論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Speech Recognition -Past, Present, and Future -
著者(和文)	古井 貞熙
Authors(English)	Sadaoki Furui
出典(和文)	, Vol. 7, No. 2, pp. 13-18
Citation(English)	NTT Review, Vol. 7, No. 2, pp. 13-18
発行日 / Pub. date	1995, 3

Speech Recognition — Past, Present, and Future —

Sadaoki Furui

After a brief review of its history, this paper describes the state of the art and discusses the future prospects of speech and speaker recognition technology. The recent incorporation of several new techniques, such as hidden Markov models (HMMs), delta-cepstral parameters, and stochastic language modeling, has greatly improved the accuracy of speaker-independent, large-vocabulary, continuous speech recognition. Speaker recognition accuracy has also been improved by using HMMs, likelihood normalization, and the text-prompted method. However, further research based on the inter-disciplinary aspects of speech is still needed before systems can approach the capability of human beings.

Introduction

Speech recognition is a knowledge-based process for automatically extracting various information from speech uttered by human beings. Speech recognition can be classified into the process of recognizing the linguistic meaning or content of speech (speech recognition in the narrow sense) and that of recognizing who is speaking (speaker recognition). The research into automatic speech recognition systems has been conducted by many researchers in many countries for more than 40 years since the 1950's. Almost a quarter of a century has passed since I entered this field.

During the last 25 years, many robots and computers have appeared in science fiction books and movies that can recognize speech, such as the computer HAL in Stanley Kubrick's famous movie "2001—A Space Odyssey" and the robot C3PO in George Lucas's "Star Wars". These have excited many people. The desire to make intelligent machines that can recognize spoken language and understand its meaning has been a dream of many researchers. However, despite the significant technological progress that has been achieved during these 25–40 years, the goal of understanding speech on arbitrary topics uttered by arbitrary speakers in all kinds of environments is still far away.

Since speech cannot be directly seen by eyes, it is difficult to intuitively understand the difficulty of speech recognition. However, the difficulty of recognizing speech is comparable to that of reading Chinese characters handwritten in cursive (running) style and deciding who wrote them. In cursive style, each character varies slightly depending on the neighboring characters. In speech, this phenomenon is called coarticulation. With written characters, the variation among different writers and due to additive noise, such as spots on the paper, is very large. There is similar variation in conversational speech; everyday conversational speech includes botched utterances and repetition more frequently than writing. Robustness against these problems is crucial in order to build speech recognition systems that are useful in real fields.

The number of researchers in speech recognition is much higher than that in speaker

recognition. Although there are some differences between the difficulties of these two tasks, they have many techniques and problems in common. This paper begins with a brief history, then describes the state of the art of both speech and speaker recognition technology, and finally discusses future prospects.

1 History of Speech Recognition Research

1.1 Speech Recognition in the Narrow Sense

The first paper on a speech recognition system was written by Davis et al. of AT&T Bell Labs in the US and published in 1952. It presented their system "Audrey"[1]. This recognized single digit utterances spoken in isolation by a single speaker. Later in the 1950's, studies for recognizing isolated syllables and phonemes embedded in isolated words were conducted at RCA Lab in the US and at University College in the UK, respectively. An important aspect of the latter research was that it used stochastic information concerning possible phoneme strings in English. This work pioneered stochastic language modeling, which is one of the major techniques used nowadays. The first trial on speaker-independent speech recognition was the vowel recognizer built at MIT Lincoln Lab in the US in 1959.

In the 1960's, several research laboratories in Japan, such as Radio Research Lab (currently Communications Research Lab), Kyoto Univ., and NTT Labs, started speech recognition trials, and they achieved good results in digit recognition and phoneme recognition including vowels.

