ISSN 1883-1974 (Print)
ISSN 1884-0787 (Online)
National Institute of Informatics News

NII Today

National Institute of Informatics News

51
Oct. 2014

Feature

Synthesis and Recognition of Speech

Creating and Listening to Speech

A digital book version of "NII Today" is now available.
http://www.nii.ac.jp/about/publication/today/

This English language edition
NII Today corresponds to
No. 65 of the Japanese edition

NII Interview

A Combination of Speech Synthesis and Speech Recognition Creates an Affluent Society

More and more people have begun to use smartphones or tablet devices with voice control. In addition, more people are operating smartphones while they listen to and communicate with the voice the smartphones generate. The former is an example of speech recognition technology and the latter is an example of speech synthesis technology. The two technologies once seemed to be somewhat elusive. During the past few years, however, they have rapidly become more feasible. We recently spoke with Nobutaka Ono, an associate professor at NII (National Institute of Informatics) who works on speech separation useful for speech recognition, and Junichi Yamagishi, an associate professor at NII who studies speech synthesis, who are both on the front lines of speech synthesis and speech recognition.

Ohkawara Attention to speech recognition and speech synthesis has been growing quickly of late. What do you think is the reason for this?

Ono For one thing, we cannot overlook the emergence of the smartphone, which is the most suitable device for speech recognition. A smartphone is very effective as a device for accurate speech recognition since a microphone can be very close to the user's mouth while the user is speaking. This is far more advantageous than a car navigation system or electric appliances that recognize speech with a distant microphone. For another, a large volume of data can be collected through a variety of devices. I strongly feel that speech recognition technology has advanced very rapidly in the past few years. When I write an e-mail with a smartphone, I use speech input more often.

Yamagishi HMM (Hidden Markov Model) is the mainstream approach in current speech recognition technology. This method employs a statistical approach, accumulating a large volume of data to build a speech recognition model based on the data.

This advance in technology is supported by new trends including the use of big data and deep learning in recent years.

Ono Yet, to advance speech recognition technology, a large volume of data is necessary. Therefore, companies or organizations that have big data have become dominant, and so more data accrues to those companies and organizations. Since this cycle is repeated, it is becoming more difficult for other companies or organizations to enter this field.

Ohkawara Japan's speech recognition technology was more advanced than other countries' previously, but Japanese companies or research institutes are no longer in that position. Is it because they put emphasis on the volume of data?

Yamagishi We have to admit that. On the other hand, Japan's speech synthesis is advanced to the level of being something of a fine art. My current study uses a large database of speech to create an average voice based on a function of the frequency of phonemes in a verb or the frequency of the voice while reading and that of an angry voice, or that of something in between. Then, with the average voice, data that indicate the difference in the voice of each individual are combined to create a voice that sounds similar to the voice of a specific person in only about 10 minutes. The voice of people who have been left speechless due to amyotrophic lateral sclerosis (ALS) or as a result of surgery for cancer can be created based only on small volume of data of their voice.

Ohkawara In the meantime, these days, two technologies, speech recognition and speech synthesis, are strongly linked.

Yamagishi Speech synthesis and speech recognition were entirely different technologies. However, each technology has advanced and has adopted a statistical approach (Hidden Markov Model). This leads researchers to come and go between each community, which has started a chemical reaction. The development process of an average voice by speech synthesis that I am currently studying is based on a speech recognition technology.

Ono However, there is controversy as to the difference in technical elements required for speech synthesis and speech recognition.

Yamagishi Many of the technologies can be shared now, but when we focus on the finer details, different elements are required. Understanding meaning is sufficient for speech recognition and recognition of fine nuance is not necessary. Still, speech synthesis is required to re-create fine nuance. Even when the same statistical model is used, the method or granularity of learning is different.

Ohkawara What can be done when technologies for speech recognition and speech synthesis are combined?

Yamagishi One result is a speech translation system. This system recognizes speech and translates it using machine translation to synthesize speech, also automatically translating it into every language to speak. Moreover, the speech is created with a human voice. In second language learning, you can understand how you should pronounce it with your own voice. If this is further advanced, the system could have an actor in a movie speak in a different language with the actor's own voice.

