

This paper was presented at a colloquium entitled “Human–Machine Communication by Voice,” organized by Lawrence R. Rabiner, held by the National Academy of Sciences at The Arnold and Mabel Beckman Center in Irvine, CA, February 8–9, 1993.

Toward the ultimate synthesis/recognition system

SADAOKI FURUI

Nippon Telegraph and Telephone (NTT) Human Interface Laboratories, 3-9-11 Midori-cho, Musashino-shi, Tokyo, 180 Japan

ABSTRACT This paper predicts speech synthesis, speech recognition, and speaker recognition technology for the year 2001, and it describes the most important research problems to be solved in order to arrive at these ultimate synthesis and recognition systems. The problems for speech synthesis include natural and intelligible voice production, prosody control based on meaning, capability of controlling synthesized voice quality and choosing individual speaking style, multilingual and multidialectal synthesis, choice of application-oriented speaking styles, capability of adding emotion, and synthesis from concepts. The problems for speech recognition include robust recognition against speech variations, adaptation/normalization to variations due to environmental conditions and speakers, automatic knowledge acquisition for acoustic and linguistic modeling, spontaneous speech recognition, naturalness and ease of human–machine interaction, and recognition of emotion. The problems for speaker recognition are similar to those for speech recognition. The research topics related to all these techniques include the use of articulatory and perceptual constraints and evaluation methods for measuring the quality of technology and systems.

VISION OF THE FUTURE

For the majority of humankind, speech production and understanding are quite natural and unconsciously acquired processes performed quickly and effectively throughout our daily lives. By the year 2001, speech synthesis and recognition systems are expected to play important roles in advanced user-friendly human–machine interfaces (1). Speech recognition systems include not only those that recognize messages but also those that recognize the identity of the speaker. Services using these systems will include database access and management, various order-made services, dictation and editing, electronic secretarial assistance, robots (e.g., the computer HAL in *2001—A Space Odyssey*), automatic interpreting (translating) telephony, security control, and aids for the handicapped (e.g., reading aids for the blind and speaking aids for the vocally handicapped) (2). Today, many people in developed countries are employed to sit at computer terminals wearing telephone headsets and transfer information from callers to computer systems (databases) and vice versa (information and transaction services). According to the basic idea that boring and repetitive tasks done by human beings should be taken over by machines, these information-transfer workers should be replaced by speech recognition and synthesis machines. Dictation or voice typewriting is expected to increase the speed of input to computers and to allow many operations to be carried out without hand or eye movements that distract attention from the task on the display.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Fig. 1 shows a typical structure for task-specific voice control and dialogue systems. Although the speech recognizer, which converts spoken input into text, and the language analyzer, which extracts meaning from text, are separated into two boxes in the figure, it is desirable that they perform with tight mutual connection, since it is necessary to use semantic information efficiently in the recognizer to obtain correct texts. How to combine these two functions is a most important issue, especially in conversational speech recognition (understanding). Then, the meanings extracted by the language analyzer are used to drive an expert system to select the desired action, to issue commands to various systems, and to receive data from these systems. Replies from the expert system are transferred to a text generator that constructs reply texts. Finally, the text replies are converted into speech by a text-to-speech synthesizer. “Synthesis from concepts” is performed by the combination of the text generator and the text-to-speech synthesizer.

Fig. 2 shows hierarchical relationships among the various types of speech recognition, understanding, synthesis, and coding technologies. The higher the level, the more abstract the information. This figure is closely related to Fig. 1; speech recognition/understanding is the process progressing upward from the bottom to one of the higher levels of Fig. 2, and speech synthesis is the process progressing downward from one of the higher levels to the bottom. Historically, speech technology originated from the bottom and has developed toward the extraction and handling of higher-level information. Some of the technology indicated in the figure remains to be investigated. Ultimate speech synthesis/recognition systems that are really useful and comfortable for users should match or exceed human capability. That is, they should be faster, more accurate, more intelligent, more knowledgeable, less expensive, and easier to use. For this purpose the ultimate systems must be able to handle conceptual information, the highest level of information in Fig. 2.

It is, however, neither necessary nor useful to try to use speech for every kind of input and output in computerized systems. Although speech is the fastest and easiest input and output means for simple exchange of information with computers, it is inferior to other means in conveying complex information. It is important to have an optimal division of roles and cooperation in a multimedia environment that includes images, characters, tactile signals, handwriting, etc. (5). HuMaNet, built by AT&T Bell Laboratories, is one such advanced experimental multimedia communication system (3).

