5

R&D Trends in Speech Recognition / Synthesis and Natural Language Processing
— Challenges toward the Establishment of User-Friendly Human Interfaces —

MASAO WATARI
Information and Communications Research Unit

5.1 Introduction

Speech recognition and synthesis, and natural language processing have long been research subjects, as they are input/output technologies that allow people to operate devices in a natural manner. On early computers, the human interface method was text command input by the user and text message output by the computer. What followed was the graphical user interface that enabled icon-oriented displays and selection using a mouse pointer. More recently, advances in computer graphics as well as in image, audio and other multimedia processing have led to the creation of more diverse interfaces. In addition, in an effort to improve their ease of use, studies on screen designs and various kinds of input devices are ongoing.

Despite such activities, human interfaces furnished on information appliances still require certain levels of skill for users, not reaching a level of natural human communications. Ideal forms of human interfaces may be interacting with an information system through voice or natural language (sentences used in our daily communication) in some cases to pick up information from a foreign language with the aid of a machine translation system.

Interfaces using speech and natural language have long been studied and have evolved to the point where they have found application in limited areas. Voice input for word processors (dictation), and translation systems for roughly reading foreign-language information on the Web are already available. However, the current state leaves much room for improvement, because a recognition rate of daily conversation remains low and translation quality for complex sentences is insufficient.

Meanwhile, broadening users of the Internet, various kinds of information devices such as personal computers, mobile phones and personal digital assistances have diffused, increasing the demand for an interface that allows "anyone," inclusive of not only sophisticated users but also novices and the elderly, to use such devices "any time" "with ease." With this in view, the Council for Science and Technology Policy of the Cabinet Office is aiming at developing "human interface technology to give machines advanced communication skills to understand and interact with humans" in ten years as exploratory researches that will lead to next-generation breakthroughs.

This paper first describes the history and the current status of speech and natural language research, and discusses differences of approaches adopted by Japan and the U.S. to promote research projects. Then, it intends to suggest challenges to be addressed for facilitating research on next-generation human interfaces.

5.2 The development and the current state of human interface technologies

5.2.1 Speech recognition

(1) History of development

The history of speech recognition study, through which researchers have been trying to create a system capable of recognizing words spoken by humans, started back in 1952, when Davis and others at the Bell Laboratories made attempts toward recognition of spoken numerals by using the zero-crossing rate*1. Subsequently, in 1959, the research for the "phonetic typewriter," a device that can recognize monosyllables, was conducted at Kyoto University. A breakthrough that led to commercialization came in the 1970s, when the DP matching method*2, in which the variation in utterance duration is normalized through dynamic programming, was proposed concurrently in Japan and Russia, followed by another proposal by Japan concerning a two-level DP-matching algorithm in order to recognize continuous digit words. By commercialization of this technique, a minicomputer-based continuous-word recognizer was introduced onto the market in 1978 to help those operators who have to input data with their hands busy for sorting packages.

During the 1970s a statistical method, the Hidden Markov Model (HMM) was studied in the U.S. It became a standard technique for spoken word recognition in the 1980s. From the late 1980s through the early 1990s, the Defense Advanced Research Projects Agency (DARPA) conducted dictation projects. In this project, the n-Gram method, a technique in which the statistical probability among n words was proposed. It brought the realization of large vocabulary continuous speech recognition. On the basis of this achievement and the improved performance of PCs, dictation software for large vocabulary continuous speech recognition was first marketed in 1997 in the U.S. At the same time, the Japanese large vocabulary continuous speech recognition software appeared in the Japanese market.

In the first half of the 1990s, through DARPA projects focused on spoken dialogue question and answering (Q&A) systems, investigation of dialogue processing technology to handle Q&A started. The outcome was commercialized in 1998, as the automatic spoken dialogue processing system for telephone-based reservation/inquiry services (call center services).

On the other hand, Japanese researchers directed their commercialization efforts to voice recognition technology applicable to car navigation systems. To allow drivers to input their destinations and commands using their voice, an algorithm that can reduce processing loads without degradation in recognition capability was devised, and first built into products in 1995.

