Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 10023–10030, October 1995
Colloquium Paper

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.

What does voice-processing technology support today?

RYOHEI NAKATSU* AND YOSHITAKE SUZUKI

Nippen Telegraph and Telephone Corp., 3-9-11, Midori-cho, Musashino-shi, Tokyo 180, Japan

ABSTRACT This paper describes the state of the art in applications of voice-processing technologies. In the first part, technologies concerning the implementation of speech recognition and synthesis algorithms are described. Hardware technologies such as microprocessors and DSPs (digital signal processors) are discussed. Software development environment, which is a key technology in developing applications software, ranging from DSP software to support software also is described. In the second part, the state of the art of algorithms from the standpoint of applications is discussed. Several issues concerning evaluation of speech recognition/synthesis algorithms are covered, as well as issues concerning the robustness of algorithms in adverse conditions.

Recently, voice-processing technology has been greatly improved. There is a large gap between the present voice-processing technology and that of 10 years ago. The speech recognition and synthesis market, however, has lagged far behind technological progress. This paper describes the state of the art in voice-processing technology applications and points out several problems concerning market growth that need to be solved.

Technologies related to applications can be divided into two categories. One is system technologies and the other is speech recognition and synthesis algorithms.

Hardware and software technologies are the main topics for system development. Hardware technologies are very important because any speech algorithm is destined for implementation on hardware. Technology in this area is advancing quickly. Almost all speech recognition/synthesis algorithms can be used with a microprocessor and several DSPs. With the progress of device technology and parallel architecture, hardware technology will continue to improve and will be able to cope with the huge number of calculations demanded by improved algorithms of the future.

Also, software technologies are an important factor, as algorithms and application procedures should be implemented by the use of software technology. In this paper, therefore, software technology will be treated as an application development tool. Along with the growth areas of application of voice-processing technology, various architectures and tools that support applications development have been devised. Also, when speech processing is the application target, it is important to keep in mind the characteristics peculiar to speech. Speech communication basically is of a nature that it should work in a real-time interactive mode. Computer systems that handle speech communications with users should have an ability to cope with these operations. Several issues concerning real-time interactive communication will be described.

For algorithms there are two important issues concerning application. One is the evaluation of algorithms, and the other is the robustness of algorithms under adverse conditions. Evaluation of speech recognition and synthesis algorithms has been one of the main topics in the research area. However, to consider applications, these algorithms should be evaluated in real situations rather than laboratory situations, which is a new research trend.

Also, the robustness of algorithms is a crucial issue because conditions in almost all real situations are adverse, and algorithms should therefore be robust enough for the application system to handle these conditions well.

Finally, several key issues will be discussed concerning advanced technology and how its application can contribute to broadening the market for speech-processing technology.

SYSTEM TECHNOLOGIES

When speech recognition or synthesis technology is applied to real services, the algorithms are very important factors. The system technology—how to integrate the algorithms into a system and how to develop programs for executing specific tasks—is similarly a very important factor since it affects the success of the system. In this paper we divide the system technology into hardware technology and application—or software←technology and describe the state of the art in each of these fields.

Hardware Technology

Microprocessors. Whether a speech-processing system utilizes dedicated hardware, a personal computer, or a workstation, a microprocessor is necessary to control and implement the application software. Thus, microprocessor technology is an important factor in speech applications. Microprocessor architecture is categorized as CISC (Complex Instruction Set Computer) and RISC (Reduced Instruction Set Computer) (1). The CISC market is dominated by the Intel x86 series and the Motorola 68000 series. RISC architecture was developed to improve processing performance by simplifying the instruction set and reducing the complexity of the circuitry. Recently, RISC chips are commonly used in engineering workstations. Several common RISC chips are compared in Fig. 1. Performance of nearly 300 MIPS is available. The processing speed of CISC chips has fallen behind that of the RISC in recent years, but developers have risen to the challenge to improve the CISC's processing speed, as the graph of the performance of the x86 series in Fig. 2 shows. Better microprocessor performance makes it possible to carry out most speech-processing algorithms using standard hardware, whereas formerly dedicated hardware was necessary. In some applications, the

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.

*Present address: ATR Media Integration and Communications Research Laboratories, Seika-cho Soraku-gun, Kyoto, 619-02 Japan.

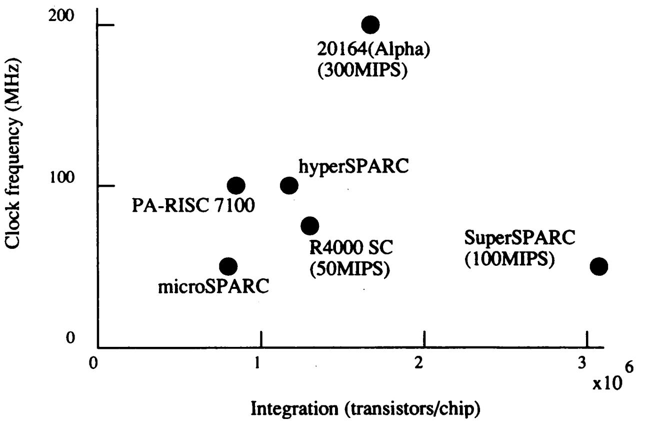

FIG. 1. Performance of RISC chip.

complete speech-processing operation can be carried out using only a standard microprocessor.