One of the most important achievements between the late 60's and the early 70's was the linear predictive coding (LPC) technique and the dynamic time warping (DTW) technique using a dynamic programming (DP) procedure[2]. LPC was first proposed by Itakura and Saito of NTT Labs as an efficient speech analysis technique based on a speech production model. DTW is a very efficient method for calculating the similarity between two time sequences and it can cope with temporal expansion and contraction of speech. This technique was proposed independently at almost the same time at NEC and in the Soviet Union (currently Ukraine), and its basic idea is still widely used in current speech recognition technology. Itakura achieved very high recognition performance by combining these two methods, LPC and DTW, while he was at AT&T Bell Labs as a visiting researcher in the early 1970's. This work triggered a resurgence of speech recognition research at Bell Labs, where it had been suspended for several years.

Important milestones in the 1970's include algorithms for connected word recognition, including the two-stage DP method proposed at NEC, continuous speech recognition algorithms investigated at Carnegie-Mellon Univ. (CMU) and IBM in the US, speaker-independent recognition techniques proposed at Bell Labs, and recognition algorithms using phonemes and syllable units proposed at NTT Labs.

The 1980's can be characterized by a shift from template (reference pattern)-matching-based methods to statistical-modeling-based methods, such as those using hidden Markov models (HMM). Although HMM-based methods were already well known at a few laboratories including IBM, the methods only started to be used at research laboratories worldwide in the middle of the 1980's, when actual methods and theories were widely published. In language modeling, stochastic methods started to be used in place of the production rule-based methods which had been used before, and a method based on word concatenation probabilities was intensively investigated at IBM.

The 1980's were also the period when a project sponsored by Defence Advanced Research Projects Agency (DARPA, currently ARPA) for recognizing continuous speech with a 1,000-word vocabulary started to play an important role in the US as a driving force for speech recognition research. Under this project, many important research achievements were obtained at SRI, MIT, CMU, BBN, MIT Lincoln Labs, etc., being accelerated by the intense competition. They include the construction of large-scale spoken language databases and techniques for using detailed

phoneme models and stochastic language models based on these databases.

In Japan, a speech recognition research group was established in ATR in western Japan (Kansai) under the leadership of NTT researchers on secondment there. They started their research with the ambitious target of automatic translation telephony. One typical advanced research activity in Japan in the 1980's was the search for speaker adaptation techniques conducted at ATR, Shinshu Univ., NTT Labs, etc.

1.2 Speaker Recognition

Among the various applications of speaker recognition technology, the first to be tried was forensics. In 1660, speech was used as the key to detect a criminal for the first time in the case concerning the death of Charles I of England. However, at that time, the decision was made subjectively based on human hearing. The first scientific approach to voice individuality was used in the kidnapping of Lindbergh's son in the 1930's. In the 1940's, the sound spectrograph was invented by Potter at Bell Labs, and this made it possible to automatically draw voice prints (visible speech). In the 1960's, Kersta at Bell Labs reported the possibility of speaker recognition using voice prints, and they were first used in court in the US.

The first paper on automatic speaker recognition was written by Pruzansky of Bell Labs in the 1960's. From then until the beginning of the 1970's, speaker recognition research was started at several laboratories, including IBM, TI, and NTT, and both text-independent and text-dependent methods were investigated[2]. NTT Labs focused on the superiority of cepstral parameters from the beginning of their research and took the lead in using these parameters.

I tried to use polynomial expansion coefficients of time sequences of cepstral coefficients as dynamic features in combination with cepstral coefficients, and obtained high recognition accuracies at the end of the 1970's, while I was visiting Bell Labs. The method of combining cepstral coefficients and their polynomial expansion coefficients (they are now called delta- and delta-delta-cepstra) was first applied to speech recognition (in the narrow sense) at NTT Labs in the 1980's. Because of its excellent features, this method is now widely used in almost all speech recognition systems around the world.

In the 1980's, speaker recognition research also started to use stochastic approaches such as HMM, and a large improvement in recognition accuracy was achieved. One of the important achievements in 1980's was the proposal of the likelihood normalization method by ITT in the US. In this method, the variation of likelihood values due to inter-session and text-dependent variabilities of feature parameters is normalized based on the idea of likelihood ratio, in order to set a stable threshold for making a decision in speaker verification.