(2) The current state and challenges

The latest speech recognition technologies have reached the point where they are able to transcribe speech almost correctly as long as the speaker vocalizes words clearly. Moreover, their application to car navigation systems demonstrates that these recognition technologies are robust to certain noises.

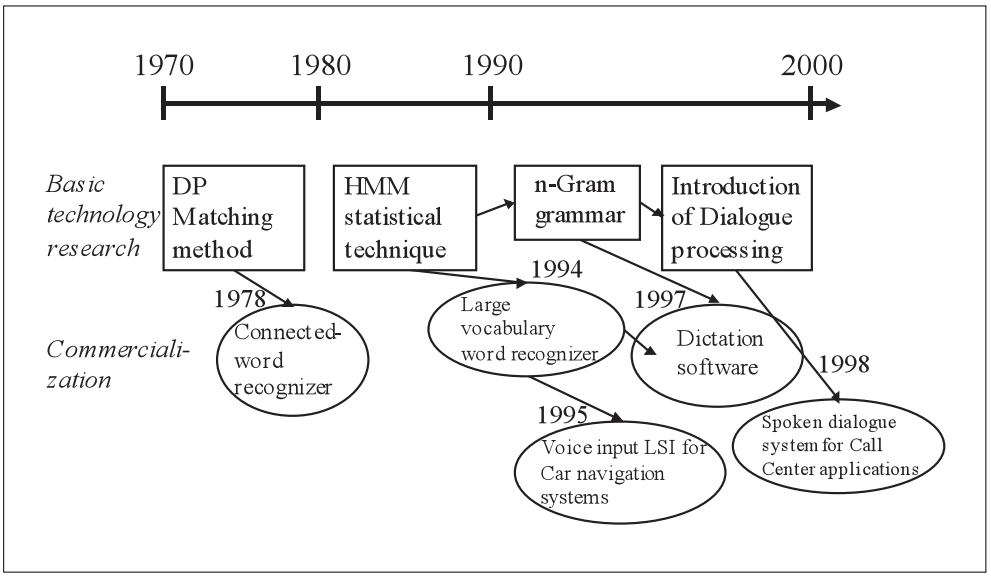
However, in areas such as hesitant "spontaneous speech" and "conversational speech" between friends, recognition performance remains poor. Also, the current technologies are not able to extract the speaker's intention or to make an adequate estimation of the speaker's situation.

5.2.2 Speech synthesis

(1) History of development

Research on speech synthesis has its roots in the 1950s as same as research on speech recognition, when researchers started seeking to output spoken messages from machines. K. Stevens at the Massachusetts Institute of Technology (MIT), G. Fant at the Royal Institute of Technology, Sweden (KTH), and others proposed the vocal-tract analog speech synthesizer — a device that can reproduce the acoustic characteristics of the vocal tract through an electric equivalent circuit. In the 1970s, NTT Electrical Communication Laboratories proposed speech synthesizer by using the linear prediction coding (LPC) to significantly reduce signal process calculations. In 1978, Texas Instruments (TI) succeeded a commercialization of a game device, Speak&Spell which can produce a certain number of spoken messages by using LPC method.

Figure 1: History of speech recognition research and commercialization



A technology to synthesize speech from arbitrary text was first developed by Klatt at MIT through the description of prosodic and phonological rules based on his knowledge and expertise. The achievement caught the attention of the Digital Equipment Corporation (DEC), and led to commercial introduction of a product named DECTalk in 1983. In the years that followed, as enhanced computer processing ability permitted acoustic waveforms to be edited and processed, waveform manipulation methods were pursued, resulting in improvements in clearness of synthesized speech.

During the 1990s, in order to produce natural, fluent speech synthesis having smooth concatenation of speech segments, considerable research energy was extended to make new model for prosodic and phonological rules derived from actual data. In the latter half of the 1990s, the HMM method, a basic technique for speech recognition, was used for segmentation of speech into phonemes so that the phoneme data for synthesis could be gotten automatically from real speech. As a result, a significant part of the preparation of data for synthesis process was automated, making it easier to create a speech synthesis system that can generate synthesized speech of a specific person's voice by collecting the person's basic voice data.

