent of the listeners. When filtered and coded by a hardware prototype (bottom panel), the test word *shoes* was never identified correctly by any of the listeners (i.e., 0 percent correct). Instead, they identified it as *choose*, the other response alternative available in the forced choice DRT test format.

Affricates, such as the first consonant in *choose*, are produced with interrupted airflow because of a momentary but complete closure of the vocal tract. Affricates generally have aperiodic energy that is shorter in duration, and more abrupt in amplitude onset, than fricatives. Our example is the first consonant in "shoes," for which airflow is uninterrupted during consonant production. Behavioral research in speech perception has shown these acoustic characteristics are critical for the perceptual distinction between fricatives and affricates.

A waveform comparison shows that the acoustic characteristics of the coded speech—related to the poor intelligibility scores for the hardware coder—involve both rise-time and duration distortions. The major cause of these problematic acoustic characteristics was that low energy was being further rounded down by the hardware coder. Because of these analyses, more precision was provided for encoding low energy, and identification

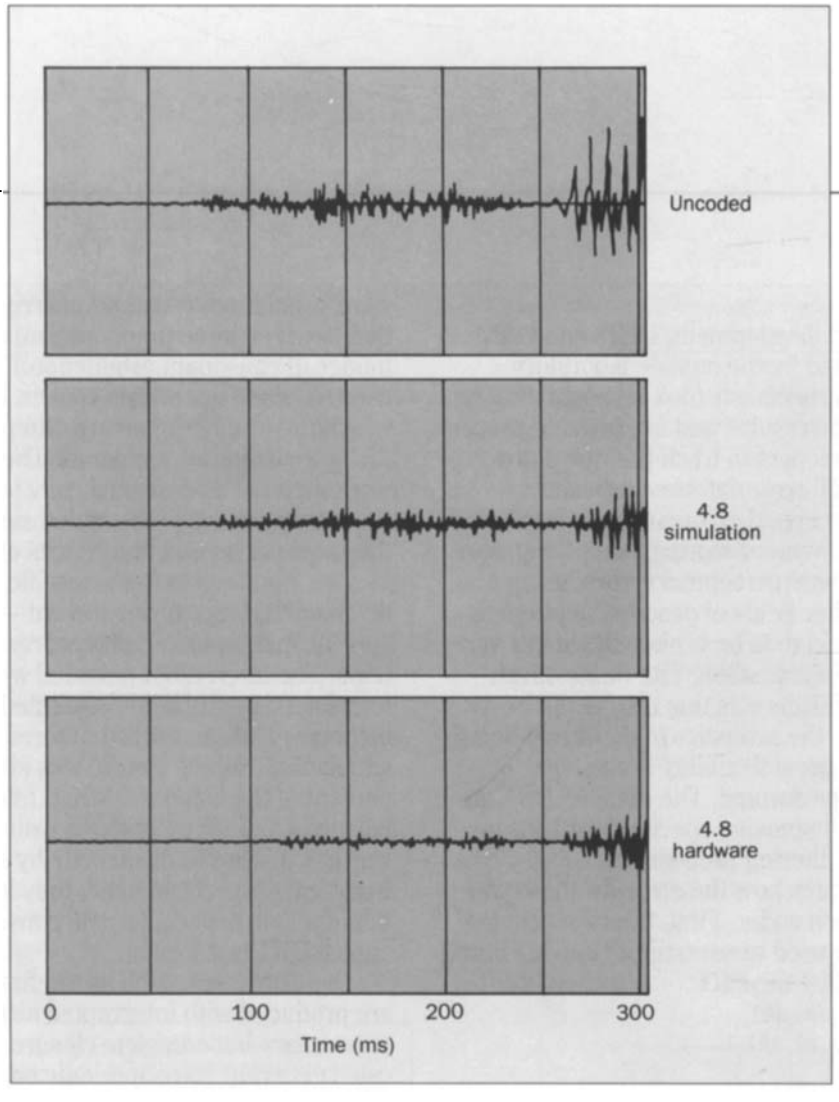

Figure 4. Acoustic analysis of sustention errors for the word *shoes*.

accuracy improved. During the 13-month development period over which these analyses were used, overall DRT intelligibility scores went from 88.8 percent correct to 92.1 percent correct, an increase of over four standard errors.