Ohkawara Can the human ear and the human mouth be the target for speech recognition and speech synthesis, respectively?

Ono Speech recognition or speech synthesis does not aim to mimic the human ear or the human mouth. There are things humans can do and these technologies cannot do, but there are also things the technologies can do that are out of reach of humans. For example, in recognizing speech, humans can recognize speech even when there is considerable noise or a speaker is at a considerable distance, whereas in these situations a speech recognition system has a long way to go. On the other hand, I study the extraction of a specific sound by using multiple microphones as preprocessing to speech recognition. Humans cannot extract pure sound from mixed sound and listen to that sound.

Ohkawara I would like your comments on the future challenges.

Ono The challenge for speech recognition is how close it will come to humans in the distant speech case. If this study is advanced, it will be possible to summarize the contents of a meeting and to automatically take the minutes. If a robot understands the contents of conversations by multiple people in a natural way, a world in science fiction will be realized.

Yamagishi The challenge for speech synthesis is how to develop expressive power. Based on the current statistical approach, only average expressions are created. Therefore, this is suitable for narration, but is weak for "voice art" – for example, the part of a movie where an actor says a line. Unless this is solved, it will be difficult to exert expressive power for 30 minutes to one hour that does not bore listeners. On the other hand, preparing for the time when speech can freely be synthesized, we have to build a secure environment to prevent synthesized speech from being used for fraud. If we fail to do this, the technology will not be able to be developed into a technology that you can use freely when needed and for its intended purpose. I think that this is the challenge.

A Word from the Interviewer

Speech synthesis and speech recognition "seemed so close, but were so far away" several years age. However, the two technologies have come closer to sparking the kind of chemical reaction that will accelerate their advance. We cannot overlook the close links among important technological trends in the IT field including mobile, cloud, big data, social, and analytics that are behind this. I hope that the two technologies will make our life more affluent.

Katsuyuki Ohkawara
Journalist

Born in 1965. From Tokyo. Formerly editor-in-chief of an IT industry journal, he became a freelance journalist in 2001. Has written on wider areas centered on the IT industry for more than 25 years. Currently a freelance writer for business magazines, PC magazines, and web media.

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

Junichi Yamagishi
Associate Professor, Digital Content and Media Sciences Research Division, National Institute of Informatics
Associate Professor, Department of Informatics, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies

NII Special 1 Speech Synthesis

"Statistical Speech Synthesis" Technology with a Rapidly Growing Application Area

What is a Natural Speech Synthesis That Uses a Statistical Approach, HMM?

Speech synthesis technology today has advanced from the past unnatural "alien voice" to a high-quality voice that is hard to distinguish from the speech of ordinary humans. Behind this is progress in speech synthesis technology that uses a statistical approach. Due partly to a huge reduction in learning data and calculation data from past volumes, the application of speech synthesis has been rapidly expanding into areas including digital signage and robotics, support for the disabled, and mobile device navigation systems. We recently spoke with three of the world's leading researchers on the front lines of speech synthesis, including re-creation of the voices of people who have lost their power of speech and reading of translations with the voice of the original speaker.

Junichi Yamagishi

Associate Professor, Digital Content and Media Sciences Research Division, National Institute of Informatics
Associate Professor, Department of Informatics, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies

What Is a "Statistical Approach" That Has Dramatically Changed Speech Synthesis Technology?

Tokuda Researchers started working on speech synthesis in the 1950s, but speech synthesis leading to the current technology began to develop in the 1970s. During this period, research was made into what kinds of rules exist when a certain sound is followed by a next sound and how the rules should be used in speech synthesis.

Yamagishi Humans vibrate the vocal cords using air exhaled from the lung to produce a sound, and resonate the sound in the oral cavity or nasal cavity to enunciate. Humans adjust resonance frequency with the form of the tongue or the mouth and add a tone to vocalize. In the past, speech synthesis derived rules of how a sound from the vocal cords changed and rules for the change of resonance frequency, and modulated an original sound based on the rules depending on texts to synthesize a sound. The speech had a unique sound that sounded like "voice of a computer" or "voice of an alien" in an old science fiction movie.