(2) The current state and challenges

Today's speech synthesis systems enabled by PC-based software can produce voices very close to humans' in terms of intonation and articulation, reaching a level where users are likely to have little odd impressions about them. The next challenge would be the synthesis of speech in a variety of styles such as recitation and conversation, and the addition of emotions.

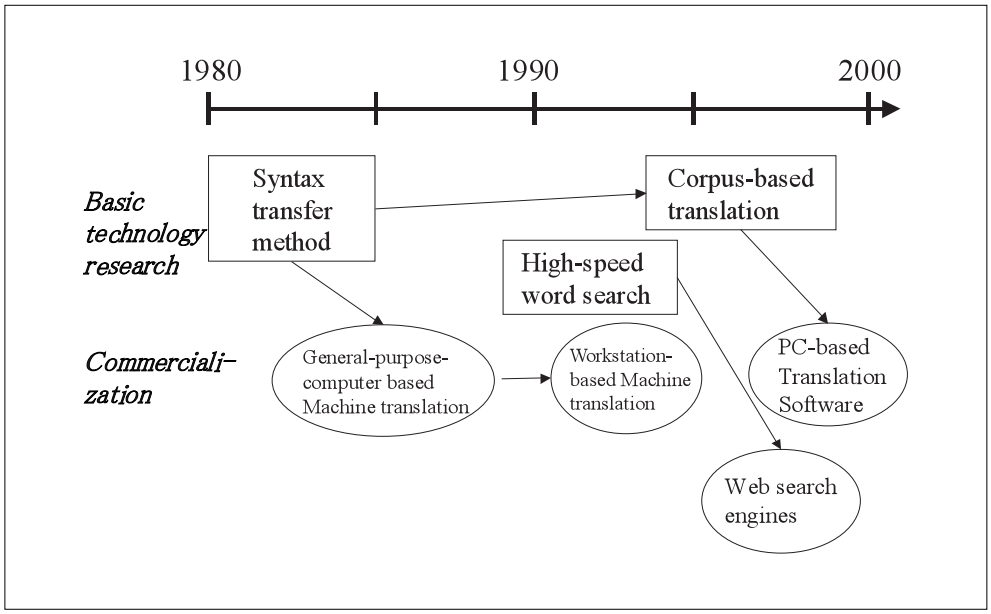
5.2.3 Natural language processing

(1) History of development

Natural language processing refers to a technology that allows computers to understand and generate natural language which people use for daily communication, and it plays the essential role in natural communication between humans and computers. The beginning of research on natural language processing goes back to the early attempts in the 1950s, which intended for translation by computers (machine translation). While U.S. researchers started with Russian-English translation, the Japanese counterparts began tackling their own subjects at Kyushu University and the Electrotechnical Laboratory.

In the U.S., the ALPAC report (written by the Automated Language Processing Advisory Committee (ALPAC) organized by the National Science Foundation) was released in 1966, which concluded that basic research was recommendable instead because machine translation was too complex to solve by using computing power and it's quality was too low to improve. Consequently, researchers shifted toward the area of basic linguistics, virtually bringing machine translation research in the country to a

Figure 2: History of natural language recognition research and commercialization



halt.

In the latter half of the 1970s, machine translation study drew strong attention in Europe and Canada, where needs were stronger. Through their research activities, the transfer method for translation of linguistically similar language pairs, which uses word-to-word correspondence between the two languages, was developed. The technology was commercialized, for example, to translate English weather forecasts into French in Canada in 1976.

In Japan, the syntax transfer method, in which the syntax of sentences was analyzed and target language sentences were reconstructed by using transferred syntax, was the mainstream because of the large linguistic distance between Japanese and English. During the first half of the 1980s, English-Japanese/Japanese-English translation systems for the abstracts of scientific papers were studied under a national project led by Kyoto University. In 1986, Japanese computer manufacturers introduced onto the market its first model of an English-Japanese/Japanese-English translation systems running on general-purpose computers. Whereas its translation quality was not enough for unedited use, the system served as a useful aid for translators to increase the efficiency.