The result of these efforts was that AT&T's product was recognized by the government for its high voice quality and intelligibility, and its new 4.8 kb/s Code-Excited Linear Prediction (CELP) coder was adopted as a new government standard. Figure 5 illustrates the improvements in voice quality from STU-II, the government's previous secure telephone, to the initial 2.4- and 4.8-kb/s coders used in it's successor, STU-III, and finally to a 4.8 kb/s algorithm that will soon be introduced. The effective maximum achievable quality score is shown by a dashed horizontal line at the top of the figure; this score was for uncoded speech with the same bandpass as the telephone network.

Text-to-Speech Synthesis

Text-to-speech (TTS) technology permits communication from computer to human by allowing the computer to generate human-like speech directly from stored text. Terminal-based uses for TTS include:

- Warning and alarm systems
- Speech-emulating terminals and training devices
- Proofreading
- Talking aids for the vocally handicapped
- Reading aids for the blind.

Audiotext services allow users to retrieve information from public or private databases using a telephone as a terminal. The information may include:

- Names and addresses from a telephone directory
- Financial accounts
- Stock quotations
- Weather reports
- Reservations
- Sales orders and inventory information
- Locations of commercial dealers.

While some of this information could be provided using stored human speech, TTS systems are appropriate when services access a large or frequently changing database. TTS reduces storage needs from 64 kb/s for stored speech to a few hundred bits for an equivalent text sentence. Maintaining the database is also simplified, because only the textual data must be updated.

Text-to-Speech Technology. A TTS system consists of several subcomponents that perform different functions. These include:

- *Text normalization.* Identifying sentence boundaries, and expanding conventional abbreviations. For example, *St.* → *Saint* or *Street*, and translating nonalphabetic characters into an appropriate pronounceable form (e.g., *$1234.56* → *one thousand, two hundred thirty-four dollars and fifty-six cents*).
- *Syntactic analysis.* Parsing the text to categorize parts of speech (e.g., noun, verb, adjective).
- *Letter-to-phoneme conversion and lexical stress assignment.* Mapping orthographic characters into the appropriate phonemes (i.e., strings of abstract sound units) and associated stress markers using a dictionary and letter-to-sound rules. *Orthography* refers to the standard spelling of words with regular alphabetic characters. For example, the orthographic string *gh* is translated to the phoneme */f/* in *tough*, the phoneme */g/* in *ghost*, and is silent in *though*. In an example of stress assignment, *record* has primary stress on the first syllable if it's a noun, but on the second syllable if it's a verb.
- *Determination of prosody.* Assigning the timing and intonation (the pitch changes) necessary for an intelligible and natural-sounding utterance.
- *Speech synthesis.* Conversion of a mathematical speech representation (such as vocal tract resonance patterns or linear predictive coding) into an audible utterance.

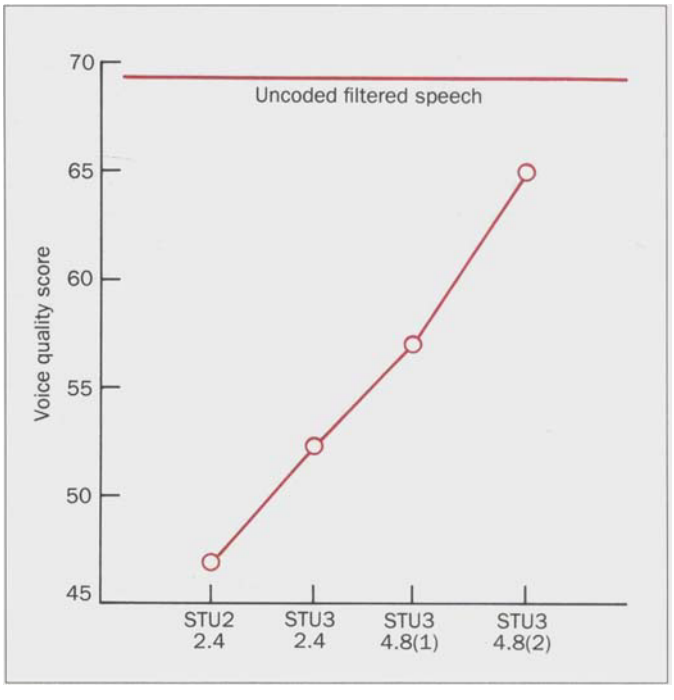

Figure 5. Voice quality (DAM) scores for low-bit-rate coders used for secure voice terminal.

