Finding Practical Application for Speech Recognition

Realizing Conversations as Smooth as Those Between Native Language Speakers

Currently, in line with the advance of computer technologies and the accumulation of a large volume of speech data, research into practical applications of speech recognition is accelerating. On the other hand, while full-scale practical use has begun and expectations have been rising, some issues have been identified. Nobutaka Ono, an associate professor who studies signal processing of sound at NII (National Institute of Informatics), recently spoke with Prof. Tatsuya Kawahara, a Kyoto University professor and an expert in speech recognition research, and Chiori Hori, Director, Spoken Language Communication Laboratory, Universal Communication Research Institute, National Institute of Information and Communications Technology (hereinafter, NICT) on the history of speech recognition technology and initiatives to find practical application, as well as the current issues.

Cloud Computing Leads to the Advance of Speech Recognition Technology

Ono First, I would like to know about the history of speech recognition technology.

Kawahara Research into speech recognition technology began more than 50 years ago. At that time, some overseas research institutes including Bell Labs were studying speech recognition technology, while pioneering research activities were also being performed in Japan. In 1962, Kyoto University developed a "phonetic typewriter." This machine recognized monosyllables including "A (ah), O (oh), and I (ee)." Later, a technology developed around 1990 serves as the basis for speech recognition technology today. The technology is based on a feature that indicates the spectral envelope* and a state transition model of statistical distribution (HMM, or Hidden Markov Model). More than 20 years have passed since then, but the basic framework of speech recognition is nearly unchanged.

*Spectral envelope: A smooth spectrum pattern that is the most important of speech feature.

Ono Would you explain the actual mechanism of speech recognition?

Kawahara The main elements that are necessary for speech recognition are an acoustic model and a word dictionary/language model (Fig. 1). The acoustic model stores patterns of the frequency of each phoneme in Japanese and the language model stores typical sequence of words in Japanese.

Hori As you see, speech recognition combines multiple technologies to extract speech as text data. For example, when a sequence of sounds is replaced with words, a sufficient estimation cannot be made with a word dictionary alone. Consequently, candidate words are picked according to the sequence of sounds and, based on the language model, the probability of the word sequence is also considered to select the closest stochastic candidate.

Ono And in recent years there have been some technological breakthroughs.

Kawahara That's right. Among them are the sophistication of the statistical model, including the acoustic model, larger-scale training data, and remarkable improvements in the processing capacity of computers. Following the downsizing and performance enhancements of computers, the performance of mobile terminals including smartphones has also been enhanced. On the other hand, with faster networks, a cloud-server system has been achieved. Using these huge servers and data, speech recognition is currently being conducted not by a terminal, but by a background server. High-precision processing that was not possible previously is now being achieved.

Hori "Big data" has become one of the keywords. To achieve speech recognition that can be utilized in the real world, a database that stores a large vocabulary is essential. Currently, enormous amounts of text and speech data exist on the web. Utilizing this massive new store of information, research institutes and other organizations have been researching and developing technologies for subtitling, indexing for search, and translating videos and audio data from all over the world.

Interviewer
Nobutaka Ono
Associate Professor, Principles of Informatics Research Division, National Institute of Informatics
Associate Professor, Department of Informatics, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies

Advance of Application in a Variety of Situations

Ono With the continuous advances in speech recognition technologies, how well are they being applied to practical systems?

Kawahara Major applications can be divided in two, i.e., "speech interface," in which a machine is instructed to perform some task by speaking to it, and "speech content," in which a natural conversation between humans is automatically transcribed or subtitled. The former application has been in practical application for about 10 years in dictation software for computers including voice typing/voice-input word processors and voice command systems for car navigation systems. It is also used in voice access to information including reservations/inquiries by phone or mobile phone.

In recent years, in line with the improved performance of speech recognition by cloud servers and the improved performance and widespread use of smartphones, needs for voice input for mobile terminals have been growing (Fig. 2). In the meantime, the latter application, "speech content" is effectively used in subtitling on television or transcription of parliamentary meetings.

Fig. 1: Mechanism of speech recognition. W represents a sequence of words, P a sequence of phonemes, and X acoustic feature.

