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.

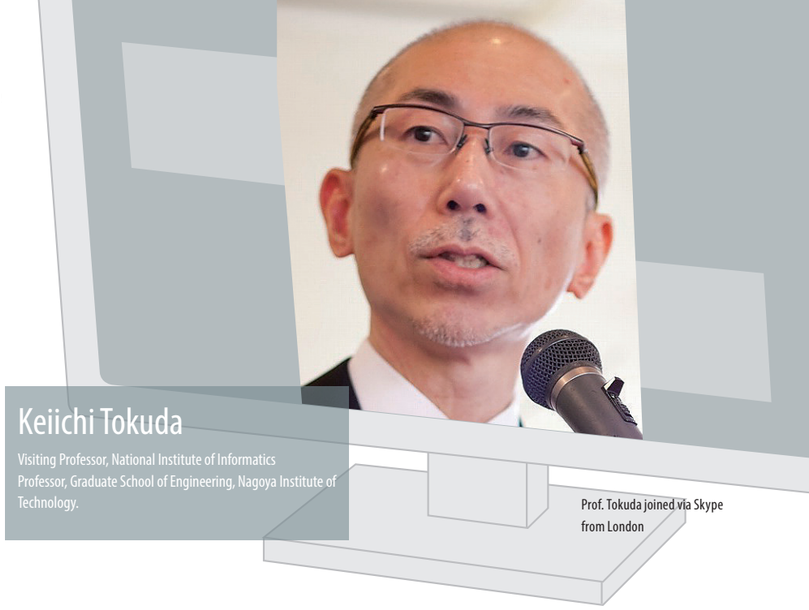
Tokuda If you want to produce a more natural speech, you need a huge amount of speech data and the post-processing including indexing is a massive job. When we assume a machine that can engage in natural conversation, for example, Doraemon or Her, a movie in which a person falls in love with artificial 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.



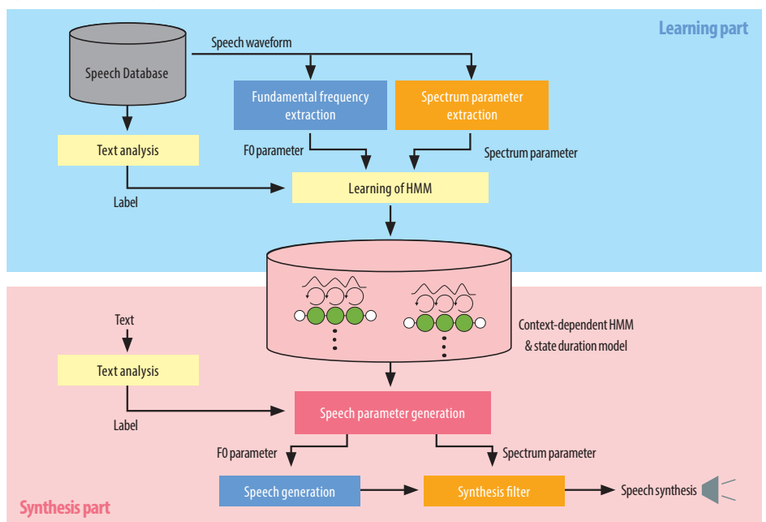
Keiichi Tokuda

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/

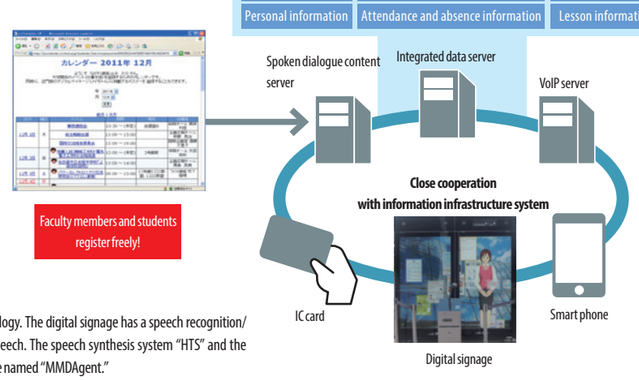


Photo 1: A digital signage constructed for navigation of the campus of Nagoya Institute of Technology. The digital signage has a speech recognition/speech synthesis/CG system. A CG character communicates interactively with visitors through speech. The speech synthesis system "HTS" and the speech recognition 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.