From the human interface point of view, future computerized systems should be able to automatically acquire new knowledge about the thinking process of individual users, automatically correct user errors, and understand the intention of users by accepting rough instructions and inferring details. A hierarchical interface that initially uses figures and images (including icons) to express global information and

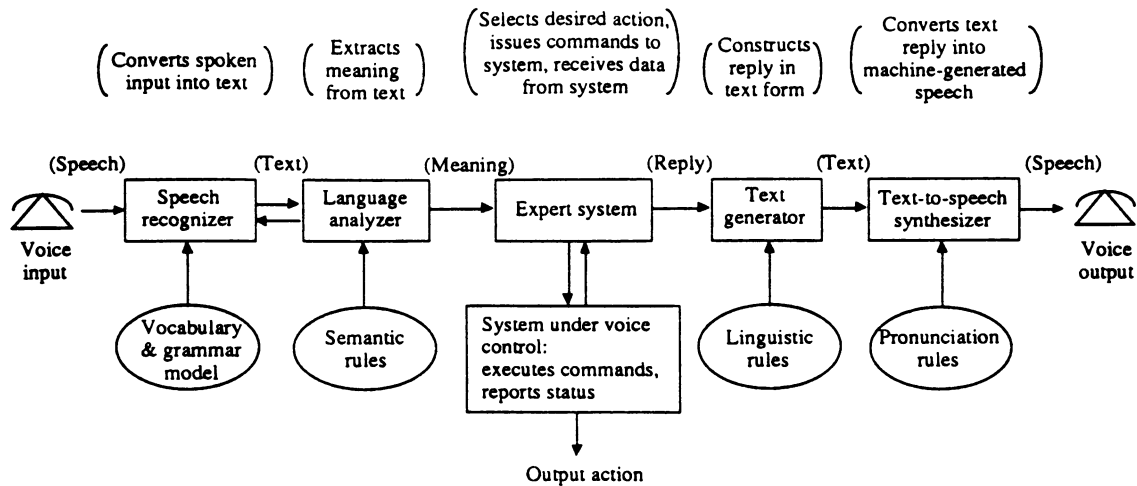


FIG. 1. Typical structure for task-specific voice control and dialogue systems. [Modified from Flanagan (3).]

then uses linguistic expression, such as spoken and written languages, for details would be a good interface that matches the human thinking process.

Ultimate communication systems are expected to use “virtual reality” technology. Fig. 3 shows the developing stages of the video and audio interfaces in teleconferencing. A teleconferencing system that uses virtual reality will become possible as these systems evolve from the present concentration type to the projection type and then to the three-dimensional type. In virtual reality systems the participants are not necessarily real human beings. They can be robots or electronic secretaries incorporating speech recognizers, synthesizers, and expert systems. It is interesting to consider the roles of speech synthesis and recognition technologies in these systems.

FUTURE SPEECH SYNTHESIZERS

Future speech synthesizers should have the following features:

- Highly intelligible (even under noisy and reverberant conditions and when transmitted over telephone networks)
- Natural voice sound
- Prosody control based on meaning

- Capable of controlling synthesized voice quality and choosing individual speaking style (voice conversion from one person’s voice to another, etc.)
- Multilingual, multidialectal
- Choice of application-oriented speaking styles, including rhythm and intonation (e.g., announcements, database access, newspaper reading, spoken e-mail, conversation)
- Able to add emotion
- Synthesis of voice from concepts

In present commercial speech synthesizers, voice quality can be selected from male, female, and children’s voices. No system has been constructed, however, that can precisely select or control the synthesized voice quality. Research into the mechanism underlying voice quality, inclusive of voice individuality, is thus needed to ensure that synthesized voice is capable of imitating a desired speaker’s voice or to select any voice quality such as harshness or softness.