Digital Signal Processors. A DSP, or digital signal processor, is an efficient device that carries out algorithms for speech and image processing digitally. To process signals efficiently, the DSP chip uses the following mechanisms: high-speed floating point processing unit, pipeline multiplier and accumulator, and parallel architecture of arithmetic-processing units and address calculation units.

DSPs of nearly 50 MFLOPS are commercially available. Recent complicated speech-processing algorithms need broad dynamic range and high precision. The high-speed floating point arithmetic unit has made it possible to perform such processing in minimal instruction cycles without loss of calculation precision. Increased on-chip memory has made it possible to load the whole required amount of data and access internally for the speech analysis or pattern-matching calculations, which greatly contributes to reducing instruction cycles. Usually more cycles are needed to access external memory than are needed for accessing on-chip memory. This is a bottleneck to attaining higher throughput. The amount of memory needed for a dictionary for speech recognition or speech synthesis is too large to implement on-chip though; several hundred kilobytes or several megabytes of external memory are required to implement such a system.

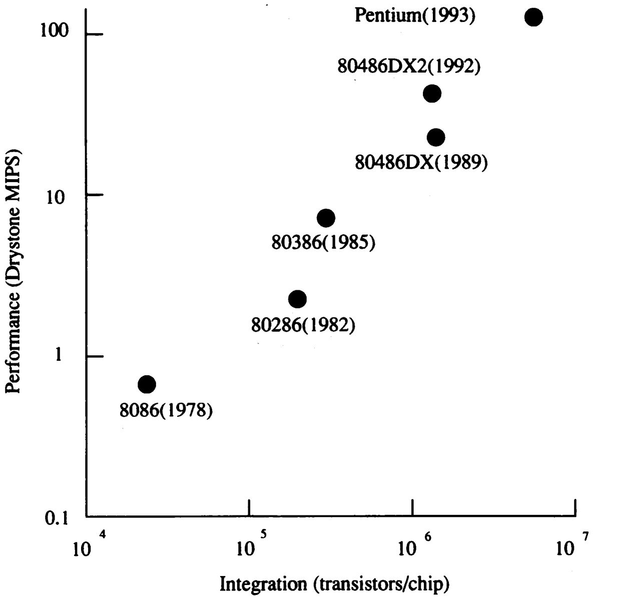

FIG. 2. Performance improvement of ×86 microprocessor.

Most traditional methods of speech analysis, as well as speech recognition algorithms based on the HMM (hidden Markov model), can be carried out in real time by a single DSP chip and some amount of external memory. On the other hand, recent speech analysis, recognition, and synthesis algorithms have become so complex and time consuming that a single DSP cannot always complete the processing in real time. Consequently, to calculate a complex algorithm in real time, several DSP chips work together using parallel processing architecture.

Equipment and Systems. Speech-processing equipment or a speech-processing system can be developed using either microprocessors or DSPs. They can use one of the following architectures: (a) dedicated, self-contained hardware that can be controlled by a host system, (b) a board that can be plugged into a personal computer or a workstation, and (c) a dedicated system that includes complete software and other necessities for an application.

In the early days of speech recognition and synthesis, devices of type (a) were developed because the performance of microprocessors and DSPs was not adequate for real-time speech processing (and were not so readily available). Instead, the circuits were constructed using small-scale integrated circuits and wired logic circuits. Recently, however, DSPs are replacing this type of system. A type (a) speech-processing application system can be connected to a host computer using a standard communication interface such as the RS-232C, the GPIB, or the SCSI. The host computer executes the application program and controls the speech processor as necessary. In this case the data bandwidth is limited and is applicable only to relatively simple processing operations.

As for case (b), recent improvements of digital device technology such as those shown in Figs. 1 and 2 have made it possible to install a speech-processing board in the chassis of the increasingly popular personal computers and workstations. The advantages of this board-type implementation are that speech processing can be shared between the board and a host computer or workstation, thus reducing the cost of speech processing from that using self-contained equipment; speech application software developed for a personal computer or workstation equipped with the board operates within the MS-DOS or UNIX environment, making it easier and simpler to develop application programs; and connecting the board directly to the personal computer or workstation bus allows the data bandwidth to be greatly widened, which permits quick response—a crucial point for a service that entails frequent interaction with a user.

In some newer systems only the basic analog-to-digital and digital-to-analog conversions are performed on the board, while the rest of the processing functions are carried out by the host system.

Method (c) is adopted when an application is very large. The speech system and its specific application system are developed as a package, and the completed system is delivered to the user. This method is common for telecommunications uses, where the number of users is expected to be large. In Japan a voice response and speech recognition system has been offered for use in public banking services since 1981. The system is called ANSER (Automatic answer Network System for Electrical Request). At first the system had only the voice response function for Touch-Tone telephones. Later, a speech recognition function was added for pulse, or rotary-dial, telephone users. After that, facsimile and personal computer interface capabilities were added to the system, and ANSER is now an extensive system (2) that processes more than 30 million transactions per month, approximately 20 percent of which are calls from rotary-dial telephones. The configuration of the ANSER system is shown in Fig. 3.