Tokuda There was some research that led to very good results, but rule-making was influenced by the personality of the researcher. In addition, even if a rule for one person could be made, it took many years to formulate rules for people with diversified attributes, such as male, female, young, or old. The method lacked flexibility.

This situation was changed with computer technology that began to be developed rapidly at around that time.

A larger volume of data than ever before could be processed at high speeds. As a result, a large number of sounds are recorded to make a database called a corpus, where researchers can pick up a sound to produce natural sounds by cutting and pasting (concatenative speech synthesis). With this method, not only monosyllables, but continuous sounds of certain lengths, such as a word or a clause, can be used to synthesize a more natural sound (unit selection synthesis).

This corpus-based speech synthesis began to be applied to telephone automated response systems at call centers and the like, and to text-reading software for computers in the 1980s. In the 1990s, it was more widely used and there was a huge boom. The technology at that time is used widely today. The speech synthesis technology for "Vocaloid," currently popular on the Internet, is based on the speech synthesis technology in the 1980s.

Yamagishi However, this method requires recording of, for example, more than 10 to several hundred hours of speech data, and needs a large database. Professor Tokuda, Associate Professor Toda, and I worked together at the same research institute to record 100 hours of speech, which took as much as one year. Since the human voice varies widely depending on the person's condition, recording speech that can be used takes about ten times as long as the speech itself.

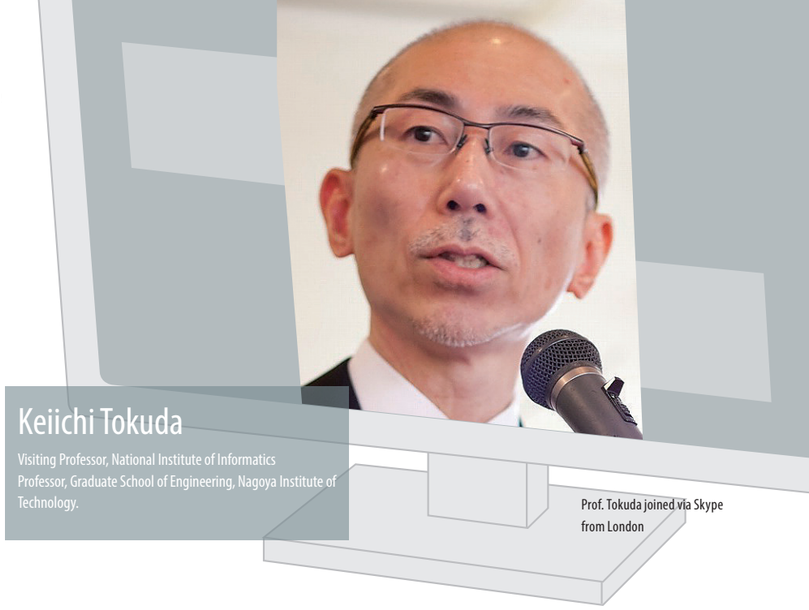
Tokuda If you want to produce a more natural speech, you need a huge amount of speech data and the post-processing including indexing is a massive job. When we assume a machine that can engage in natural conversation, for example, Doraemon or Her, a movie in which a person falls in love with artificial intelligence, an indefinite variety of speech synthesis is required and indefinite recording of data is necessary. In addition, as the size of the database becomes larger, it becomes more difficult to use in small devices, including mobile phones with limited storage capacity and low performance.

Then, as a result of research into a more flexible and efficient method for speech synthesis, we found a statistical stochastic model HMM (hidden Markov model).

Mechanism and Three Advantages of Speech Synthesis by HMM

Toda Speech synthesis that uses HMM begins to have a computer learn the correspondence between texts and speech waveforms accumulated in the corpus as a function. This is an approach that yielded results in speech recognition. This function is used to find a property that is hidden behind a speech waveform even when it has a fluctuation. For example, even the same expression "Ohayou (Good morning)" has a different waveform every time it is pronounced. Yet, a statistical method can calculate to identify a shared pattern and the correspondence to the text is always the same "Ohayou." Prof. Tokuda has developed the world's first technology for direct synthesis of speech from HMM.