For smooth and successful conversation between computerized systems and people by means of speech recognizers and synthesizers, how to prompt the users by synthesized commands or questions is crucially important. It has been reported that, when users are expected to respond to the

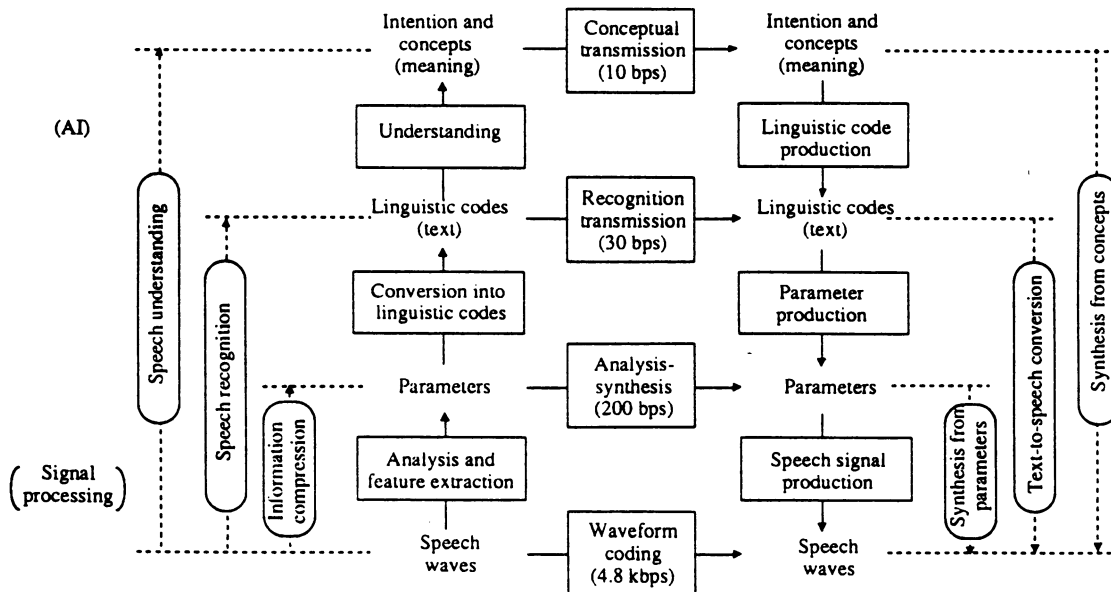


FIG. 2. Principal speech information-processing technologies and their relationship (4).

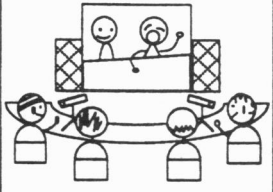
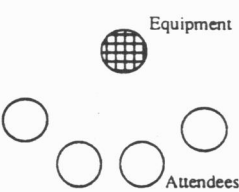
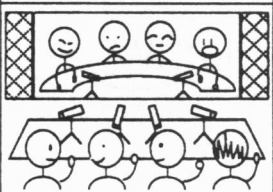
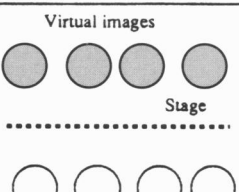
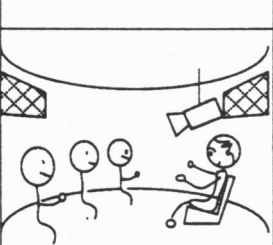
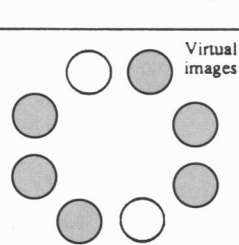
	Image	Configuration	Concept
Concentration type			One-dimensional Users are conscious of equipment
Projection type			Two-dimensional The state of the other party is reproduced on the stage
Three-dimensional type			Three-dimensional People in different places gather in a virtual conference space

FIG. 3. Evolution in teleconferencing (6).

machine with isolated words, the percentage of isolated-word responses depended strongly on the prompts made by the machine (7). It was also reported that the intonation of prompted speech as well as the content could strongly influence user responses.

FUTURE SPEECH RECOGNIZERS

Future speech recognition technology should have the following features:

- Few restrictions on tasks, vocabulary, speakers, speaking styles, environmental noise, microphones, and telephones
- Robustness against speech variations
- Adaptation and normalization to variations due to environmental conditions and speakers
- Automatic knowledge acquisition for phonemes, syllables, words, syntax, semantics, and concepts
- The ability to process discourse in conversational speech (e.g., to analyze context and accept ungrammatical sentences)
- Naturalness and ease of human-machine interaction
- Recognition of emotion (prosodic information)

Extraction and normalization of (adaptation to) voice individuality is one of the most important issues (8). A small fraction of people occasionally produce exceptionally low recognition rates (the so-called sheep and goats phenomenon). Speaker normalization (adaptation) methods can usually be classified into supervised (text-dependent) and unsupervised (text-independent) methods. Experiments have shown that people can adapt to a new speaker's voice after hearing just a few syllables, irrespective of the phonetic content of the syllables (9).