To boost the capacity to generate grammars and lexicons, which had been limited in the conventional hand-made method, what advanced in the 1990s was the technology to construct them from enormous volumes of Japanese-English bilingual translated data. The new technology, called corpus-based translation*3, contributed to enhancing translation quality. With the advancement in computer performance, workstation-based translation software became commercially available in the early 1990s, followed by PC-based products in the late 1990s.

As an application of natural language processing, keyword search technology was developed in the 1990s to enable automatic document sorting and retrieval. This technology uses morphological analysis and word segmentation, which was established through machine translation studies. In addition, Web search engines, indispensable tools to search desired Web sites, depend on a combination of morphological analysis technology and high-speed word search technology by using parallel computing.

(2) The current state and challenges

According to a report by the Asia-Pacific Association for Machine Translation (AAMT), the performance delivered by current English-Japanese translation systems is high enough to help improve reading comprehension skills of people whose English ability is equal to or below TOEIC 700 (intermediate). For another leap in machine translation capability, semantic analysis among words and context analysis between sentences, which is an unsuccessful research area, should be encouraged.

With the growth of information available on the

Web, users are demanding better means to retrieve desired information among the vast and various resources. To this end, retrieval and summarization technologies that use something beyond keywords, such as meanings or concepts, should be sought after.

5.3 Challenges in the Promotion of next-generation human interface technology

As described in the previous chapters, both speech recognition and natural language processing are attempts to enable computers to perform some human capabilities. Long-lasting basic research has led to breakthroughs that served as the foundation for application and commercialization of these new technologies. However, their current levels still fall short of satisfaction of all users. This chapter discusses research direction, research environment, and research management issues to be addressed to develop next-generation human interfaces that functions more like humans.

5.3.1 Research challenges

The performance of speech recognition and natural language processing, which has made a significant leap forward through the use of statistical techniques in the 1990s, seems to face a big barrier recently. Current systems cannot provide phonemic recognition rates as humans have and cannot handle prosodic information. They also fail to provide semantics and contextual processing as well as to interpret the speaker's intentions and situations based on dialogues.

What is desirable is to create a new model to move beyond the status quo by actively assimilating knowledge of acoustics and linguistics into the conventional statistical approaches. Some pioneering efforts on the basic component technologies are found among the representative Japanese speech recognition research projects listed in Table 2, whose outcomes are awaited with anticipation.

In the long run, more attention should be paid to the aspect that recognition and language comprehension are deeply related to the cognitive and learning functions performed by the human brain. While digital information processing technologies have made remarkable advances in the 20th century, cognitive and learning mechanisms remain to be elucidated. Researchers will realize the need to look deeper understanding of the human brain mechanism, and this issue will stand as a major challenge in the 21st century.

5.3.2 Construction of databases for common use

To build the foundation of speech recognition and natural language research, data collection is crucial. However, the task requires considerable staff hours, it is difficult for a single research institute to make various kinds of speech databases. More specifically, in addition to the collection of huge volumes of data, manually checking analysis results which are attached for each data component is such enormous work that it cannot be handled by one laboratory alone.

A public-aided membership consortium that aims to construct linguistic databases to be shared among members was established in the U.S. in 1992 as the Linguistic Data Consortium (LDC), and in Europe in 1995 as the European Language Resource Association (ELRA), and both are continuing their activities to date. Many experts are working for these organizations to collect, maintain and distribute the databases. Their databases are offered inexpensively for academic research purposes, while their commercial use is offered at higher prices.

Japan has also seen its own attempts to create common databases, none of which have lasted long or led to sustainable activities thus far. Databases constructed through a project most likely become unavailable as soon as the project ends, due to a lack of financial support to maintain and upgrade the databases afterwards.

In 1999, following the U.S. and European precedents, the Language Resource Consortium (GSK; acronym for the Japanese phrase Gengo Shigen Kyoyukiko), an organization to collect, maintain and expand databases for common use, was formed in Japan, even though it has yet to start any substantial activity due to the lack of funds. Raising funds for building the foundation of research is not easy, as the foundation itself would not yield any immediate practical outcome.

A common database not only serves as a research foundation, but also allows fair performance evaluation of the systems by using the same database. From this perspective, construction of common databases is hoped for as a means to facilitate fair competition among research institutions.