Tokuda I have been working on HMM since 1995. To describe the mechanism simply, we derive the fundamental frequency (that corresponds to vocal intensity or intonation) of speech and spectrum *1 (equivalent to resonance in the vocal tract) from the training data and make a model (set a function) of the correspondence to the text by HMM. The results of analysis of the text read are checked against the database to generate a sound source by using a fundamental frequency parameter that is the closest to the answer statistically and in the same way, the spectrum parameter is generated to synthesize a tone (refer to Fig. 1). This method has the following three advantages.



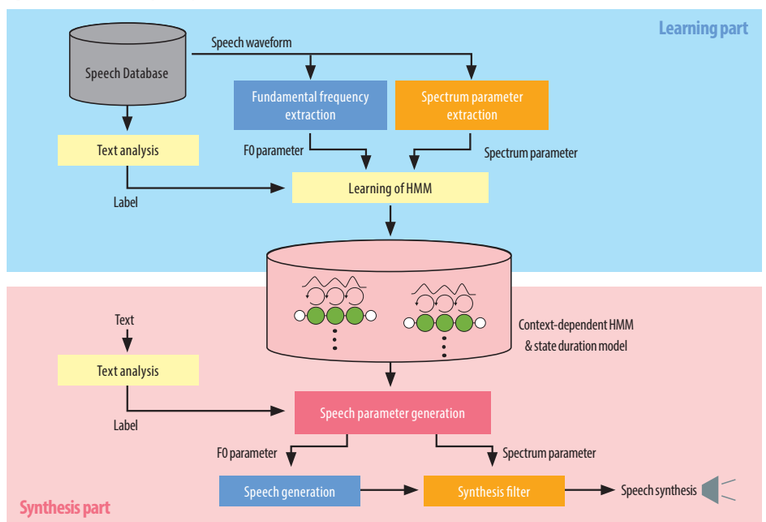
Keiichi Tokuda

Visiting Professor, National Institute of Informatics
Professor, Graduate School of Engineering, Nagoya Institute of Technology.

Prof. Tokuda joined via Skype from London

***1. Spectrum** Intensity distribution of frequency component contained in sound in the order of wavelength.

Fig 1: Framework of HMM speech synthesis



Tomoki Toda
Visiting Associate Professor, National Institute of Informatics
Associate Professor, Graduate School of Information Science, Nara Institute of Science and Technology

Speech Synthesis That Contributes to the Support of the Disabled

- **Memory Efficient Speech Synthesis System Loadable to Mobile Devices**

The volume of data needed for natural speech synthesis is reduced dramatically to only about 1 to 2 MB. Digital signage, mobile devices, or mascot robots can synthesize speech easily within their devices (Photo 1).

- **No language-dependency. Application to Multiple Languages Is Easy. Usable for Singing Voice Synthesis**

Since language-dependency is almost non-existent, speech synthesis software developed for a certain language can be used with little modification to other languages. Currently, this technology is applied to more than 40 languages.

With this flexible technology, when "Speech" and "Text" are replaced with "Singing voice" and "Music note with lyrics," respectively, the synthesis of a singing voice is made possible in the same mechanism. When Mr. A has a machine learn several songs that he has sung and enters the music note with the lyrics of another song into the machine, the machine sings the song with the voice of Mr. A. In addition, vocal assistance with one's own voice is possible. This technology is released to the general public as free/paid software as "CeVIO Creative Studio" (Photo 2).

- **Adjustment of Parameters Allows "Mimicking," "Mixing," and "Creating" Speech**

Simply adjusting parameters allows the creation of a variety of voices, which is also a characteristic of this technology. Users can add expressions of emotions, imitate the voice of another person, mix the voices of multiple individuals, or create speech that does not exist in the real world.

Toda High expectations are placed on technology that allows the creation of a variety of voices with only a change to parameters in the field of medical care and welfare in addition to entertainment. In the examples so far, a machine is assumed to read texts, but based on similar technology, for example, the speech of Mr. A is picked up with a microphone and is converted into the voice of Mr. B in real-time. When this technology is applied, a foreign language that a foreigner pronounces is converted to another Japanese speaker's voice as if it is uttered by the Japanese speaker. Also, in a similar way, speech that is difficult to listen to as it is can be converted to natural, clear speech.

