Pursuing Elegance and Creating a Research-goal Umbrella

Hirokazu Kameoka
Senior Distinguished Researcher, NTT Communication Science Laboratories

Overview

According to a survey on clumsiness while speaking of about 1800 undergraduate and graduate students of Japanese universities, approximately 30% answered that they have, or have to some extent, a problem with their pronunciation in everyday conversation, and those who were aware of the difficulty in pronunciation tended to feel that they are often asked to repeat what they said [1]. Hirokazu Kameoka, a senior distinguished researcher at NTT Communication Science Laboratories, aims to eliminate various obstacles in communication by analyzing, synthesizing, and converting voices and speaking styles. We asked him about his current research and his attitude as a researcher.

Keywords: decomposition of acoustic signals, acoustic scene analysis, speech-to-speech conversion

Development of technology for understanding a situation from sounds and converting voices according to that situation

—Please tell us about the research you are currently working on.

With the aim of creating a means by which people can communicate without inconvenience in a variety of situations, I am engaged in research on technology for decomposition of acoustic signals and acoustic scene analysis as well as speech-synthesis technology focused on high quality and naturalness. It is generally not easy to decompose a mixture into its constituents, e.g., extracting certain juices from a mixture of juices. In contrast, humans have the ability to understand acoustic scenes by distinguishing between sounds and by reading various nonlinguistic information contained in people's voices such as the tone and intention of the speaker. This ability plays an important role in people's social lives, especially in communication. For research on decomposition of acoustic signals and acoustic scene analysis, my aim is to build mathematical models and algorithms for computers to perform decomposition of acoustic signals and acoustic scene analysis.

Specifically, we, at NTT Communication Science Laboratories (CS Labs), have been working on several tasks: (i) sound-source separation, which separates and extracts multiple sounds contained in a mixed sound; (ii) sound-source identification, which identifies the nature of the target sound; (iii) voice activity detection, which estimates when the target sound is being emitted, (iv) sound-source localization, which estimates where the target sound is coming from; and (v) speech enhancement, which removes reverberation and noise to emphasize a specific voice. Conventionally, these tasks have been

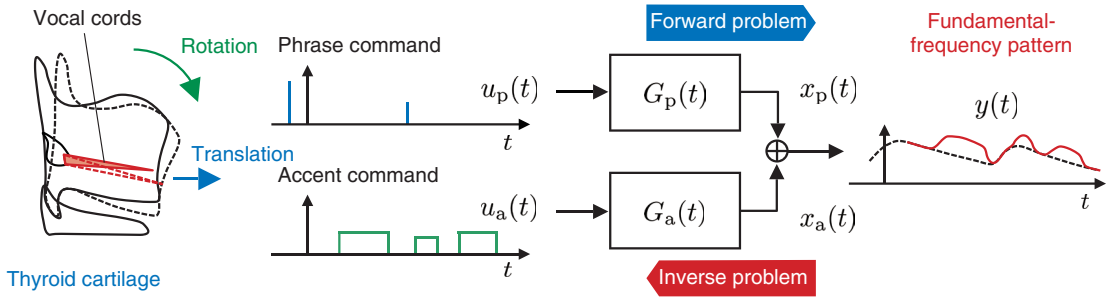

Fig. 1. Process of generating a fundamental-frequency pattern of speech and its inverse problem.

dealt with as individual research subjects. However, they are not actually independent but interdependent. For example, if sounds could be identified in advance, it would be relatively easy to separate them, and if the sounds could be separated in advance, it would be easy to localize them. In other words, the solution to one problem is a clue to solving another. From this viewpoint, instead of considering these problems as individual problems, I have approached them as joint optimization problems and devised a method of solving them collectively.

Just as multiple sounds are mixed into external sounds, various components are mixed into each voice. During conversation, we convey nonlinguistic information such as tone and intention to the other party by using the pitch of the voice together with the linguistic information corresponding to the words spoken. Speech contains elements such as phonemes related to linguistic information as well as phrase and accent components related to nonlinguistic information. The fundamental-frequency pattern, which represents the temporal change in pitch of the voice, is controlled by the thyroid cartilage that applies tension to the vocal cords, and phrase and accent components are considered associated with the translational and rotational movements of the thyroid cartilage. Correct estimation of the timing and intensity of these components can provide important physical quantities for quantifying nonlinguistic information; however, inverse estimation of these has long been a difficult problem (**Fig. 1**). In my research on decomposition and scene analysis concerning speech, I focused on the problem of decomposing the fundamental-frequency pattern into phrase and accent components and developed an efficient and accurate method for solving that problem based on a statistical signal-processing approach [2].