Automatic knowledge acquisition is very important in achieving systems that can automatically follow variations in tasks, including the topics of a conversation. Not only the linguistic structures but also the acoustic characteristics of a speech vary according to the task. Since it is impossible to

collect a large database for every kind of task, the recognizers should be able to automatically acquire new knowledge about these features and trace these changes (10–12).

Table 1 indicates broad projections for speech recognition technology that is/will become available in commercial systems in the next decade. The ultimate systems should be capable of robust speaker-independent or speaker-adaptive, continuous speech recognition. They should have no restrictions on vocabulary, syntax, semantics, or task. These systems will probably be made possible by implementing automatic learning systems. For the projections in the table to come about, we need continued research in many aspects of speech recognition technology.

The following are also important from the viewpoint of applications:

- Incentive for customers to use the systems
- Low cost
- Creation of new revenues for suppliers
- Cooperation on standards and regulation
- Quick prototyping and development

One of the most useful applications of speech recognition technology in telecommunication networks is the directory assistance service. For this application, systems based on recognizing spoken spelled names are being investigated at many laboratories. However, it is not easy for users to correctly spell the names of persons whose telephone numbers are unknown. In addition, there are several sets of letters having similar pronunciations, such as the English E-rhyme set, and pronunciation of the spelling of other persons' names is often unstable, since this is not a familiar task for us. Therefore, it is not easy for recognizers to correctly recognize alphabetically spelled names. A more successful approach might be to recognize naturally spoken names using the most advanced speech recognition technology, even if the machines have to recognize hundreds of thousands of names (14).

The requirements that future speaker recognizers should satisfy are similar to those for future speech recognizers. They include the following:

Table 1. Projections for speech recognition

Year	Recognition capability	Vocabulary size	Applications
1990	Isolated/connected words Whole-word models, word spotting, finite-state grammars, constrained tasks	10–30	Voice dialing, credit card entry, catalog ordering, inventory inquiry, transaction inquiry
1995	Continuous speech Subword recognition elements, stochastic language models	100–1,000	Transaction processing, robot control, resource management
2000	Continuous speech Subword recognition elements, language models representative of natural language, task-specific semantics	5,000–20,000	Dictation machines, computer-based secretarial assistants, database access
2000+	Continuous speech Spontaneous speech grammar, syntax, semantics; adaptation, learning	Unrestricted	Spontaneous speech interaction, translating telephony

Modified from Rabiner and Juang (13).

- Few restrictions on text, speaking style, environmental noise, microphones, and telephones
- Robustness against speech variations
- Adaptation and normalization to variations due to environmental conditions and speakers
- Automatic acquisition of speaker-specific characteristics
- Naturalness and ease of human-machine interaction
- Incentive for customers to use the systems
- Low-cost creation of new revenues for suppliers
- Cooperation on standards and regulation
- Quick prototyping and development

One of the most serious problems arises from variability in a person's voice. In speaker recognition there are always time intervals between training and recognition, and it is unrealistic to ask every user to utter a large amount of training data. Therefore, the variability problem is more serious for speaker recognition than for speech recognition (15). Speaker normalization (adaptation) and recognition methods should be investigated using a common approach. This is because they are two sides of the same problem: how best to separate the speaker information and the phoneme information in speech waves or how best to extract and model the speaker-specific phoneme information (16).

TOWARD ROBUST SPEECH/SPEAKER RECOGNITION UNDER ADVERSE CONDITIONS

As described in the previous sections, robustness against speech variations is one of the most important issues in speech/speaker recognition (17–20). Methods that are not robust in actual use cannot be considered authentic methods. There are many reasons for speech variations. Even the psychological awareness of communicating with a speech recognizer could induce a noticeable difference in the talker's

speech. The main causes of speech variation can be classified according to whether they originate in the speaking and recording environment, the speakers themselves, or the input equipment (Table 2; ref. 17).