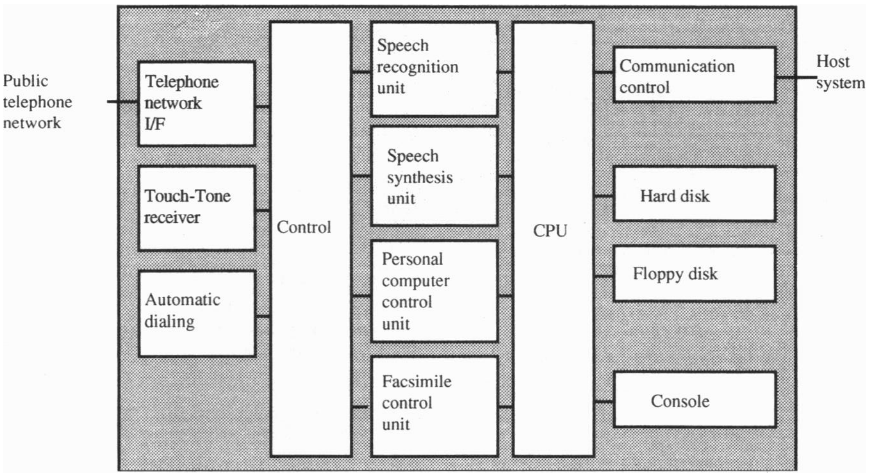

FIG. 3. Configuration of the ANSER system (3).

Application Technology Trend

Development Environment for DSP. As mentioned above, a DSP chip is indispensable to a speech-processing system. Therefore, the development environment for a DSP system is a critical factor that affects the turnaround time of system development and eventually the system cost. In early days only an assembler language was used for DSPs, so software development required a lot of skill. Since the mid-1980s, the introduction of high-level language compilers (the C cross compiler is an example) made DSP software development easier. Also, some commonly used algorithms are offered as subroutine libraries to reduce the programmer's burden. Even though the cross compilers have become more popular these days, software programming based on an assembler language may still be necessary in cases where real-time processing is critical.

Application Development Environment. Environments in application development take various forms, as exemplified by the varieties of DSP below:

(a) Application software is developed in a particular language dedicated to a speech system.

(b) Application software is developed using a high-level language such as C.

(c) Application software is developed using an "application builder," according to either application flow or application specifications.

Application software used to be written for each system using assembler language. Such cases are rare recently because of the inadequacy of a low-level programming language for developing large and complex systems. However, assembler language is still used in special applications, such as games or telephones, that require compact and high-speed software modules.

Recently, applications have usually been described by a high-level language (mainly C). Speech recognition and synthesis boards are commonly plugged into personal computers, and interface subroutines with these boards can be called from an application program written in C.

As an application system becomes larger, the software maintenance becomes more complicated; the user interface becomes more important; and, eventually, control of the speech system becomes more complicated, the specifications become more rigid, and improvement becomes more difficult. Therefore, the application development should be done in a higher-level language. Moreover, it is desirable for an "Application Builder" to adhere to the following principles: automatic program generation from a flowchart, automatic generation of an application's grammar, and graphical user interface for software development. Such an environment is called "Application Builder." Development of the Application Builder itself is an important and difficult current theme of automatic software generation research. So far, few systems have been implemented using the Application Builder, though some trials have been done in the field of speech-processing systems (4).

Speech Input/Output Operating Systems. Availability of a speech input/output function on a personal computer or workstation is a requisite for popularizing the speech input/output function. It is adequate to implement the speech input/output function at the operating system level. There are two ways: (a) developing a dedicated operating system with speech input/output function and (b) adding a speech input/output function as a preprocessor to an existing operating system.

Method (a) best utilizes speech recognition and synthesis capabilities. With this method a personal computer or workstation system with a speech input/output function can be optimally designed. However, no operating system that is specific to speech input/output has yet been developed. In the field of character recognition, on the other hand, pen input specific operating systems—called "Pen OS"—have begun to appear. These operating systems are not limited to handwritten character recognition; they are also capable of pen-pointing and figure command recognition. In other words, character recognition is not the primary function but just one of the functions. This grew out of the realization that recognition of handwritten characters is not yet satisfactory and that handwriting is not necessarily faster than keyboard input. Optimum performance was achieved by combining several data input functions such as keyboard input, mouse clicking, or pen input. This principle should be applied to the development of speech operating systems.

In method (b) there is no need to develop a new operating system. For example, a Japanese kana-kanji preprocessor, which converts keyboard inputs into 2-byte Japanese character codes, is widely used for Japanese character input. The same concept can be applied to speech input/output. An advantage of this method is that speech functions can easily be added to popular operating systems such as MS-DOS.

Real-Time Application Support. Real-time dialogue-processing function is a future need in speech applications. The speech system must carry out an appropriate control of the speech recognizer and the speech synthesizer to realize the real-time dialogue between humans and computers. Also, an easy method to describe this control procedure needs to be developed and opened to users so that they can develop dialogue control software easily. These points would be the main problems when developing an application system in the near future. The state of the art and the required technologies for the real-time application system are described below.