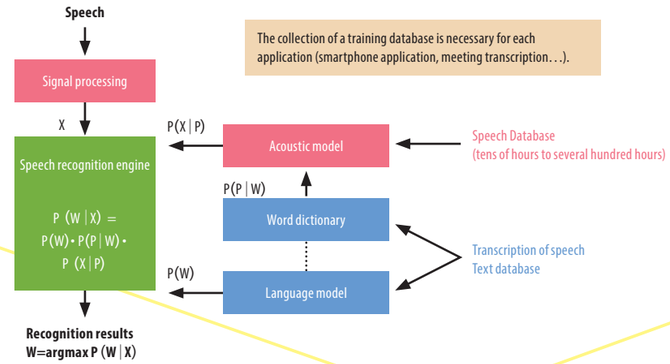
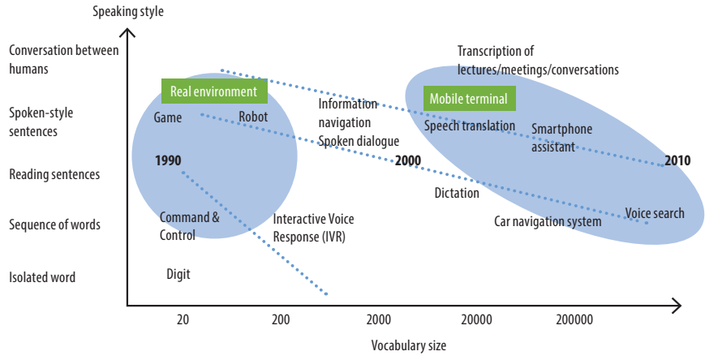


Fig. 2: Example of an application which utilizes speech recognition technology



The Spoken Language Communication Laboratory at NICT, where I belong, focuses on two areas of research. First, we are studying speech interfaces for simple communication between humans, and between humans and machines. Second, we conduct research on subtitling audio data on the web and automatic indexing technology for search. For speech interfaces, we have been developing a speech translation system that uses speech translation technology to achieve communication between speakers of different languages. We have also been working on a system that uses speech recognition and synthesis technologies to support communication between people with hearing difficulties and people who have full hearing. As for our subtitling research, we have been developing technologies such as recognition and translation of multilingual broadcast news audio into Japanese, automatic indexing, which enables you to search speech based on search words, and the extraction of non-speech acoustic events.

Since it is not an easy task to conduct NICT's research on multilingual speech processing solely in Japan, we aim to achieve a global network of speech translation research and to realize multilingual communications by using speech recognition through U-STAR (Universal Speech Translation Advanced Research), a consortium of 23 countries and 28 research institutes worldwide.

Tatsuya Kawahara
Professor, Graduate School of Informatics/Academic Center for Computing and Media Studies, Kyoto University

Challenges in Recognizing Spoken Language

Ono What are the challenges for further practical application?

Kawahara Most current speech recognition systems assume that users speak previously prepared content as a simple sentence politely and clearly. In this case, the recognition accuracy has reached 90%. However, the situation is different in conversations between humans. The accuracy of speech recognition becomes satisfactory in public speaking such as lectures and parliamentary meetings, especially when the speech is recorded in a studio or using a headset microphone, but it still is difficult to improve accuracy for conversations in a noisy environment, including a home or urban area, or for everyday conversation where speech is diversified.

Hori The challenge that we face is to recognize ambiguous speech from conversations as done between people that share the same native language; where one would speak while thinking, and also being able to recognize those that are not clearly pronounced.

Kawahara Unlike language processing done by humans, current speech recognition technology does not actually comprehend the meaning of the words. As stated earlier, collecting data for model construction is the key in speech recognition. However, a huge variety of data exists in conversations and the solution is not merely as simple as accumulating a large volume of this data. The immediate challenge is to realize a technology that can accurately handle diversified data.

Hori On the other hand, it is also important to create a technology that can enhance precision with limited resources, enabling us to achieve the same performance with very limited training data as that achieved with a large volume of data. While until recent years we had no choice but to only deal with superficial areas due to the limitations in the performance of computers, nowadays, large-scale computers can undertake vast calculations, and it is also essential to incorporate mass information to train a more high-precision model.

Kawahara To enhance speech recognition to the level of conversations in native speakers, one or two further breakthroughs are necessary. For this advance, we need to continue to pursue a statistical learning theory.

Hori To improve speech recognition, it must be used and therefore it is essential that we continue to apply and permeate the technology into our daily lives. Furthermore, by expanding speech recognition technology worldwide, we shall collect feedback from users, even identify new issues from them, and we shall be ready to resolve the issues when they are revealed. For this purpose, cooperation among each of the relevant organizations is essential to further foster and develop the technology.