Additive noises can be classified according to whether they are correlated or uncorrelated to speech. They can also be classified as stationary or nonstationary. The most typical nonstationary noises are other voices. Although various kinds of signal-processing methods have been proposed to suppress additive noises, we still need to develop more flexible and effective methods, especially for nonstationary noises.

Although the physical phenomena of variation can be classified as either noise addition or distortion, the distinction between these categories is not clear. When people speak in a noisy environment, not only does the loudness (energy) of their speech increase, but the pitch and frequency components also change. These speech variations are called the Lombard effect (21). Several experimental studies have indicated that these indirect influences of noise have a greater effect on speech recognition than does the direct influence of noise entering microphones (17).

Recognition performance under noisy conditions is often impaired by variations in the amount of speech-quality degradation rather than by the degradation itself. Problems are created, for example, by the variation of noise level associated with variations in the distance between the speaker and the microphone. To cope with these variations, it is essential to develop methods for automatically adapting to and normalizing these effects.

When recognizing spontaneous speech in dialogues, it is necessary to deal with variations that are not encountered when recognizing speech that is read from texts. These variations include extraneous words, out-of-vocabulary

Table 2. Main causes of speech variation

Environment	Speaker	Input equipment
Speech-correlated noise— reverberation, reflection	Attributes of speakers— dialect, gender, age	Microphone (transmitter) Distance to the microphone
Uncorrelated noise— additive noise (stationary, nonstationary)	Manner of speaking— breath and lip noise, stress, Lombard effect, rate, level, pitch, cooperativeness	Filter Transmission system— distortion, noise, echo Recording equipment

words, ungrammatical sentences, botched utterances, re-starts, repetitions, and style shifts. It is crucially important to develop robust and flexible parsing algorithms that match the characteristics of spontaneous speech. Instability in the detection of end points is frequently observed. Additionally, the system is required to respond to the utterance as quickly as possible. To solve these problems, it is necessary to establish a method for detecting the time at which sufficient information has been acquired instead of detecting the end of input speech. How to extract contextual information, predict users' responses, and focus on key words are very difficult and important issues.

Style shifting also is an important problem in spontaneous speech recognition. In typical laboratory experiments, speakers read lists of words rather than try to accomplish a real task. Users actually trying to accomplish a task, however, use a different linguistic style.

SPEECH AND NATURAL LANGUAGE PROCESSING

Speech (acoustic) processing and language processing have usually been investigated in isolation, and the technologies of these two areas have merely been combined to obtain a final decision in speech recognition and understanding. However, the methods produced from the results obtained from natural-language-processing research are not always useful in speech recognition. Therefore, it has recently become important to investigate new models that tightly integrate speech- and language-processing technology, especially for spontaneous speech recognition (22, 23).

These new models should be based on new linguistic knowledge and technology specific to speaking styles, which are very different from read speech. It will be necessary to properly adjust the methods of combining syntactic and semantic knowledge with acoustic knowledge according to the situation. How to extract and represent concepts in speech, that is, how to map speech to concepts, and how to use conceptual associations in recognition processes are important issues in linguistic processing for spontaneous speech (12).

Statistical language modeling, such as bigrams and trigrams, has been a very powerful tool, so it would be very effective to extend its utility by incorporating semantic knowledge. It will also be useful to integrate unification grammars and context-free grammars for efficient word prediction. Adaptation of linguistic models according to tasks and topics is also a very important issue, since collecting a large linguistic database for every new task is difficult and costly (24).

USE OF ARTICULATORY AND PERCEPTUAL CONSTRAINTS

Speech research is fundamentally and intrinsically supported by a wide range of sciences. The intensification of speech research continues to underscore an even greater interrelationship between scientific and technological interests (3). Although it is not always necessary or efficient for speech synthesis/recognition systems to directly imitate the human speech production and perception mechanisms, it will become more important in the near future to build mathematical models based on these mechanisms to improve performance (4, 10, 25).

For example, when sequences of phonemes and syllables are produced by human articulatory organs, such as tongue, jaw, and lips, these organs move in parallel, asynchronously, and yet systematically. Present speech analysis methods, however, convert speech signals into a single sequence of instantaneous spectra. It will become important to decom-

pose speech signals into multiple sources based on the concealed production mechanisms (26). This approach seems to be essential for solving the coarticulation problem, one of the most important problems in both speech synthesis and recognition.