As an example application of this method, we implemented a demonstration system that converts standard-intonational Japanese speech into Kansai-accent intonation at events such as NTT CS Labs Open House and NTT R&D Forum. Since the content of the demonstration was familiar, it was very well received by many visitors and the press and widely covered in television programs, newspapers, and Internet articles.

This research lasted about 10 years, including the time I was in graduate school before joining the company. During that time, I was grateful to be acknowledged for many efforts. For example, I received the Signal Processing Society Young Author Best Paper Award from the Institute of Electrical and Electronics Engineers (IEEE) in 2009 and the Young Scientist's Prize of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2018. I heard that my IEEE award was the first such award to be received by a Japanese researcher. These awards improved my presence in the research field, so when I look back, I feel that they were significant events.

In addition to the research that I described above, we are working on speech synthesis with high quality and naturalness by using an approach based on deep learning. In particular, we are focusing on research on a speech-to-speech conversion technique that can flexibly convert various speech features (such as voice characteristic and rhythm) as well as fundamental-frequency patterns. There are many situations in which it is difficult to communicate smoothly. Some examples are conversations in an unfamiliar language, giving a presentation in a tense state of mind, and for people with impairment or deterioration of the vocal or hearing functions. Through research on this speech-to-speech conversion technique, we are aiming to remove such obstacles that can hinder smooth communication (**Fig. 2**).

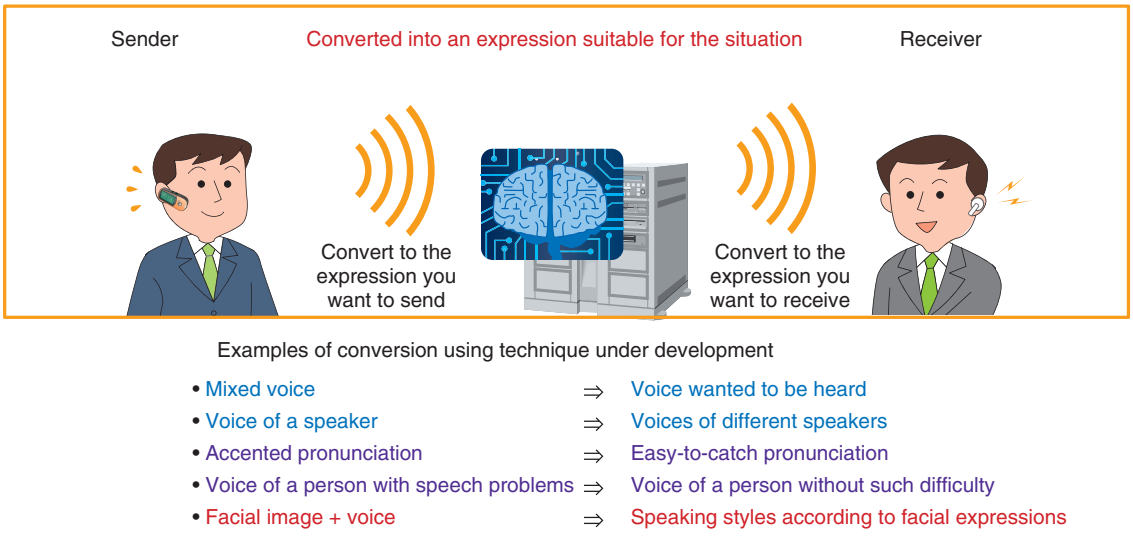

Fig. 2. Specific examples of communication-function-expansion technology.

A new perspective from the questions received at NTT CS Labs Open House

—You achieved great success. Please tell us a little more about your recent research on speech-to-speech conversion.

The fact that we decided to pursue deeper research into speech-to-speech conversion was actually triggered by an event at NTT CS Labs Open House that I mentioned earlier. Although the technique we introduced at that exhibition was used to convert an accent of the speech into another accent, one of the visitors asked the question, "Is it also possible to convert an accent of the speech uttered by a non-native speaker of English into an accent of a native speaker?" "Accent" is generally imagined as meaning a pronunciation accent of one's speech. Confusingly, in the field of speech research, it is also used as a term that means one component in the fundamental-frequency pattern, which is the meaning that we intended at that time. In fact, to convert the accent of a pronunciation in English, it is not enough to change the fundamental-frequency pattern; it is also necessary to change the *voice characteristic*, that is, the characteristic corresponding to the timbre of the voice. In other words, the technique that we introduced at that time could not handle conversion of the accent of an English pronunciation. That being said, that question got me a bit more interested in the problem of translating English-pronunciation accents. In the field of speech research, a technique called voice conversion, which converts an input voice into another person's voice, has been developed. In anticipation that this technique is probably the most relevant to the problem of translating English-pronunciation accents, I implemented and experimented with a current method of voice conversion.