5.3.3 Difference between Japanese and U.S. research projects

The gap between Japan and the U.S. is said to be widening in terms of the speed of commercialization from R&D results. This can be seen in some cases of the commercialization of speech recognition technologies.

In the U.S, DARPA played an important role in R&D of speech recognition. As shown in Table 1, the organization concretized projects goal with clear application concepts. In these projects, DARPA funded two or more research institutions to achieve the same goal, so that they studied to reach the goal in competition with others. In more detail, a database for evaluations was created at the point of project start and shared among participants. The institutions involved, most of them universities, contended for higher performance through making prototype systems with integration of many component technologies to demonstrate possible applications. After the project, the resulting intellectual property rights (IPRs) were transferred to the respective institutions, which were then allowed to transfer the technologies to private enterprises such as venture firms, solely through their own decisions when necessary. An example can be seen in the DARPA-funded research project on spoken dialog technology for air travel information, conducted from 1990 through 1994. In 1995, researchers at the institutions in charge — MIT, Carnegie Mellon University (CMU), and SRI International— established two venture firms in Boston and Silicon Valley. They built phone-based automatic reservation/inquiry systems by using speech recognition technology, which was utilized to support 24-hour call center services with minimal staff. Today, in the U.S., voice automation systems for call centers are expanding their scope of application beyond flight reservations, to a variety of reservation/information services. The majority of the market is controlled by the two venture companies, which have now grown to considerable scale.

In Japan, the major national projects on speech recognition as shown in Table 2 are in progress. Except for the "Spoken Language Translation Research Project" by ATR and the recent project for "The Realization of Advanced Spoken Language Information Processing from Prosodic Features," all of them aim to improve the recognition performance for either read text or spoken dialogue, as in the case with U.S. DARPA projects. While the "Research on Human-Machine Dialogue System through Spoken Language" project members created prototypes for evaluation, the other projects have carried out only component technologies so that each university can focus on its part. This method contributed to the development of constituent technologies, but did not clearly indicate the overall performance level of each resulting system due to the absence of an evaluation scheme for the integrated outcome. On the other hand, industry has been making their

Table 1: U.S. DARPA projects and commercialization of their results

Period	Project	Project objective	Result in commercialization
Latter half of 1980s First half of 1990s	Dictation projects Resource Management Task (small-scale dictation) (1987-1990) Dictation of WSJ newspaper (1991-1996)	Dictation of reading aloud newspaper	Release of the PC-based "Dictation" software (from IBM, etc.)
Latter half of 1990s	Q&A dialogue projects Air travel information system: ATIS (1990-1994) Web information retrieval: TIDES (1998-2000) Communicator (1999-2003)	Flight reservation and weather forecast inquiry in spoken dialogue.	Spoken dialogue systems for call centers (by two U.S. venture firms)
Early 2000s	Talking-agent project: Human Centered System (2002-2007)	Construction of a talking digital secretary agent.	

Table 2: Major voice recognition projects in Japan

Term	Project Name	Description	Funding
FY1986-1989	Advanced Man-Machine Interface through Spoken Language	Research on speech recognition of recitation.	Grant-in-Aid for Scientific Research on Priority Areas by MEXT
FY1993-1995	Research on Understanding and Generating Dialogue by Integrated Processing of Speech, Language and Concept	Research on dialog comprehension through integrated processing of speech and language.	Grant-in-Aid for Scientific Research on Priority Areas by MEXT
FY1993-1999	Spoken Language Translation Research Project	Automatic translation of conversation for hotel reservations between English and Japanese speakers.	ATR (former ATR Interpreting Telecommunication Research Laboratories)
FY1996-2000	Research on Man-Machine Dialogue System Through Spoken Language	Creation of a spoken dialogue system to retrieve academic literature from the Internet.	Research for the Future Program by the Japan Society for the Promotion of Science
FY1997-1999	IPA Japanese Dictation free software project	Development of continuous speech recognition software for Japanese.	Project of the Information-technology Promotion Agency (IPA)
FY1999-2003	Spontaneous Speech: Corpus and Processing Technology	Establishment of the technology for spontaneous speech recognition, understanding and summarization	Science and Technology Agency Priority Program, Organized Research Combination System
FY1999-2003	Integrated Understanding of Multidimensional Acoustic Signals	Research on sound spatiality, speech analysis/synthesis, speech recognition, spoken dialogue, and recognition of sound information.	COE formation program by MEXT
FY2000-2002	Development of the Anthropomorphic Dialogue Agent	Development of the basic software for the technology to generate human-like dialogues by machines.	Project of the Information-technology Promotion Agency (IPA)
FY2000-2003	The Realization of Advanced Spoken Language Information Processing from Prosodic Features	Integration of prosodic studies, ranging from basics to application.	Grant-in-Aid for Scientific Research on Priority Areas (B) by MEXT