Development of a new sound source model that precisely represents the actual source characteristics, as well as research on the mutual interaction between the sound source and the articulatory filter, would seem to be needed for faster progress in speech synthesis.

Psychological and physiological research into human speech perception mechanisms shows that the human hearing organs are highly sensitive to changes in sounds, that is, to transitional (dynamic) sounds, and that the transitional features of the speech spectrum and the speech wave play crucially important roles in phoneme perception (27). The length of the time windows in which transitions of sounds are perceived has a hierarchical structure and ranges from the order of several milliseconds to several seconds. The hierarchical layers correspond to various speech features, such as phonemes, syllables, and prosodic features. It has also been reported that the human hearing mechanism perceives a target value estimated from the transitional information extracted using dynamic spectral features.

The representation of the dynamic characteristics of speech waves and spectra has been studied, and several useful methods have been proposed. However, the performance of these methods is not yet satisfactory, and most of the successful speech analysis methods developed thus far assume a stationary signal. It is still very difficult to relate time functions of pitch and energy to perceptual prosodic information. Discovery of good methods for representing the dynamics of speech associated with various time lengths is expected to have a substantial impact on the course of speech research. This research is closely related to the analysis method based on the speech production mechanism described above.

The human hearing system is far more robust than machine systems—more robust not only against the direct influence of additive noise but also against speech variations (i.e., the indirect influence of noise), even if the noise is very inconsistent. Speech recognizers are therefore expected to become more robust when the front end uses models of human hearing. This can be done by imitating the physiological organs (28) or by reproducing psychoacoustic characteristics (29).

Basic speech units for speech synthesis and speech/speaker recognition should be studied from the following perspectives:

- Linguistic units (e.g., phonemes and syllables)
- Articulatory units (e.g., positions and motion targets for the jaw and tongue)
- Perceptual units (e.g., targets and loci of spectral movement and distinctive features)
- Visual units (features used in spectrogram reading)
- Physical units (e.g., centroids in vector/segment quantization)

These units do not necessarily correspond to each other. It will be important to establish new units based on combinations of these viewpoints.

Humans skillfully combine a wide variety of linguistic knowledge concerned with syntax and semantics according to the difficulty and characteristics of given sentences. It is necessary to investigate how to achieve these capabilities in speech recognition. The use of constraints imposed by articulatory and perceptual systems will also be useful for making speech synthesis/recognition systems more natural for the users.

EVALUATION METHODS

It is important to establish methods for measuring the quality of speech synthesis/recognition systems. Objective evaluation methods that ensure quantitative comparison of a broad range of techniques are essential to technological development in the speech-processing field. Evaluation methods can be classified into the following two categories (30):

- *Task evaluation*: creating a measure capable of evaluating the complexity and difficulty of tasks.
- *Technique evaluation*: formulating both subjective and objective methods for evaluating techniques and algorithms for speech processing.

Task evaluation is very important for speech recognition, since the performances of recognition techniques can be compared only when they are properly compensated for the difficulty of the task. Although several measures for task evaluation have already been proposed, such as word and phoneme perplexity, none of them is good enough at evaluating the difficulty in understanding the meanings of sentences. It may be very difficult to achieve a reliable measure for such purposes, since it involves quantifying all sources of linguistic variability. Nevertheless, we should try to accomplish this target step by step, by creating several measures, such as "meaning perplexity" and "concept perplexity," since these steps are highly related to the basic principles pertaining to modeling the meanings and concepts conveyed by speech.

Technique evaluation must take the viewpoint of improving the human-machine interface (31). Ease in human-machine interaction must be properly measured. Recognition systems having minimum recognition errors are not always the best. There are various trade-offs among the categories of errors, such as substitution, insertion, and deletion. Even if the error rate is relatively high, systems are acceptable if the error tendency is natural and matches the principles of human hearing and perception. It is crucially important that recognition errors are easy to correct and the system does not repeat the same errors.

To correctly evaluate technologies and to achieve steady progress, it is important to comprehensively evaluate a technique under actual field conditions instead of under a single controlled laboratory condition. Even if recognition systems perform remarkably well in laboratory evaluations and during demonstrations to prospective clients, they often do not perform nearly as well in the "real world." This is mainly because the speech that actually has to be recognized varies for many reasons, as mentioned above, and therefore usually differs from training speech. Recognition performance also generally varies with the motive and experience of users.