Barge in. "Barge in" means an overlap or collision of utterances, which frequently occurs in human speech conversation. This phenomenon is actually an important factor in making the communication smooth. The time delay on an international telecommunication line makes the conversation awkward because it interferes with this barge in. It is pointed out that the same phenomenon readily occurs in dialogues between humans and computers. The following techniques allow for barge in in a speech system:

- The system starts recognition as soon as it starts sending a voice message. An echo canceller can be applied to subtract the synthesizer's voice and to maintain recognition performance.
- The system may start sending a voice message as soon as the key word is extracted even if the speaker's voice has not been completed yet.

Although the former function has been implemented in some systems, no application has yet incorporated the latter function. This function can be carried out in a small-vocabulary system, so the human-machine interface should be researched in this case.

Key word extraction. Even in applications based on word recognition, humans tend to utter extra words or vocal sounds and sometimes even entire sentences. When the recognition vocabulary is limited, a simple key word extraction or spotting technique is used. (If the vocabulary size is huge or unlimited, the speech must be recognized in its entirety.) Several word-spotting algorithms have been proposed (5, 6) and proven efficient for extracting only key words.

ALGORITHMS

Recognition of large spoken vocabularies and understanding of spontaneous spoken language are studied eagerly by many speech recognition researchers. Recent speech synthesis research focuses on improvement of naturalness and treatment of prosodic information. The state of the art of speech recognition/synthesis technologies is described further elsewhere in this volume in papers by Carlson (7) and Makhoul and Schwartz (8). In most of these studies, rather clean and closed speech data are usually used for both training and evaluation. However, field data used in developing applications are neither clean nor closed. This leads to various problems, which are described below:

- *Databases.* A speech database that includes field data rather than laboratory data is necessary both for training and evaluating speech recognition systems.
- *Algorithm assessment.* Evaluation criteria other than those that have been used to undertake algorithm evaluation in the laboratory should be used to assess the feasibility of speech recognition/synthesis algorithms and systems to be used in real applications.
- *Robustness of algorithms.* In a real environment a broad spectrum of factors affect speech. Speech recognition/synthesis algorithms should be robust under these varied conditions.

These topics will be discussed in the following sections.

Databases

Databases for Research. Currently, in speech recognition research, statistical methods such as the HMM are widely used. One of the main characteristics of a statistical method is that its performance generally depends on the quantity and quality of the speech database used for its training. The amount of speech data collected is an especially important factor for determining its recognition performance. Because construction of a large database is too big a job for a single researcher or even a single research institute, collaboration among speech researchers to construct databases has been very active. In the United States, joint speech database construction is undertaken in Spoken Language programs supported by DARPA (Defense Advanced Research Projects Agency) (9, 10). In Europe, under several projects such as ESPRIT (European Strategic Program for Research and Development in Information Technology), collaborative work has been done for constructing large speech databases (11, 12). Also, in Japan large speech databases are under construction by many researchers at various institutes (13).

For speech synthesis, on the other hand, concatenation of context-dependent speech units has recently proven efficient for producing high-quality synthesized speech. In this technology selection of the optimum unit for concatenation from a large set of speech units is essential. This means that a large speech database is necessary in speech synthesis research also, but a trend toward collaborative database construction is not yet apparent in this area. The main reason is that the amount of speech data needed for speech synthesis is far less than that for speech recognition because the aim of speech synthesis research is not to produce various kinds of synthesized speech.

Databases for Application. The necessity of a speech database is not yet being discussed from the standpoint of application. Two reasons are as follows:

- The application area both for speech recognition and speech synthesis is still very limited. Accordingly, there is not a strong need for constructing speech databases.
- Applications are tied closely to business. This means that vendors or value-added resellers who are constructing speech databases do not want to share the information in their databases.

For telephone applications, on the other hand, disclosure of information about databases or the databases themselves is becoming common. In this area, applications are still in the early stages and competition among vendors is not strong yet. Another reason is that in telephone applications the amount of speech database necessary for training a speaker-independent speech recognition algorithm is huge, which has led to a consensus between researchers that collaboration on database construction is necessary.

In the United States, Texas Instruments is trying to construct a large telephone speech database that is designed to provide a statistically significant model of the demographics of the U.S. population (14). Also, AT&T is actively trying to adapt word-spotting speech recognition technology to various types of telephone services. For that purpose, the company has been collecting a large amount of speech data through telephone lines (15). Several other institutes are also constructing various kinds of telephone speech databases (16, 17). In Europe, various trial applications of speech recognition technology to telephone service are under way (18, 19). In Japan, as discussed earlier, the ANSER speech recognition and response system has been widely used for banking services since the beginning of the 1980s. For training and evaluating the speech recognition algorithm used in the ANSER system, a large speech database consisting of utterances of more than 3000 males and females ranging in age from 20 to 70 was constructed (20).

As telephone service has been considered a key area for the application of speech recognition technology, various trials have been undertaken and various issues are being discussed. Several comments on these issues, based on the experiences of developing the ANSER system and participating in the operation of the banking service, are given below.