However, it turned out that this method could change only the speaker identity; it could not change the speaking style such as the accent of a pronunciation. At this time, I felt that this problem was deeper and more interesting than I had imagined, so I decided to tackle the problem of voice conversion more seriously and focus on the broader problem of speech-to-speech conversion. Around that time, with the advent of deep learning (so-called artificial intelligence (AI)), I started working on speech synthesis and conversion using the deep-learning approach. In the four years that have passed since then, I have investigated various approaches in collaboration with colleagues and trainees, and we have created many high-quality and flexible speech-to-speech conversion techniques that can convert not only English-pronunciation accents but also a wider range of speech features.

—Why did you decide to work on acoustic signals and speech synthesis?

When I was a student, I played guitar in a band. If I had a favorite song, I often practiced it after listening

to the sound and wrote it down as a musical score, so-called "playing by ear." For my graduation thesis, since I wanted to work on a theme related to my interest, I decided to research an algorithm that automatically plays by ear. That decision gave me the chance to knock at the gate of a laboratory that studies sound. I knew nothing about that field of research, but I remember being excited when imagining how to use my knowledge of signal processing and statistics that I learned at university. At the time, I had a longing to sing better, but I didn't really feel confident because I didn't really like my singing voice. Therefore, besides automating playing by ear, I wondered if I could automatically convert my voice into a good singing voice. Many researchers who study speech are interested in music, and quite a few of them started their research for similar reasons. In my case too, such a simple motive has led to my current research. Looking back now, I feel that the dual motivations to automate playing by ear and improve my singing voice have something in common with my current research—which aims to support human hearing and vocalization functions.

What we can do musically is restricted by our musical skills, such as playing by ear, playing musical instruments, and singing. In the same way, in our daily communications, we face various restrictions due to physical and psychological conditions. My current interest is to remove such restrictions through the power of machine learning (AI) and signal processing and create an environment in which everyone can communicate comfortably and freely. To reach this goal, I think two technologies will be key. One is scene analysis, which captures the situation and environment in which the sender and receiver are located, and the other is media conversion, which converts the information that the sender wants to convey to the receiver into an expression suitable for the situation. I also want to explore the possibilities of new communication method that makes effective use of not only sounds but also multiple media, such as video images and text, and create the basic technology to implement them.

Expanding what you can do, what you want to do, and what is required while creating a research-goal umbrella

—Has anything changed since you became a senior distinguished researcher?

My research life has not changed much so far, but I want to focus on the following two objectives. The first is the objective that I have attempted to achieve as a researcher, that is, to expand the range of "what I can do," "what I want to do," and "what is required." The second is to create a research-goal *umbrella*. I learned the importance of having such an umbrella in the group to which I was assigned when I joined NTT. Having a clear research goal on which everyone can agree on its usefulness to society will allow researchers to focus on the difficult issues at hand and be confident in the direction they should take. Research activities are the work of steadily accumulating research results, of which each one is not necessarily a big success. Therefore, we may sometimes suddenly become anxious that what we are doing is nothing special. When I joind NTT, I was able to study with confidence under the umbrella that my senior colleagues created. This is exactly like being protected by an umbrella of a research goal. Therefore, as a senior distinguished researcher, I have become even more strongly motivated to make sure everything is crystal clear to junior researchers and that everyone involved can feel at ease and make progress without hesitation. I believe that to expand the range of "what I can do," "what I want to do," and "what is required," and create a research-goal umbrella will only be possible if I continue to improve and grow. With that in mind, I will continue to work hard and not accept things as they are.

On the contrary, I also try to keep in mind that one matter should remain unchanged. That is my research style, namely, *pursuing elegance*, which is also my policy as a researcher. This policy was formed during my university days when I was greatly influenced by the research style and thoughts of my supervisor who was a former NTT researcher. Elegance is hard to define, but it's an aesthetic sense for a skillful approach that sees the true nature of things. It may be similar to what we feel when we come across elegant problems, solutions, and proofs in mathematics. By pursuing elegance in everything we do, I think we should be able to sharpen our thoughts and climb to higher levels. In my life as a researcher, I have felt confident in my achievements, but I still feel they are not enough, so I want to continue pursuing elegance and continue writing as many papers as possible.