It was reported that people change their speaking styles when they notice that they are conversing with computers (32). This is another reason why it is important to develop experimental systems and test them under actual field conditions.

CONCLUSION

This paper predicts speech recognition and synthesis technology for the year 2001 and describes the most important research problems to be solved for accomplishing those ultimate recognition and synthesis systems. The problems include automatic knowledge acquisition, speaking style control in synthesis, synthesis from concepts, robust speech/speaker recognition, adaptation/normalization, language processing, use of articulatory and perceptual constraints, and evaluation methods.

One important issue that is not included in this paper is language identification. It is usually assumed that the language

of input speech for recognition, whether English, Japanese, French, etc., is known beforehand. However, in several cases, such as multilanguage interpreting systems, it is necessary to automatically identify the language of input speech. Methods that can be used for this purpose will probably be related to speaker recognition technology.

Although speech synthesis and recognition research have thus far been done independently for the most part, they will encounter increased interaction until commonly shared problems are investigated and solved simultaneously. Only then can we expect to witness tremendous speech research progress and hence the fruition of widely applicable beneficial techniques.

1. Wilpon, J. G. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 9991-9998.
2. Levitt, H. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 9999-10006.
3. Flanagan, J. L. (1991) *Proceedings of EUROSPEECH 91*, pp. 7-22.
4. Furui, S. (1989) *Digital Speech Processing, Synthesis, and Recognition* (Dekker, New York).
5. Cohen, P. R. & Oviatt, S. L. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 9921-9927.
6. Koizumi, N. (1991) *J. Inst. Television Eng. Jpn.* **45**, 474-479.
7. Basson, S. (1992) in *Proceedings of the COST 232 Workshop on Speech Recognition Over the Telephone Line* (Rome).
8. Furui, S. (1992) in *Advances in Speech Signal Processing*, eds. Furui, S. & Sondhi, M. M. (Dekker, New York), pp. 597-622.
9. Kato, K. & Furui, S. (1985) *IEICE Tech. Rep.* H85-5.
10. Atal, B. S. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 10046-10051.
11. Bates, M. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 9977-9982.
12. Marcus, M. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 10052-10059.
13. Rabiner, L. R. & Juang, B. H. (1993) *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).
14. Minami, Y., Shikano, K., Yamada, T. & Matsuoka, T. (1992) in *Proceedings of the IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, VII.1.
15. Rosenberg, A. E. & Soong, F. K. (1992) in *Advances in Speech Signal Processing*, eds. Furui, S. & Sondhi, M. M. (Dekker, New York), pp. 701-737.
16. Matsui, T. & Furui, S. (1993) in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis*, pp. II-391-394.
17. Furui, S. (1992) in *Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions, Cannes-Mandelieu*, pp. 31-42.
18. Juang, B. H. (1991) *Comput. Speech Lang.* **5**, 275-294.
19. Makhoul, J. & Schwartz, R. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 9956-9963.
20. Weinstein, C. J. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 10011-10016.
21. Junqua, J. C. (1992) in *Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions, Cannes-Mandelieu*, pp. 43-52.
22. Moore, R. C. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 9983-9988.
23. *Proceedings of the Speech and Natural Language Workshop* (Kaufmann, San Mateo, CA).
24. Kuhn, R. & DeMori, R. (1990) *IEEE Trans. Pattern Anal. Machine Intell., PAMI-12* **6**, 570-583.
25. Carlson, R. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 9932-9937.
26. Atal, B. S. (1983) in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Boston), 2.6.
27. Furui, S. (1986) *J. Acoust. Soc. Am.* **80**, 1016-1025.
28. Ghitza, O. (1992) in *Advances in Speech Signal Processing*, eds. Furui, S. & Sondhi, M. M. (Dekker, New York), pp. 453-485.
29. Hermansky, H. (1990) *J. Acoust. Soc. Am.* **87**, 1738-1752.
30. Furui, S. (1991) in *Proceedings of the Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assessment Methods* (Chiavari, Italy).
31. Kamm, C. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 10031-10037.
32. Wilpon, J. G. & Roe, D. B. (1992) in *Proceedings of the COST 232 Workshop on Speech Recognition Over the Telephone Line* (Rome).