Simulated Telephone Lines. Because of the difficulties underlying the collection of a large speech corpus through telephone lines, there has been a discussion that a synthetic telephone database, constructed by passing an existing speech

Table 1. Milestones of speech recognition in ANSER

		Training data		
		Line	No. speakers	Region
Spring 1981	Service started in Tokyo area	Pseudo-telephone lines	500	Tokyo
Autumn 1981	Service recognizer retrained	Telephone line	500	Tokyo
Spring 1982	Service area widened to Osaka and Kyushu areas	Telephone line	500	Tokyo
Autumn 1982	Speech recognizer retrained	Telephone line	1500	Tokyo, Osaka, Kyushu

database through a simulated telephone line, could be used as an alternative (21), and a prototype of such a simulated telephone line has been proposed (22).

Table 1 summarizes the use of speech data to improve service performance in the ANSER system. In the first stage of the ANSER services, a speech database from simulated telephone lines was used. This database was replaced by real telephone line data because recognition based on the simulated telephone data was not reliable enough. From these experiences it was concluded that simulated telephone data are not appropriate for real telephone applications. The main reasons are as follows:

- Characteristics of real telephone lines vary greatly depending on the pass selected. It is difficult, therefore, to simulate these variations using a simulated line.
- In addition to line characteristics, various noises are added to speech. Again, it is difficult to duplicate these noises on a simulated line.

To overcome the difficulty of collecting a large speech corpus through real telephone lines, NYNEX has constructed a large database, called NTIMIT (23), by passing recorded speech through a telephone network.

Dialects. It is considered important for speech data to include the various dialects that will appear in real applications. This dialect factor was taken into consideration in the construction of several databases (14, 15). The ANSER service used speech samples collected in the Tokyo area when the real telephone line database was introduced. As the area to which banking service is offered expanded to include the Osaka and Kyushu areas, it was pointed out that the performance was not as good in these areas as it was in the Tokyo area. In response to the complaints, telephone data were collected in both the Osaka and Kyushu areas. The speech recognizer was retrained using all of the utterances collected in all three areas, and recognition performance stabilized.

Assessment of Algorithms

Assessment of Speech Recognition Algorithms. In the early days of speech recognition, when word recognition was the research target, the word recognition rate was the criterion for assessment of recognition algorithms. Along with the shift of research interest from word recognition to speech understanding, various kinds of assessment criteria, including linguistic processing assessment, have been introduced. These assessment criteria are listed in Table 2. Regarding application, the following issues are to be considered.

Assessment Criteria for Applications. In addition to the various criteria used for the assessment of recognition methods at the laboratory level, other criteria should be introduced in real applications. Some of these assessment criteria are listed in Table 3. Recently, various kinds of field trials for telephone speech recognition have been undertaken in which several kinds of assessment criteria are used (3, 18, 24, 25). Among these criteria, so far, the task completion rate (TCR) is considered most appropriate (24, 26). This matches the experience with the ANSER service. After several assessment criteria had been measured, TCR was determined to be the best criterion. TCR was under 90 percent, and there were many complaints from customers. As TCR exceeded 90 percent, the number of complaints gradually diminished, and complaints are rarely received now that the TCR exceeds 95 percent.

Assessment Using the "Wizard of Oz" Technique. When application systems using speech recognition technology are to be improved in order to raise user satisfaction, it is crucial to pinpoint the bottlenecks. The bottlenecks could be the recognition rate, rejection reliability, dialogue design, or other factors. It is nearly impossible to create a system to evaluate the bottlenecks totally automatically because there are too many parameters and because several parameters, such as recognition rate, are difficult to change. The "Wizard of Oz" (WOZ) technique might be an ideal alternative to this automatic assessment system. When the goal is a real application, reliable preassessment based on the WOZ technique is recommended. One important factor to consider is processing time. When using the WOZ technique, it is difficult to simulate real-time processing because humans play the role of speech recognizer in the assessment stage. Therefore, one must factor in the effect of processing time when evaluating simulation results.

Assessment of Speech Synthesis Technology. Speech synthesis technology has made use of assessment criteria such as phoneme intelligibility or word intelligibility (27). What is different from speech recognition technology is that almost all the assessment criteria that have been used are subjective criteria. Table 4 summarizes these subjective and objective assessment criteria. As environments have little effect on synthesized speech, the same assessment criteria that have been used during research can be applied to real use. However, as applications of speech synthesis technology rapidly diversify, new criteria for assessment by users arise:

a. In some information services, such as news announcements, customers have to listen to synthesized speech for lengthy periods, so it is important to consider customers' reactions to listening to synthesized speech for long periods.

b. More important than the intelligibility of each word or sentence is whether the information or meaning is conveyed to the user.