(Interview/Report by Hideki Ito)

Chiori Hori
Director, Spoken Language Communication Laboratory, Universal Communication Research Institute
National Institute of Information and Communications Technology (NICT)

Research by Prof. Tatsuya Kawahara

Speech Recognition Technology that Plays an Active Role in Making Transcripts in Parliament

In Parliament, meeting records were created by stenographers. Yet, with the abolition of training for new stenographers, a system that employs the speech recognition technology of Prof. Kawahara and others was introduced in the House of Representatives in 2011. The system recognizes every speech recorded through the speaker's microphone in all plenary sessions and committee meetings to create a draft for the meeting record (transcript). This is the world's first system that directly recognizes meeting speech in national Parliament.

Prof. Kawahara explains the mechanism of the system as follows:

"We first constructed a database (corpus) that consisted of the speech of meetings in the House of Representatives and a faithful transcript (actual utterance). We then analyzed the difference with the sentences in the meeting records to create a statistical model. As a result, we discovered that approximately 13% of the words were different, mainly in the elimination of redundant words such as 'Eh ('Well' in English)' or 'Desune (an end-of-sentence expression).' Based on this statistical model, we have constructed a language model that predicts the actual utterance from a large volume of texts in meeting records consisting of approximately 200 million words during the past 10 years or so."

In addition, applying this language model to actual speech, we constructed an acoustic model from approximately 500 hours of recordings of meetings. These models are trained and updated in a semi-automatic manner. With a future general election or cabinet reshuffle, the models will reflect the change in the set of speakers and continuously improve their performance.

Prior to the full deployment, the performance of the system was evaluated in 2010. The accuracy of speech recognition was 89% in terms of character correctness against meeting records. The speech recognition result is corrected and edited by a stenographer using a special editor. The usefulness of this system that creates draft transcripts was verified and full-scale operation of the system began.

"The average character correct rate in 118 meetings in 2011 was 89.8%. The rate did not fall below 85% in almost any of the meetings. As far as plenary sessions are concerned, the rate is nearly 95%. However, we are not satisfied. By improving the performance a level higher, we think that we can port the system into other applications" (Prof. Kawahara)

Meeting transcription system for House of Representatives

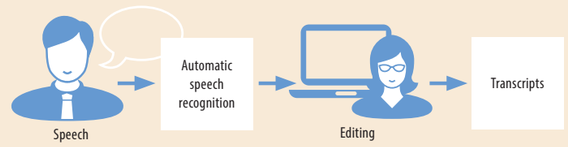


Image of the transcription system that utilizes the speech recognition technology. Starting in FY2011, this system processes meetings in all plenary sessions and committee meetings in the House of Representatives. This is the world's first speech recognition system deployed and operated that directly recognizes meeting speech in national Parliament.

Research by Dr. Chiori Hori

Development of Speech Translation Application that Overcomes Language Barriers

Ono Currently, in U-STAR (Universal Speech Translation Advanced Research), a consortium of 23 countries and 28 research institutes worldwide, NICT is conducting research and development of an automatic speech translation system through international collaboration. In 2012, U-STAR developed a multilingual speech translation system that covers approximately 95% of the world's population and its official languages. Dr. Hori takes the leading role in its development. Would you explain a little about it?

Hori Through collaboration with research institutes in each country, U-STAR is currently developing a multilingual speech translation system that is available for 17 speech-input languages, 27 text-input languages, and 14 speech-output languages. This system implements a technology that NICT standardized internationally (compliant with ITU-T Recommendations F.745 and H.625) in 2010. A control server that NICT operates and servers for speech recognition, machine translation, and speech synthesis that each member research institute operates are mutually connected over a network-based speech translation communication protocol. Multilingual speech translation service is then provided to users through a client application.

Ono Is the system open for general users?

Hori Yes. The speech translation application; VoiceTra4U, has been publicly released for iOS devices as part of our field experiment. VoiceTra4U recognizes spoken speech and translates the contents of conversations with simple operations. This application covers more than 30 languages in Asia and Europe. In addition to the "Single Mode" where you can translate and display the results on a single device, the "Chat Mode" allows up to five iPhones to be connected to conduct a chat conversation on a real-time basis. The contents of the conversations are translated into each designated language of the speakers.

Ono This application is very useful. Would you explain your ideas on how the application advances from here?

Hori VoiceTra4U allows speech-based conversations with people with visual impairments or text-based conversations with people with hearing difficulties. We aim to overcome communication-modality barriers as well as language barriers.

Outline of network-based speech translation by U-STAR