For example, many of the people who underwent vocal cord cordectomy because of diseases including laryngeal cancer practice "esophageal speech" in which speech is uttered by vibrating the esophageal entrance well or voice production with an auxiliary called an electrolarynx, but the speech produced by these methods is very unnatural and hard to listen to compared to normal speech. In this situation, if we can obtain a sample of the voice of the person when he/she was healthy, a voice changer can convert the voice produced by these methods into a more natural voice, which is closer to the original voice of the person.

Yamagishi The point is that even if the volume of sample speech is small, the statistical speech synthesis technology can imitate the voice of the person very naturally. If recorded data of about 10 minutes is available, speech synthesis with the person's original voice is possible. In addition, it is easy to create an "average voice" from the voices of multiple individuals.

This is very helpful for supporting people with dysarthria. Patients with diseases including amyotrophic lateral sclerosis (ALS), in which dysarthria rapidly progresses, suffer from being less and less able to communicate with their own voice, and people around them are concerned about not being able to understand their speech. When these patients use a conversation support device that uses speech synthesizers built using recorded data of their voice before they became ill to output speech, their speech—which is becoming less and less easy to listen to—is corrected and output as clear speech. This required only several minutes of recorded data.

For a patient with ALS in Scotland, 20 people living in the vicinity cooperated in recording speech. The average parameter of their recording led to the successful creation of speech that was comfortably close to the pronunciation of the patient. Since the patient lived in an area with a thick accent, the general average voice was not satisfactory. Yet, with this technology, the patient's past way of speaking with an accent could be recreated and the patient was pleased to be able to recover their identity.

Everyone would like to speak with his/her own voice as much as possible. A conversation support device using speech synthesis technology will be able to be had at a low cost, but accumulating the speech data of many people as widely as possible is vital for each disabled person to put the device to good use. The "Voice Bank" project where speech data are collected worldwide and are made available is underway. In Japan, NII is currently promoting "Japanese Voice Bank Project"*2.

Clearer in Lamprophony Than Humans, the Future That Speech Synthesis Opens

Tokuda Associate Professor Toda is studying and developing support technology for people who can perform speech input, while Associate Professor Yamagishi is studying and developing a support technology for people who have trouble enunciating. These studies are expanding the area of application of speech synthesis more widely. The time will soon come when artificial intelligence makes a natural conversation with humans. I will accelerate technological development for a casual conversation with machines, not in an unpleasant voice, just as we enjoy with humans.

Yamagishi In a listening contest where researchers on speech synthesis worldwide evaluate the naturalness of synthesized speech, speech synthesis that uses HMM has been approved for the first time in the world as "having the same intelligibility as humans." It is also evaluated as "more intelligible caught in noisy condition than a human voice." In a sense, we have obtained a voice of higher quality than a human voice.

(Interview/Report by Masahiro Doi)

Photo 2 : "CeVIO Creative Studio" - "singing voice synthesis" software that utilizes speech synthesis technology that uses HMM. Character voice packages including "Satou Sasara" are available. With CeVIO Creative Studio, users can synthesize text speech and synthesized singing voice and also add feelings parameter including "cheer," "anger," or "sorrow," and also modify voice quality or vocal sound by changing timing, pitch, or volume. (Left: example of an editing screen of talk tracks, Right: example of an editing screen of song tracks)

Image supplier: CeVIO Project (Distributor) Frontier Works Inc. (Illustration) Masatsugu Saito http://cevio.jp/others/CCS/