Table 2. Assessment criteria for speech recognition

Phoneme level	Word level	Sentence level
Correct segmentation rate Phoneme recognition rate	Isolated word: Word recognition rate Connected word: Word recognition rate (including insertion, deletion, and substitution) Word sequence recognition rate Key-word extraction rate	Word recognition rate (after linguistic processing) Sentence recognition rate Correct answer rate

Table 3. Assessment criteria of speech recognition from user's side

Objective criteria	Subjective criteria
Task completion rate	Satisfaction rating
Task completion time	Fatigue rating
Number of interactions	Preference
Number of error correction sequences	

Table 4. Assessment criteria of speech synthesis

Intelligibility	Naturalness
Phoneme intelligibility	Preference score
Syllable intelligibility	MOS
Word intelligibility	
Sentence intelligibility	

c. In interactive services such as reservations, it is important for customers to realize from the beginning that they are interacting with computers, so for some applications synthesized speech should not only be intelligible but should also retain a synthesized quality.

Several studies have been done putting emphasis on the above points. In one study, the rate of conveyance of meaning was evaluated by asking simple questions to subjects after they had listened to several sentences (28). Another study assessed the effects on listener fatigue of changing various parameters such as speed, pitch, sex, and loudness during long periods of listening (29).

Robust Algorithms

Classification of Factors in Robustness. Robustness is a very important factor in speech-processing technology, especially in speech recognition technology (30). Speech recognition algorithms usually include training and evaluation procedures in which speech samples from a database are used. Of course, different speech samples are used for training procedures than for evaluation procedures, but this process still contains serious flaws from the application standpoint. First, the same recording conditions are used throughout any one database. Second, the same instruction is used for all speakers in any one database.

This means that speech samples contained in a particular database have basically similar characteristics. In real situations, however, speakers' environments vary. Moreover, speech characteristics also tend to vary depending on the environments. These phenomena easily interfere with recognition. This is a fundamental problem in speech recognition that is hard to solve by algorithms alone.

Robustness in speech recognition depends on development of robust algorithms to deal with overlapping variations in speech. Factors that determine speech variations are summarized in Table 5.

Environmental Variation. Environmental variations that affect recognition performance are distance between the speech source and the microphone, variations of transmission characteristics caused by reflection and reverberation, and microphone characteristics.

Distance Between Speaker and Microphone. As the distance between the speaker and the microphone increases, recognition performance tends to decrease because of degradation of low-frequency characteristics. Normalization of transmission characteristics by using a directional microphone or an array microphone has been found to be effective in compensating for these phenomena (31).

Reflection and reverberation. It has been known that the interference between direct sound and sound reflected by a desk or a wall causes sharp dips in the frequency. Also, in a closed room the combination of various kinds of reflection causes reverberation. As application of speech recognition to conditions in a room or a car has become a key issue, these phenomena are attracting the attention of speech recognition researchers, and several studies have been done. Use of a directional microphone, adoption of an appropriate distance measure, or introduction of adaptive filtering are reported to be effective methods for preventing performance degradation (32, 33).

Microphone Characteristics. Each microphone performs optimally under certain conditions. For example, the frequency of a close-range microphone flattens when the distance to the mouth is within several centimeters. Accordingly, using different microphones in the training mode and the recognition mode causes performance degradation (32, 34). Several methods have been proposed to cope with this problem (35, 36).

Noise. All sounds other than the speech to be recognized should be considered noise. This means that there are many varieties of noise, from stationary noise, such as white noise, to environmental sounds such as telephone bells, door noise, or speech of other people. The method of compensating varies according to the phenomenon.

For high-level noise such as car noise or cockpit noise, noise reduction at the input point is effective. A head-mounted noise-canceling microphone or a microphone with sharp directional characteristics is reported effective (37, 38). Also, several methods of using microphone arrays have been reported (39, 40).

If an estimation of noise characteristics is possible by some means such as use of a second microphone set apart from the speaker, it is possible to reduce noise by calculating a transverse filter for the noise characteristics and applying this filter to the input signal (41, 42).

An office is a good target for diversifying speech recognition applications. From the standpoint of robustness, however, an office does not provide satisfactory conditions for speech recognition. Offices are usually not noisy. Various kinds of sounds such as telephone bells or other voices overlap this rather calm environment, so these sounds tend to mask the speech to be recognized. Also, a desk-mounted microphone is

Table 5. Factors determining speech variation

Environment	Noise	Speaker
Reflection	Stationary or	Interspeaker variation
Reverberation	quasi-stationary	Dialect
Distance to microphone	noise	Vocal tract
Microphone	White noise	characteristics
characteristics	Car noise	Speaking manner
	Air-conditioner noise	Coarticulation
	Nonstationary noise	Intraspeaker variation
	Other voices	Emotion
	Telephone bells	Stress
	Printer noise	Lombard effect

preferable to a close-range microphone from the human interface standpoint. Use of a comb filter has been proposed for separating object speech from other speech (43).

Speaker Variation. There are two types of speaker variation. One is an interspeaker variation caused by the differences among speakers between speech organs and differences in speaking manners. The other is intraspeaker variation. So far, various kinds of research related to interspeaker variations have been reported because this phenomena must be dealt with in order to achieve speaker-independent speech recognition. Intraspeaker variations, however, have been regarded as small noise overlapping speech and have been largely ignored. When it comes to applications, however, intraspeaker variation is an essential speech recognition factor.

Human utterances vary with the situation. Among these variations, mannerisms and the effects of tension, poor physical condition, or fatigue are difficult to control. Therefore, speech recognition systems must compensate for these variations.