2 Current Speech and Speaker Recognition Techniques

Speech recognition techniques used commercially these days are classified into isolated word recognition and key word spotting from continuous speech. The ANSER (Automatic Answer Network System for Electrical Request) system developed at NTT Labs more than 10 years ago uses the former technique, and the automated operator service recently developed at Bell Labs uses the latter technique. A wide variety of research and development is now being actively performed at many laboratories around the world, ranging from fundamental research for recognizing large-vocabulary continuous speech uttered by unlimited speakers to the development of commercial systems for specific applications.

The structure of a typical large-vocabulary continuous speech recognition system currently under investigation is shown in Fig. 1. In this system, a speech wave is first converted into digital form in the signal processing part, and then converted into a time series of feature parameters, such as cepstra and delta-cepstra, in the feature extraction part. Although various methods for extracting and using prosodic features, that is, time functions of voice pitch (height) and amplitude, have been investigated, no satisfactory method has yet been

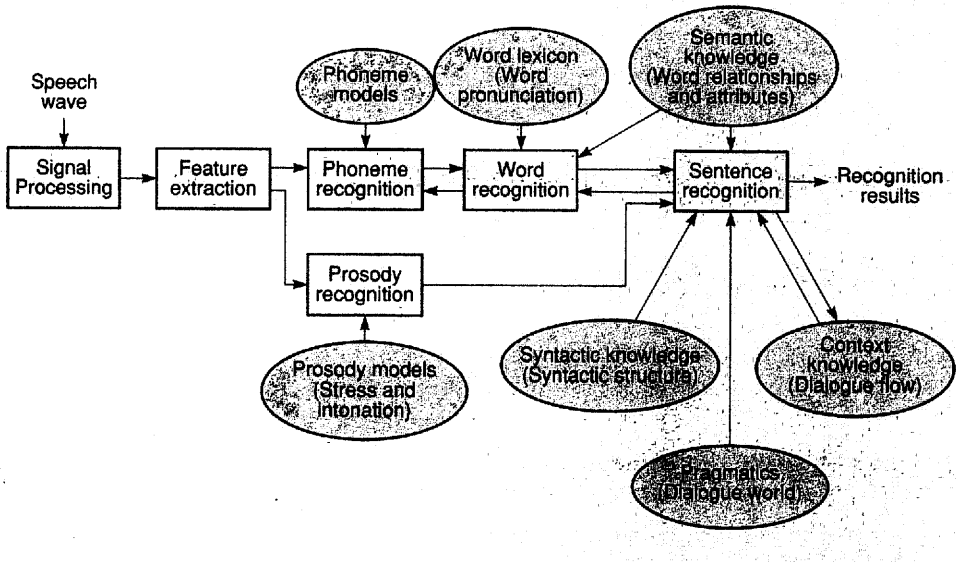

Figure 1. Typical Structure of an Up-to-Date Speech Recognition System

invented.

In typical up-to-date continuous speech recognition systems, the system predicts a sentence (hypothesis) that is likely to be spoken by the user, based on the current topic, the meaning of words, and language grammar, and represents the sentence as a sequence of words. This sequence is then converted into a sequence of phoneme models which were created beforehand in a training stage. Each phoneme model is typically represented by an HMM. The likelihood (probability) of producing the time series of feature parameters from the sequence of the phoneme models is calculated, and combined with the linguistic likelihood (appropriateness) of the hypothesized sentence to calculate the overall likelihood (probability) that the sentence was uttered by the speaker. The (overall) likelihood is calculated for other sentence hypotheses, and the sentence with the highest likelihood score is chosen as the recognition result.