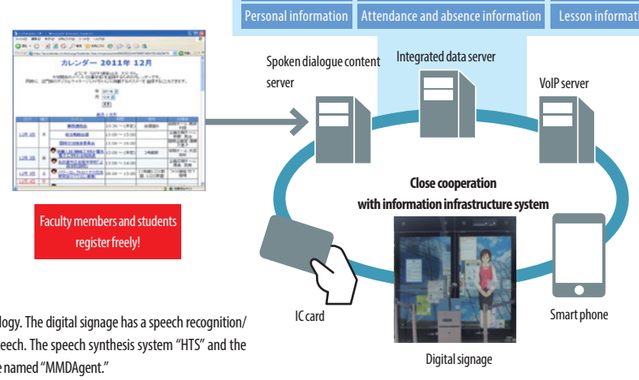


Photo 1: A digital signage constructed for navigation of the campus of Nagoya Institute of Technology. The digital signage has a speech recognition/speech synthesis/CG system. A CG character communicates interactively with visitors through speech. The speech synthesis system "HTS" and the speech recognition engine "Julius" that are used in this system are released as open-source software named "MMDAgent."

***2 Voice Bank Project** A project in which the voices of participants other than patients are collected to improve the quality of life for vocally impaired patients. The voice data are mixed so as to be used as a template, which allows the easy and swift construction of a speech synthesis system with the patients' own voice.
http://www.nii.ac.jp/research/voicebank/

Click below to access a movie/audio file.

Movie Speech Information Processing for Communication with Your Own Voice — Junichi Yamagishi Associate Professor
http://www.yourepeat.com/watch/?v=CSPP_z0GfzQ

Movie [MMDAgent] Created Software That Can Talk with Hatsune Miku — Keiichi Tokuda Professor
https://www.youtube.com/watch?v=hGiDMVakggE

Audio Augmented Speech Production Based on Statistical Voice Conversion — Tomoki Toda Associate Professor
http://isw3.naist.jp/~tomoki/NII/DemoVC_Toda@NAIST.pptx
※ PPT file is downloaded.

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.

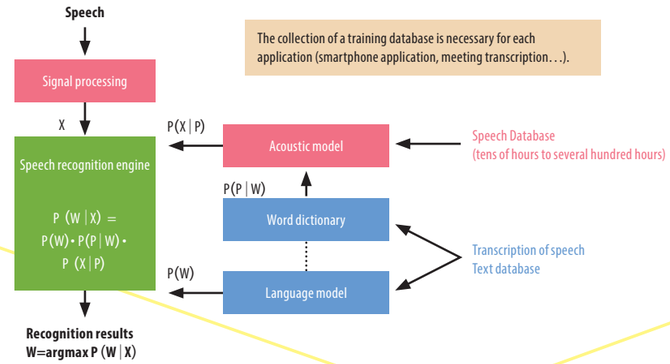
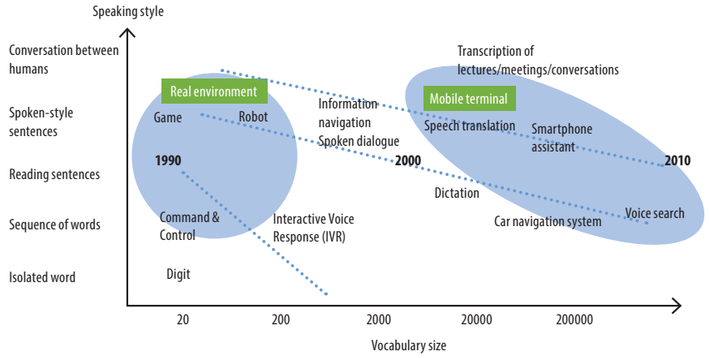


Fig. 2: Example of an application which utilizes speech recognition technology



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

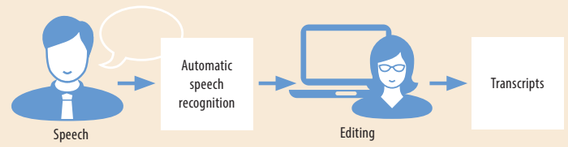


Image of the transcription system that utilizes the speech recognition technology. Starting in FY2011, this system processes meetings in all plenary sessions and committee meetings in the House of Representatives. This is the world's first speech recognition system deployed and operated that directly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

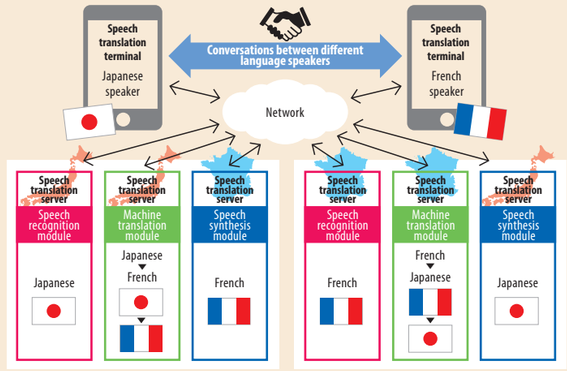
Ono Is the system open for general users?