Several studies of intraspeaker variation have been undertaken. One typical intraspeaker variation is known as the "Lombard effect," which is speech variation caused by speaking under very noisy conditions (44). Also, in several studies utterances representing various kinds of speaking mannerisms were collected, and the HMM was implemented to recognize these utterances (45, 46).

SPEECH TECHNOLOGY AND THE MARKET

As described in this paper, practical speech technologies have been developing rapidly. Applications of speech recognition and speech synthesis in the marketplace, however, have failed to keep pace with the potential. In the United States, for example, although the total voice-processing market at present is over $1 billion, most of this is in voice-messaging services. The current size of the speech recognition market is only around $100 million, although most market research in the 1980s forecasted that the market would soon reach $1 billion (47). And the situation is similar for speech synthesis. In this section the strategy for market expansion is described, with emphasis on speech recognition technology.

Illusions About Speech Recognition Technology

In papers and surveys on speech recognition, statements such as the following have been used frequently:

> "Speech is the most convenient method of communication for humans, and it is desirable to achieve oral communication between humans and computers."
> "Speech recognition is now mature enough to be applied to real services."

These statements are basically correct. However, when combined, they are likely to give people the following impression:

> "Speech recognition technology is mature enough to enable natural communication between computers and humans."

Of course, this is an illusion, and speech researchers or vendors should be careful not to give users this fallacious impression. The truth could be stated more precisely as follows:

a. The capacity to communicate orally is a fundamental human capability, which is achieved through a long learning process that begins at birth. Therefore, although the technology for recognizing natural speech is advancing rapidly, there still exists a huge gap between human speech and the speech a computer is able to handle.

b. Nevertheless, speech recognition technology has reached a level where, if applications are chosen appropriately, they can provide a means for communication between computers and humans that—although maybe not natural—is at least acceptable.

Strategy for Expanding the Market

Market studies carried out by the authors and others have identified the following keys to expansion of the speech recognition market, listed in descending order of importance (47, 48):

- *Applications and marketing*. New speech recognition applications must be discovered.
- *Performance*. Speech recognition algorithms must perform reliably even in real situations.
- *Capabilities*. Advanced recognition capabilities such as continuous speech recognition must be achieved.

Based on the above results and also on our experiences with developing and operating the ANSER system and service, the following is an outline of a strategy for widening the speech recognition market.

Service Trials. Because practical application of speech recognition to real services is currently limited to word recognition, which is so different from how humans communicate orally, it is difficult for both users and vendors to discover appropriate new applications. Still it is necessary for vendors to try to offer various kinds of new services to users. Although many of them might fail, people would come to recognize the capabilities of speech recognition technology and would subsequently find application areas suited to speech recognition.

Telephone services might be the best, because they will offer speech recognition functions to many people and help them recognize the state of the art of speech recognition. Also, as pointed out earlier, the interesting concept of a "Speech OS" is worth trying.

Robustness Research. It is important for speech recognition algorithms to be made more robust for real situations. Even word recognition with a small vocabulary, if it worked in the field as reliably as in the laboratory, would have great potential for application. It is delightful that recently the importance of robustness has attracted the attention of speech researchers and that various kinds of research are being undertaken, as described in the previous section.

One difficulty is that the research being done puts too much emphasis on stationary or quasi-stationary noise. There are tremendous variations of noise in real situations, and these real noises should be studied. Also, as stated before, intraspeaker speech variation is an important factor to which more attention should be paid.

Long Term Research. At the same time, it is important to continue research of speech recognition to realize more natural human-machine communication based on natural conversation. This will be long-term research. However, as oral communication capability arises from the core of human intelligence, basic research should be continued systematically and steadily.

CONCLUSION

This paper has briefly described the technologies related to speech recognition and speech synthesis from the standpoint of practical application.

First, systems technologies were described with reference to hardware technology and software technology. For hardware technology, along with the rapid progress of technology, a large amount of speech processing can be done by personal computer or workstation with or without additional hardware dedicated to speech processing. For software, on the other hand, the environment for software development has improved in recent years. Still further endeavor is necessary for vendors to pass these improvements on to end users so they can develop application software easily.

Then, several issues relating to the practical application of speech recognition and synthesis technologies were discussed. Speech databases for application and evaluation of these technologies were described. So far, because the range of applications of these technologies is limited, criteria for assessing the applications are not yet clear. Robustness of algorithms applied to field situations also was described. Various studies are being done concerning robustness.

Finally, reasons for the slow development of the speech recognition/synthesis market were discussed, and future directions for researchers and vendors to explore were proposed.