Thus, in most of the current advanced systems, the recognition process is performed top-down, that is, driven by linguistic knowledge. Based on this principle, a very large vocabulary continuous speech recognition system with an 80,000-word vocabulary was built as a key element of a multi-modal dialogue system for telephone directory assistance at NTT Labs. It is based on the HMM-LR algorithm using HMMs as phoneme models and a generalized LR parser as a language model. To cope with the problem of background noise, an HMM composition technique was investigated and incorporated into the system. The results of experiments give the sentence understanding rate as 58% and the task completion rate as 99%[4]. (See "Speech Recognition Technology" in this Special Feature)

The only speaker recognition techniques that have been put to practical use are text-dependent ones. However, since these limit the applications, text-independent and text-prompted methods are now under investigation for various purposes. The text-prompted method was recently proposed at NTT Labs to cope with the problem that conventional systems are easily defeated by a recorded voice[5]. In this method, a new text is prompted by the system every time the system is used. The system accepts the input utterance only when it determines that the registered speaker uttered the prompted sentence. That is, the system not only recognizes speakers, but also rejects utterances whose text differs from the prompted text, even if it is uttered by a registered speaker.

Table 1. Prospects for the Development of Speech Recognition Systems

Year	1990	1995	2000	2000+
	Isolated/connected words	Continuous speech	Continuous speech	Continuous speech
Recognition Capability	Whole word models, word spotting, finite-state grammars, Constrained tasks	Sub-word recognition elements, stochastic language models	Sub-word recognition elements, language models representative of natural language, task-specific semantics	Spontaneous speech grammar, syntax, semantics; adaptation; learning
Vocabulary Size	10 – 30	100 – 1,000	5,000 – 20,000	Unrestricted
Processor Requirements	2 – 4 DSPs (25 Mips/Chip)	4 – 10 DSPs (50 Mips/Chip)	5 – 50 DSPs (200 Mips/Chip)	20 – 60 DSPs (1,000 Mips/Chip)
Applications	Voice dialing, credit-card entry; catalog ordering, inventory inquiry, transaction inquiry	Transaction processing, robot control, resource management	Dictation machines, computer-based sacretarial assistants, database access	Spontaneous speech interaction, translation telephony

Because the vocabulary is unlimited, prospective impostors cannot know in advance the sentence they will be prompted to say. A recorded and played-back voice can thus be correctly rejected. (See "Speaker Recognition Technology" in this Special Feature)

3 Future Prospects of Speech and Speaker Recognition Techniques

Table 1 shows the prospects of speech recognition systems at the level of practical use. This table was originally made by Rabiner and Juang[1], and has been modified by me[3]. Major applications of speech recognition include database (information) retrieval, guidance and transactions, automated reservations, automated ordering, dictation, voice-activated word processors, electronic secretaries, robots, automated translation telephony, and aids for the handicapped.

Speaker recognition techniques are expected to be widely used in the future as methods of verifying the claimed identity in telephone banking and shopping services, information retrieval services, remote access to computers, credit-card calls, etc.

The capability of current speech and speaker recognition techniques is very limited compared with that of human beings. In order for this technology to become widely used, it is essential for it to progress step by step, and for useful applications that match the current level (capability and limitation) of the technology to be created. The most important issues include how to create language models (rules) for spoken language, and how to establish methods that are robust against voice variation due to individuality, the physical and psychological condition of the speaker, telephone sets, network characteristics, additive background noise, and so on.

For these purposes, it is necessary to promote sure and steady research and development by grasping the essence of speech phenomena, instead of developing methods by simply looking at them superficially. Speech recognition technology is related to many scientific and engineering fields as shown in Fig. 2, and has an inter-disciplinary nature. It can also be said that speech research exists at the boundary between natural science and engineering. Knowledge and technology from a wide range of areas will be necessary to develop speech recognition technology. Even though nobody can become an expert in all these areas, it is essential for speech researchers to thoroughly understand the fundamentals of these related areas.

For example, when several phonemes or syllables are continuously spoken, as in the case of usual sentence speech, the tongue, jaw, lips, etc. move asynchronously in parallel, and yet with coupled relationships. Current speech analysis techniques, however, represent speech as a simple time series of spectra. It will become necessary to analyze speech by decomposing it into several hidden factors based on speech production mechanisms. It will also be necessary to clarify the process by which human beings understand spoken lan-