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



NII essay

Voice is Conveyed or Voice Conveys?

Minoru Tsuzaki

Professor, Faculty of Music, Kyoto City University of Arts

Although the feature of this issue is "Speech Synthesis and Speech Recognition," I mainly study audition and I myself am always interested in the part that comes slightly before speech. Generally speaking, when it comes to speech, you first think of a human voice, which is differentiated from an animal call in that a human voice conveys words.

Words represent language content and it is indisputable that language content conveys meaning or will. Therefore, when a candidate in an election campaign says that "Please vote for me so that I might convey the voices of each one of you to parliament!" he does not promise to reproduce speech signals in parliament, but promises to represent the opinions of each voter and to act in parliament. The main objective of speech recognition and speech synthesis technologies is to make it possible for a machine to link linguistic information with speech signals. This technology must be convenient, so that it can be used effectively in every aspect.

Indeed, it has already begun to find application.

However, I think that there is one aspect of speech that we must never forget. It is an aspect that is more fundamental and since everybody enjoys it unconsciously, it tends to be hidden behind the aspect of linguistic information. That is personality. In a natural environment, a voice has an owner, and that is never an extra aspect of a voice.

There is a possible origin for person or "persona", i.e. per (through, by way of) and sona (sound). When you look up "persona" in an English-Japanese dictionary, it has a meaning of "Kamen (Mask in English)". This meaning originates from the aspect that in Greek times, an actor was identified only by his voice in a mask and that a skilled actor freely changed his voice depending on the mask that he wore. In short, the strategy of using a voice to identify a person has existed since ancient times. Animals began to produce a sound, thought to be originally a strategy for them to be able to identify individual animals in darkness.

In order for a machine that can recognize and synthesize speech to live together with humans in their living environment with high affinity, this aspect of a "persona" must be handled skillfully. If a machine that does not look like a human speaks as if it were a human, it may confuse users. Also, if the world's best speech synthesis machines have the same voice and speak on public information in various places, that would also be a problem. Even if a concierge robot properly understands what humans are saying, users might get angry if the robot's attitude is something like "By the way, who are you?"

Vocaloids, virtual characters, have won fans and I am among them. The key reasons why these female singers (?) never fall into the "uncanny valley" and are accepted are that the fans came to know the vocaloids first by their voices and a "persona" actually existed there.

Editor's postscript

The feature of this issue was speech, the area of specialty of the late Yoh'ichi Tohkura, the former deputy director of NII and a long-time chief editor of NII Today. The achievements of Prof. Tohkura are well-known. Among them, his efforts to establish Advanced Telecommunications Research Institute International, which has led cutting-edge research on speech, and on fostering young researchers, facilitated remarkable developments in speech research. Speech synthesis and recognition technologies have become important technologies in our lives. We take this opportunity to remember the great achievements of Prof. Tohkura.

Weaving Information into Knowledge
NII

National Institute of Informatics News [NII Today]
No. 51 Oct. 2014 [This English language edition NII Today corresponds to No. 65 of the Japanese edition.]
Published by National Institute of Informatics, Research Organization of Information and Systems
Address: National Center of Sciences 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430
Publisher: Masaru Kitsuregawa Editorial Supervisor: Ichiro Satoh
Cover illustration: Toshiya Shirotani Photography: Seiya Kawamoto, Yusuke Sato
Copy Editor: Madoka Tainaka Production: Nobudget Inc.
Contact: Publicity Team, Planning Division, General Affairs Department
TEL: +81-3-4212-2164 FAX: +81-3-4212-2150 E-mail: kouhou@nii.ac.jp

Bit (NII Character)

http://www.nii.ac.jp/about/publication/today/