Behavioral Science Research from AT&T Bell Laboratories. The goal of research and development in TTS is to generate intelligible and natural sounding synthetic speech. We will discuss recent linguistic and psychological contributions that behavioral scientists at AT&T Bell Laboratories have made to improve the prosody (i.e., intonation and timing) of TTS systems. These contributions have been developed from the following sources:

- *Acoustic analyses of natural speech produced by humans.* Hypotheses are then formulated about the

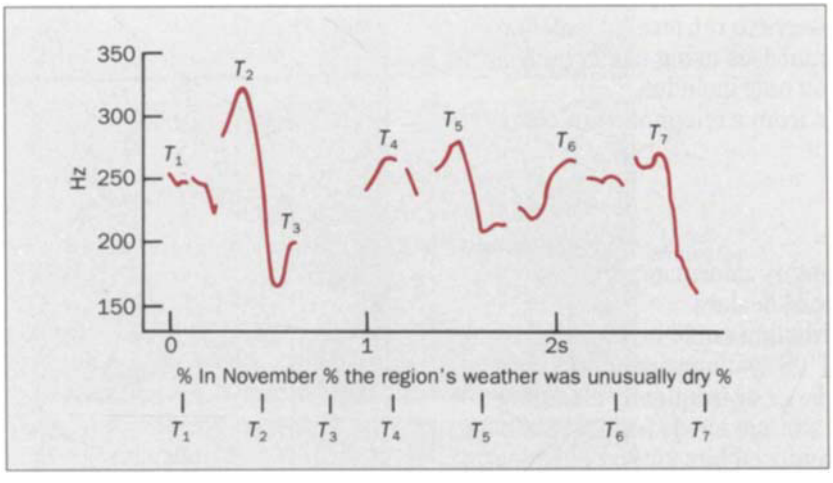

Figure 6. *F0* contour for an utterance of *In November, the region's weather was unusually dry.* *T1* through *T7* are points in the contour interpreted as *F0* targets in the present synthesis rules. The labeling under the contour indicates how these targets are aligned with syllables and phrase boundaries (%).

basic units and rules that determine the observed acoustic variations.

- *Perceptual evaluations that test hypotheses about acoustic variations.* These evaluations measure listener judgments about synthetic speech generated by competing rule systems.

Improvements in TTS prosody rules. Without prosodic rules for intonation and time, both the quality and intelligibility of synthetic speech would be impaired. Synthetic speech would lack fluency and be monotonic, jerky and difficult to understand. An *intonation contour* is the underlying pitch pattern of the fundamental frequency (*F*0) that occurs over time in speech phrases. The intonation contour distinguishes statements (falling intonation contours) from questions (rising contours). It often conveys information about syntactic structure, discourse structure, and the speaker's attitude. Figure 6 illustrates an intonation contour for the spoken sentence *In November, the region's weather was unusually dry.*

A theory of intonational structure by computational linguist Janet B. Pierrehumbert and her colleagues[5-8] provides an elegant and efficient means to achieve a variety of natural-sounding intonation contours. Pierrehumbert theorizes that intonation contours are composed of only two types of target tones, high (H) and low (L). These can be organized into many acceptable linear tonal sequences according to a grammar. Figure 7 shows a synthetic intonation contour generated by an experimental TTS system using these rules. The same sentence, depending on the intended meaning, could also be given other intonational structures. For example, when the phrase *In November* is given a rise-fall-rise intonation contour composed of a sequence of low-high-low-high abstract tones, listeners tend to interpret the sentence to mean *In November* [as opposed to some other time], *the region's weather was unusually dry.* AT&T's experimental TTS intonation rules strive for an acceptable, somewhat neutral, intonation for unrestricted text, because text is often ambiguous. However, when the text's intent is known, as in an announcement for a specific application, the intonation contour may be readily tailored for maximum naturalness and effectiveness.

Hirschberg and Pierrehumbert[9] and Silverman[10] have shown that variation in the intonational variables of *pitch range* (a specified range of *F*0 values) and *final lowering* are important in conveying discourse structure. An increase in pitch range marks the introduction of a new topic of discourse, analogous to a textual paragraph.