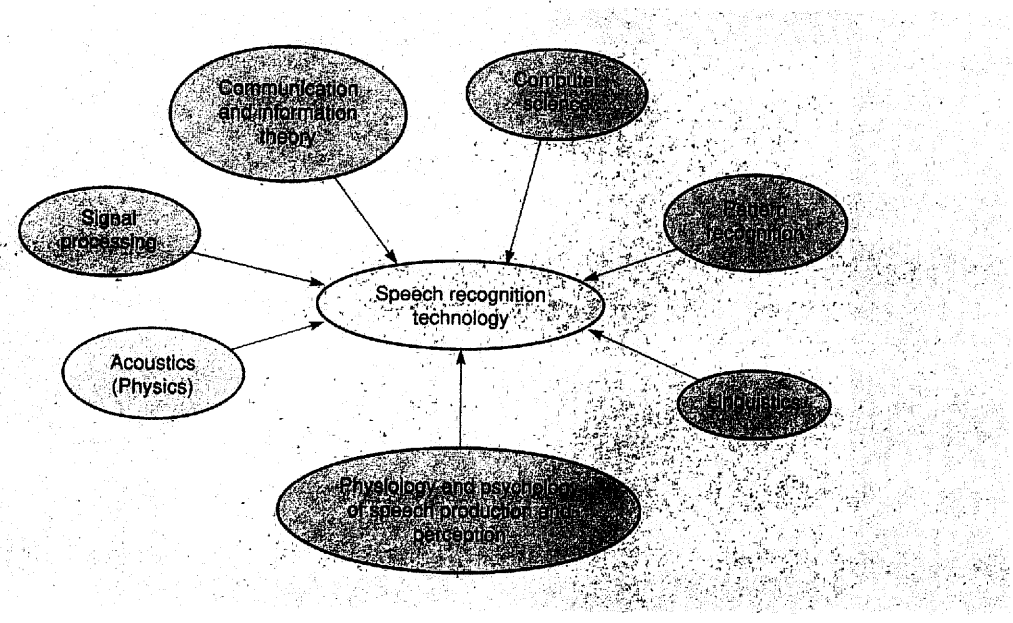

Figure 2. Science and Engineering Related to Speech Recognition

guage, in order to obtain hints for constructing language models for spoken language, which is very different from written language.

Conclusion

Speech recognition technology has made steady progress in its 40-year history, and has succeeded in creating several substantial applications. However, the current level is still far below the goal of its researchers, which is to equal or even exceed human capabilities. It is necessary to spend more effort on a wide range of speech fundamentals and technology to make real progress.

This Special Feature consists of four papers introducing recent advances and future prospects of speech and speaker recognition technology. Following this introduction, the second paper is an invited paper written by Dr. Chin-Hui Lee and Dr. Lawrence R. Rabiner of AT&T Bell Labs, who have been researching speech recognition for many years and have created various new ideas. Their paper focuses on future directions in speech recognition. The other two papers present state-of-the-art techniques for speech and speaker recognition, including acoustic modeling and linguistic modeling. I hope this special issue will provide useful information to a wide range of readers, and catch the interest of many people in speech and speaker recognition technology.

References

(1) L. R. Rabiner and B. H. Juang: "Fundamentals of Speech Recognition," Prentice-Hall, New Jersey, 1993.
(2) S. Furui: "Digital Speech Processing, Synthesis, and Recognition," Marcel Dekker, New York, 1989.
(3) S. Furui: "Toward the Ultimate Synthesis/Recognition System" *Voice Communication between Humans and Machines*, ed by D. B. Roe & J. G. Wilpon, National Academy Press, Washington D. C., pp.450-466, 1994.
(4) Y. Minami, K. Shikano, O. Yoshioka, S. Takahashi, T. Yamada and S. Furui: "A Large-Vocabulary Continuous Speech Recognition Algorithm and Its Application to a Multi-Modal Telephone Directory Assistance System," ARPA Workshop on Human Language Technology, Princeton, pp.362-367, 1994.
(5) T. Matsui and S. Furui: "Concatenated phoneme models for text-variable speaker recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Minneapolis, II-391-394, 1993.

The Author

Sadaoki Furui
Research Fellow and the Director, Furui Research Laboratory, Human Interface Laboratories, NTT.