1. Patterson, D. & Ditzel, D. (1980) *Comput. Archit. News* **8,** 25–33.
2. Nakatsu, R. (1990) *IEEE Comput.* **23,** 43–48.
3. Nakatsu, R. & Ishii, N. (1987) in *Proceedings of Speech Tech '87* (Media Dimensions, New York), pp. 168–172.
4. Renner, T. (1992) in *Proceedings of the First IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* (IEEE, New York), Session 4.
5. Kimura, T., *et al.* (1987) *Proc. IEEE ICASSP-87*, pp. 1175–1178.
6. Wilpon, J. G., *et al.* (1990) *IEEE Trans. Acoust. Speech Signal Process.* **38,** No. 11, 1870–1878.
7. Carlson, R. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9932–9937.
8. Makhoul, J. & Schwartz, R. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9956–9963.
9. MADCOW (Multi-Site ATIS Data Collection Working Group) (1992) *Proceedings of the Speech and Natural Language Workshop* (Kaufmann, San Mateo, CA) pp. 7–14.
10. Pallet, D., *et al.* (1991) Unpublished National Institute of Standards and Technology document.
11. Carre, R., *et al.* (1984) *Proc. IEEE ICASSP-84*, 42.10.
12. Gauvain, J. L., *et al.* (1990) in *Proceedings of the International Conference on Spoken Language Processing*, pp. 1097–1100.
13. Itahashi, S. (1990) in *Proceedings of the International Conference on Spoken Language Processing*, pp. 1081–1084.
14. Picone, J. (1990) *Proc. IEEE ICASSP-90*, pp. 105–108.
15. Jacobs, T. E., *et al.* (1992) in *Proceedings of the First IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* (IEEE, New York).
16. Cole, R., *et al.* (1992) in *Proceedings of the International Conference on Spoken Language Processing*, pp. 891–893.
17. Pittrelli, J. F., *et al.* (1992) in *Proceedings of the First IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* (IEEE, New York).
18. Rosenbeck, P. & Baungaard, B. (1992) in *Proceedings of the First IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* (IEEE, New York).
19. Walker, G. & Millar, W. (1989) in *Proceedings of the ESCA Workshop 2*, p. 10.
20. Nomura, T. & Nakatsu, R. (1986) *Proc. IEEE ICASSP-86*, pp. 2687–2690.
21. Rosenbeck, P. (1992) in *Proceedings of the Workshop on Speech Recognition over the Telephone Line, European Cooperation in the Field of Scientific and Technical Research*.
22. Yang, K. M. (1992) in *Proceedings of the First IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* (IEEE, New York).
23. Jankowski, C., *et al.* (1990) *Proc. IEEE ICASSP-90*, pp. 109–112.
24. Chang, H. M. (1992) in *Proceedings of the First IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* (IEEE, New York).
25. Sprin, C., *et al.* (1992) in *Proceedings of the First IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* (IEEE, New York).
26. Neilsen, P. B. & Kristernsen, G. B. (1992) in *Proceedings of the First IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* (IEEE, New York).
27. Steeneken, H. J. M. (1992) in *Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions*, pp. 1–10.
28. Hakoda, K., *et al.* (1986) in *Records of the Annual Meeting of the IEICE* (Institute of Electronics Information and Communication Engineers, Tokyo).
29. Kumagami, K., *et al.* (1989) in *Technical Report SP89-68 of the Acoustical Society of Japan*.
30. Furui, S. (1992) in *Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions*, pp. 31–42.
31. Tobita, M., *et al.* (1989) in *Proceedings of the IEEE Pacific Rim Conference on Communications, Computer and Signal Processing*, pp. 631–634.
32. Tobita, M., *et al.* (1990) in *Technical Report SP90-20 of the Acoustical Society of Japan*.
34. Acero, A. & Stern, R. M. (1990) *Proc. IEEE ICASSP-90*, pp. 849–852.
35. Acero, A. & Stern, R. M. (1991) *Proc. IEEE ICASSP-91*, pp. 893–896.
36. Tobita, M., *et al.* (1990) *Trans. IEICE Japan* (Institute of Electronics, Information, and Communication Engineers, Tokyo), Vol. J73 D-II, No. 6, pp. 781–787.
37. Starks, D. R. & Morgan, M. (1992) in *Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions*, pp. 195–198.
38. Viswanathan, V., *et al.* (1986) *Proc. IEEE ICASSP-86*, pp. 85–88.
39. Kaneda, Y. & Ohga, J. (1986) *IEEE Trans. Acoust. Speech Signal Process.* **6,** 1391–1400.
40. Silverman, H., *et al.* (1992) in *Proceedings of the Speech and Natural Language Workshop* (Kaufmann, San Mateo, CA), pp. 285–290.
41. Nakadai, Y. & Sugamura, N. (1990) in *Proceedings of the International Conference on Spoken Language Processing*, pp. 1141–1144.
42. Powell, G. A., *et al.* (1987) *Proc. IEEE ICASSP-87*, pp. 173–176.
43. Nagabuchi, H. (1988) *Trans. IEICE Japan* (Institute of Electronics, Information and Communication Engineers, Tokyo), No. 5, pp. 1100–1106.
44. Roe, D. B. (1987) *Proc. IEEE ICASSP-87*, pp. 1139–1142.
45. Lippmann, R., *et al.* (1987) *Proc. IEEE ICASSP-87*, pp. 705–708.
46. Miki, S., *et al.* (1990) in *Technical Report SP90-19 of the Acoustical Society of Japan*.
47. Nakatsu, R. (1989) in *Proceedings of Speech Tech '89* (Media Dimensions, New York), pp. 4–7.
48. Pleasant, B. (1989) in *Proceedings of Speech Tech '89* (Media Dimensions, New York), pp. 2–3.