—Please say a few words to your junior researchers.

Although research is often difficult and emotionally demanding, I think it is most important to enjoy researching at NTT and your research. Research does

not progress when you are depressed or when your thoughts are blocked by delusions; conversely, when your spirits are high, research progresses.

Since the speed of your research will change in accordance with your feelings, it is important to keep your frame of mind stable and look forward. For example, we tend to think that we want to beat someone or show off to people around us. That thinking stems from being overly conscious of others and overwhelmed by negative emotions such as jealousy and impatience. Therefore, I think it's a good idea to focus on whether you are growing day by day without worrying about problems that are beyond your control. By looking forward to your own growth, even if you hit a brick wall, you will feel like trying harder to get over it.

You should also be careful when you get too absorbed in your research. At first glance, although it seemed that everything was going well at the time, when you look back later, you often find that your research actually had stalled. I'm prone to the same thinking, but, at such times, your perspective is usually narrowed. That is a state of pouring your heart and soul into your research while believing that what is objectively not so important is important. If you are not alert, you can fall into such a state anytime, so it is necessary to cultivate a sense of calmness that allows you to look at yourself objectively at all times as well as the ability to concentrate on what is in front of you. This means it is important to have two personalities—one that works diligently and one that is calm and objective—and consciously continue a dialogue between those two personalities.

I consider researchers to be providers of wisdom to change the world for the better. Of course, NTT researchers conduct research for NTT, but from a more macroscopic perspective, the common mission of all researchers is to make the world a better place. Although we need to be strong enough to compete with others, we must also have the integrity to be respectful of the research and contributions of others with the same mission.

—Please tell us about your future goals and prospects.

I will continue to pursue research focusing on how to facilitate human communication. Looking beyond that pursuit, we're aiming to achieve *cross-media conversion*. Audio, text, and video images are media with different characteristics. For example, audio is useful when you want to express a message quickly and convey it to another party, and text is useful when you want to quickly read the main points of the message. An advantage of video images is that they can express detailed information that cannot be expressed through audio or text alone. I believe that by taking advantage of the features of each of these media and allowing the sender and receiver to flexibly select the media to use according to the situation in which they are in, it will be possible to achieve efficient and smooth communication. To make that possibility a reality, we have to enable cross-media conversion, namely, transforming each media signal into a different media signal in a manner that retains the message or content. I think this task will be a very challenging and interesting theme.

As a researcher, I'd like to pursue beneficial research themes and elegant approaches. Each field has so-called *star* research themes. However, although that star theme is important, it might be an impregnable subject that many researchers have been tackling for years. Benchmarks, evaluation systems, and data sets have already been established for these research themes, making research and experimentation relatively easy. However, it requires fairly high degrees of specialized knowledge and skills to make a difference while many researchers compete. In contrast, it is also important to work on pioneering new themes that no one has focused on. This can add value to the field and help create a new world. However, in some cases, it is necessary to construct an evaluation system and data set from scratch, so it takes much effort to launch such research. I'd like to cultivate a high degree of expertise and flexible creativity and strive to solve various problems while striking the right balance between both types of efforts.

References

[1] T. Kitamura, Y. Nota, M. Hashi, and H. Takemoto, "Survey of Japanese Undergraduate and Graduate Students' Awareness of Clumsiness while Speaking," Journal of the Acoustical Society of Japan, Vol. 75, No. 3, pp. 118–124, 2019.

[2] H. Kameoka, "Generative Modeling of Voice Fundamental Frequency Contours for Prosody Analysis, Synthesis, and Conversion," NTT Technical Review, Vol. 13, No. 11, 2015. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201511fa2.html

■ **Interviewee profile**
Hirokazu Kameoka
Senior Distinguished Researcher, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E., M.S., and Ph.D. from the University of Tokyo in 2002, 2004, and 2007. He is currently a senior distinguished researcher and senior research scientist with NTT Communication Science Laboratories and an adjunct associate professor with the National Institute of Informatics. From 2011 to 2016, he was an adjunct associate professor with the University of Tokyo. His research interests include audio and speech processing and machine learning. He has been an associate editor for the IEEE/ACM Transactions on Audio, Speech, and Language Processing since 2015 and a member of the IEEE Audio and Acoustic Signal Processing Technical Committee since 2017.