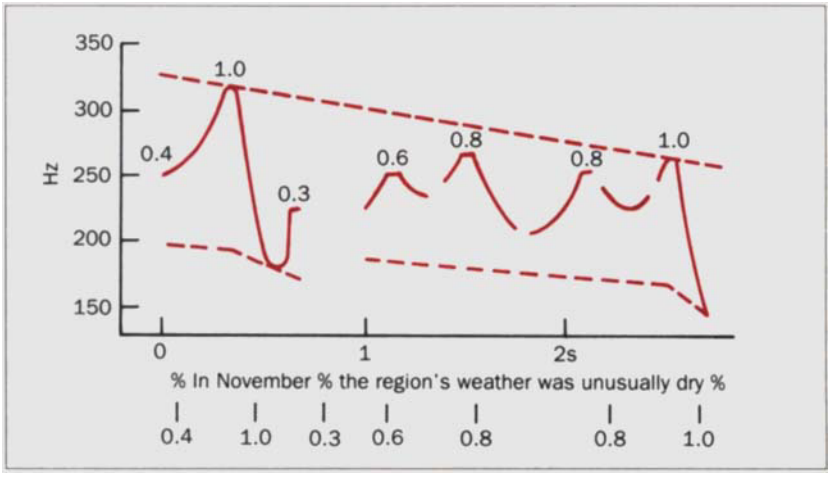

Figure 7. Synthetic version of the *F0* contour in Figure 6. The dashed lines indicate the current *F0* range at each point in time. *F0* targets and their alignment with the text are indicated by decimal numbers, which are interpreted as a proportion of the way from the bottom to the top of the *F0* range.

In final lowering—another type of local pitch range variation—the pitch range is gradually lowered and compressed during the final portion of an utterance. Final lowering marks the end of a topic or discourse. Synthetic speech generated by rule from abstract representations of knowledge, as used in artificial intelligence, can use intonational variables to organize discourse and convey information more effectively and naturally.

Studies and models of segmental durations are currently being used to improve TTS timing rules. These contribute to both intelligibility and speech quality. Gopal and Syrdal[11] and Van Santen[12] have analyzed natural speech to find better mathematical descriptions of rule systems that determine phoneme durations. Syrdal[13] developed a new TTS duration rule system using some of these results. In perceptual evaluations, this rule system was preferred over the previous rules about 80 percent of the time. It also was slightly preferred over synthetic versions that had durations equivalent to those observed in the same utterances spoken naturally.

Automatic Speech Recognition

Human-to-computer voice communication using automatic speech recognition (ASR)[14] can serve several functions.

- It can reduce the need for human attendants in uncomplicated interactions by substituting speech recognition devices.
- It can expand a user's control options when they are limited by the available input device (e.g., handset and keypad).
- It can eliminate manual control for a device.

ASR Technology. Automatic speech recognition devices can be categorized by three major dimensions that affect the user interface: training, vocabulary size, and connectedness of speech.

- *Training.* Most commercial ASR devices must be trained by collecting samples of speech. Speaker-dependent devices are trained by each individual who will use the system. Speaker-independent devices may be trained during development, using a broad sample of speakers' voices, or are developed using empirically derived models of speech.
- *Vocabulary.* Most ASR devices recognize a constrained set of words or utterances. For the more technically difficult speaker-independent systems, accurate recognition (>95 percent) can only be done today for small sets of words (2 to 50) that are phonetically distinct.

For speaker-dependent systems, vocabularies may range into the thousands. This is especially true if the application has a syntax that further constrains the word possibilities at any point in the transaction.

- *Speech connectedness.* ASR devices vary in their ability to detect and recognize continuous human speech. The simplest recognizers, *isolated word* systems, require that each utterance—either a single word or short phrase—be spoken in isolation, e.g., with enough silence on either side of the utterance for the system to detect start and stop points. More sophisticated devices allow a user to speak connectedly—for example, by speaking a string of digits, or spelling a name.

The goal of most ASR research is to allow speakers to use conversational speech without restrictions on their speed or vocabulary. This places the processing burden on the device, not the user. However, current ASR technology is not yet mature. Furthermore, there are additional constraints on the technology imposed by the 300–3,200 Hz bandwidth of most telecommunication networks. Therefore, products must be designed so users adapt as much as possible to the limitations of the technology. For example, if the ASR device expects isolated word input during a brief fixed interval, instructions must be designed to elicit that type of speech from users, even though their natural tendency is to speak in phrases. Human factors studies for trials of automated call-classification services illustrate how appropriate instructions can be designed and tested, and how different needs and limitations must be considered when designing a service.

Automating Call Classification with ASR. Many types of special billing (i.e., 0+) calls—collect, person-to-person, person-collect, bill-to-third number, and calling card—require operator assistance during call setup. However, if the desired call type can be classified by an ASR device, operator intervention is required only if the call is answered. Such an automated operator feature would require 0+ callers to be prompted by a recorded announcement to speak one of several key words, and to do so with only a single keyword at the appropriate time.