own effort to push ahead with application and commercialization on the basis of academic findings. Under the situation where industry and academia are not eager to collaborate with each other, technology transfers from a university to a venture firm, as seen in the U.S., have not occurred in Japan. In other words, current and past national projects of Japan place emphasis on the advance of basic research rather than the promotion of business-academia collaboration.

In the panel discussion on "Present State and Future of Large-Scale Projects on Speech and Language"[1] conducted by the Spoken Language Processing SIG of IPSJ, cited were problems associated with national projects, such as the difficulty to take risky and adventuresome approaches for cerate totally new innovation and absence of a support to maintain the outcomes after the project ends. This suggests what should be coped in the future Japanese project funding program, namely, fostering "a culture where serious competitiveness without fear of failure" is valued in place of the current "success-oriented culture," and utilizing the results of a project to the next project.

Another point indicated by the university professors interviewed was that, in the information research field, research staff including postdoctoral fellows, working at universities are too scarce to deal with large-scale systems and thus they tend to choose more specific component technologies as their research subjects.

In the U.S., academic organizations always make effort to get funds from external sources, as their basic funds are limited. Since a national project can provide them with the critical amount of research grants, competition to acquire them is so fierce that even a basic research project has a distinct objective. In the case of DARPA-funded speech recognition projects, fair competition is said to have accelerated research activities. A swift

and smooth technology transfer from research to commercial is also contributing to the country's overall strength in technology development.

5.4 Conclusion

Amid the spread of information devices, the next-generation human interface technology is expected to solve the digital divide and to improve machines' friendliness to users even including beginners. In an age when every kind of information is available through networks accessible via mobile devices just around the corner, ubiquitous access to information is becoming a reality. The ultimate goal should be to enable people to easily obtain desired information through natural communication with information devices.

Revitalization of basic research is one of desired way in order to attack many obstacles on the path to such goal. Now that improvement in performance has hit the ceiling in some respect, a new attempt to break through is needed.

With the growing global trend in basic researches depend more on universities' effort than business enterprises' effort, an expectation for universities basic research outcomes is increasing. To allow them to carry out research projects on the basis of risky but innovative ideas, the culture should be changed where fair competition and evaluation are ensured, and those failed are more tolerated and encouraged to make another endeavor.

Furthermore, with an eye to focusing energy on prompt commercialization of created breakthrough technologies, measures should be taken to foster activities less mature than the U.S., such as industry-academia collaboration and technology transfer.

Glossary

*1 Zero-crossing rate
This is the number of zero crossings of the waveform within a certain frame period, and indicates an approximate amount of energy of the waveform. This method was able to get characteristics of speech sound with relative ease on hardware.

*2 DP matching method
The duration of each phoneme in speech varies in each utterance. The method uses dynamic programming to normalize the variation and find the most likely pattern.

*3 Corpus-based translation
The use of corpus here refers to a large-scale database consisting of example sentences and their translations. The corpus-based machine translation uses grammars extracted from the corpus and a part of example translations in the corpus. The resulting translation quality is higher than that of conventional machine translation based on empirical methods.

References

[1] Fujisaki, "Panel: Present State and Future of Large Scale Projects on Speech and Language," Special Interest Group on Spoken Language Processing, Information Processing Society of Japan, 29-40 (1999)

(Original Japanese version: published in March 2002)