Before a 1987 field trial of this service using prototype ASR equipment, human factors studies were conducted by Gellman and Whitten[15] to select an appropriate announcement wording that would elicit the correct speech from callers.

To measure callers' responses to an automated operator, a complete simulation system was constructed, including a simulated telephone network, live attendants to act as ASR devices, and an announcement capability. Experimental subjects were told only that they would be evaluating a new (fictitious) telephone information service. They were instructed to place an operator-assisted long-distance call to reach the attendant-based service. They encountered one set of automatic operator protocols simulated on the artificial network. For example, when dialing a collect call, the subject heard an announcement similar to—*At the tone, please say collect, calling card, third number, person, or operator*—followed by a beep. The subject would respond appropriately by speaking only the single menu item requested within a few seconds after the prompt. The subject's recorded responses to the automated operator tested the experimental hypotheses about how best to cue people on *when and how* to speak, and about how much learning took place as subjects used the service repeatedly.

To test for cueing when to talk, three announcement conditions were constructed, using either of two prompt phrases, and one of three cueing methods. Table I lists the three announcement types and the percentage of correct responses. The *at the tone* condition (prior warning plus tone at the end of the prompt) was significantly better at eliciting appropriate behavior. There was no difference between the two phrases, nor were there significant interactions between prompt and cueing method. To test for learning effect, subjects made three operator-assisted calls to the fictitious service. A dramatic learning effect was observed. Across all prompting conditions and call types, 58 percent of the subjects succeeded on the first trial, 72 percent on the second, and 74 percent on the third. This showed that a reprompt proce-

Table I. Percent Correct Responses in Responding to ACTR Cueing Methods

Initial Prompt-Menu—"Now" (i.e., *Please say collect, calling card, third number, person, or operator—Now*)	65%
Initial Prompt-Menu—"Tone" (i.e., *Please say collect, calling card, third number, person, or operator—Beep*)	58%
"At The Tone"—Initial Prompt-Menu—"Tone" (i.e., *At the tone, please say collect, calling card, third number, person, or operator—Beep*)	79%

dure following an erroneous response would aid significantly in eliciting the correct response.

The results of this experiment were incorporated into an operator services field trial of the automated operator. While the results of the field trial were promising, the callers' interaction with the system clearly needed improvement. By using the pre-tested prompting announcements, about 63 percent of the callers responded appropriately with an isolated keyword in their first attempt during a session. The remaining 37 percent of the callers added extraneous utterances (e.g., *please* or *uhh*) or paraphrased the keywords (e.g., saying *reverse the charges* instead of *collect*) or said nothing. When reprompted, nearly 60 percent responded with an isolated menu item. Results such as these show the need for continued ASR algorithm development, and, more generally, for the need to adapt technology to human performance and social convention. For example, larger vocabularies are needed for synonym recognition and acceptance. Algorithms will also be needed to recognize key words embedded in longer utterances—as described by Bossemeyer et al.[16]—and refinements need to be made to the human-to-computer speech transaction.

Conclusion

Recent developments in speech processing technologies have improved basic telephone service by providing more efficient and secure speech transmission. They also have permitted human-computer dialogues where computers produce speech from stored text, or recognize and interpret human speech. Because the telecommunication network supports important business and social activities, each new product and service must be introduced with attention to network quality and integrity.

An important criterion of quality is *user acceptance*. Does the speech sound natural and intelligible? Does the service design aid the user's ability to understand a spoken message and respond appropriately? As with all services and products, speech applications must be sensitive to the user's expectations, physiological and cognitive limitations, and social constraints. Because services automated with speech technologies extend human voice communication, designers must recognize that users have well-formed—and not easily modified—expectations about speech transactions. Thus, developing, evaluating, and adapting technology to human expectations and needs must be an intrinsic part of the product life cycle, from basic research to the maintenance of mature products and services.

This paper has reviewed several contributions of behavioral scientists in developing telecommunications services using speech technology. In human-to-human communication, speech technology is used to transmit or reproduce voice input as accurately as possible. We have focused on describing the contributions of human factors specialists in evaluating the quality of the processed speech and improving the transmission technologies. In computer-to-human and human-to-computer communication, text-to-speech synthesis and automatic speech-recognition technologies must translate a message from one modality to another, from ASCII text to speech, or vice versa. From a human factors perspective, the translation algorithm must emulate some aspect of human behavior, so its behavior is natural and rational to users. With text-to-speech synthesis, we focused on how linguists and psychologists have made theoretical and empirical contributions to the development of the technology. And with ASR, we focused on how human factors specialists determine how best to adapt human behavior to the limitations of a still maturing technology.

The combined efforts of behavioral scientists, speech technologists, and product developers, can meet the challenge of harnessing computers to enhance human communication and information transfer. Just as computers have made possible revolutionary advances in speech processing, so speech processing technology

may someday revolutionize how people communicate with computers.

References

1. W. D. Voiers, "Methods Of Predicting User Acceptance of Voice Communications Systems," Final Report of Contract No. DCA100-74-C0056, prepared for the Defense Communications Agency, Reston, Virginia, July 1976.
2. L. S. DiBiaso, "Satellite User Reaction Tests: A Subjective Evaluation of Echo Control Methods," *National Technical Conference*, Vol. 3, (CH1514-9/79/0000-0212), September 1979, pp. 48.6.1-48.6.6.
3. D. O. Bowker and C. A. Dvorak, "Speech Transmission Quality of Wideband Packet Technology, *Globecom '87*, pp. 47.7.1-47.7.3.
4. A. K. Syrdal, "Methods for a Detailed Analysis of Dynastat DRT Results," AT&T Bell Laboratories Technical Memorandum, 1987.
5. J. B. Pierrehumbert, *The Phonology and Phonetics of English Intonation*, Ph.D. dissertation, Massachusetts Institute of Technology, 1980.
6. J. B. Pierrehumbert, "Synthesizing Intonation," *Journal of the Acoustical Society of America*, Vol. 40, No. 70(4), October 1981, pp. 985-995.
7. M. Y. Liberman and J. B. Pierrehumbert, "Intonational Invariance Under Changes in Pitch Range and Length," *Language Sound Structure: Studies in Phonology Presented to Morris Halle*, M. Arnoff and R. Oehrle, eds., MIT Press, Cambridge, Massachusetts, 1984, pp. 157-233.
8. M. D. Anderson, J. B. Pierrehumbert and M. Y Liberman, "Synthesis By Rule of English intonation patterns," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1984, pp. 2.8.1-2.8.4.
9. J. Hirschberg and J. B. Pierrehumbert, "The Intonational Structuring of Discourse," *Proceedings of the 24th Annual Meeting, Association for Computational Linguistics*, New York, 1986, pp. 136-144.
10. K. Silverman, *The Structure and Processing of Fundamental Frequency Contours*, Ph.D. thesis, Cambridge University, Cambridge, England, 1987.
11. 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, S16, 1987.
12. J. P. H. Van Santen and J. P. Olive, "Diagnostic Tests of Segmental Duration Models," *Journal of the Acoustical Society of America*, Vol. 85 (Supplement 1), S43, 1989.
13. A. K. Syrdal, "Improved Duration Rules for Text-To-Speech Synthesis, *Journal of the Acoustical Society of America*, Vol. 85 (Supplement 1), S43, 1989.
14. B. S. Atal and L. R. Rabiner, "Speech Research Directions," *AT&T Technical Journal*, Vol. 65, No. 5, September/October 1987, pp. 75-67.
15. R. W. Bossemeyer, J. G. Wilpon, C. H. Lee and L. R. Rabiner, "Automatic Speech Recognition of Small Vocabularies Within The Context of Unconstrained Input," *Journal of the Acoustical Society of America*, Vol. 84 (Supplement 1), S212, 1988.
16. L. H. Gellman and W. B. Whitten, "Simulating An Automatic Operator Service to Optimize Customer Success," *Proceedings of the 12th International Symposium on Human Factors in Telecommunications*, The Hague, Netherlands, May 1988.

Biographies (continued)

including duration rules and female voice synthesis. She has a Ph.D. in psychology from the University of Minnesota, Minneapolis, and joined AT&T in 1986. Judith E. Tschirgi, a supervisor in the department, joined AT&T in 1979, and works on the development of new switching services prototypes, their underlying technologies, and paths to product creation. She has a Ph.D. in experimental psychology from the University of California, San Diego. John J. Wisowaty, who joined AT&T in 1984, has a Ph.D. in experimental psychology from the University of California, San Diego, and is responsible for evaluating and improving proper name pronunciation for speech synthesis systems.

(Manuscript received June 13, 1989)