Science and Technology Are the Collective Wisdom of Our Predecessors. It Is Our Mission—the Researchers of Today—to Make Them Even Better

Hirokazu Kameoka
Senior Distinguished Researcher, NTT Communication Science Laboratories

Abstract

People have various feelings and discomforts related to speech as typified by comments such as "The voice of a cartoon character differs from what I imagined," "I'm not confident speaking owing to my stuttering," and "I want to regain my voice that I lost due to illness or injury." Hirokazu Kameoka, a senior distinguished researcher at NTT Communication Science Laboratories, aims to create an environment in which all people can communicate comfortably by removing various barriers in communication through signal-processing and machine-learning technologies, which are key elements of artificial intelligence. We interviewed him about the progress in his research and what he enjoys most about his research activities.

Keywords: voice conversion, sound-source separation, crossmodal signal generation

Pursuing crossmodal signal-generation technology to augment communication functions

—This is our second interview with you. Can you tell us about the research you are conducting?

In our daily communication with others, we may be unable to speak as we wish due to physical barriers caused by disabilities, aging, or other factors; skill constraints, such as inability to speak foreign languages; and psychological barriers such as nervousness. I'm engaged in the development of signal-processing and machine-learning technologies to overcome these various forms of barriers and constraints concerning communication.

Communication involves a sender and receiver, and my aim is to build a system that converts the signals sent by the sender into expressions appropriate to the situation in real time in a manner that enables messages to be sent and received as desired by each party. Sound-source-separation technique, which complements the auditory function of the receiver, and

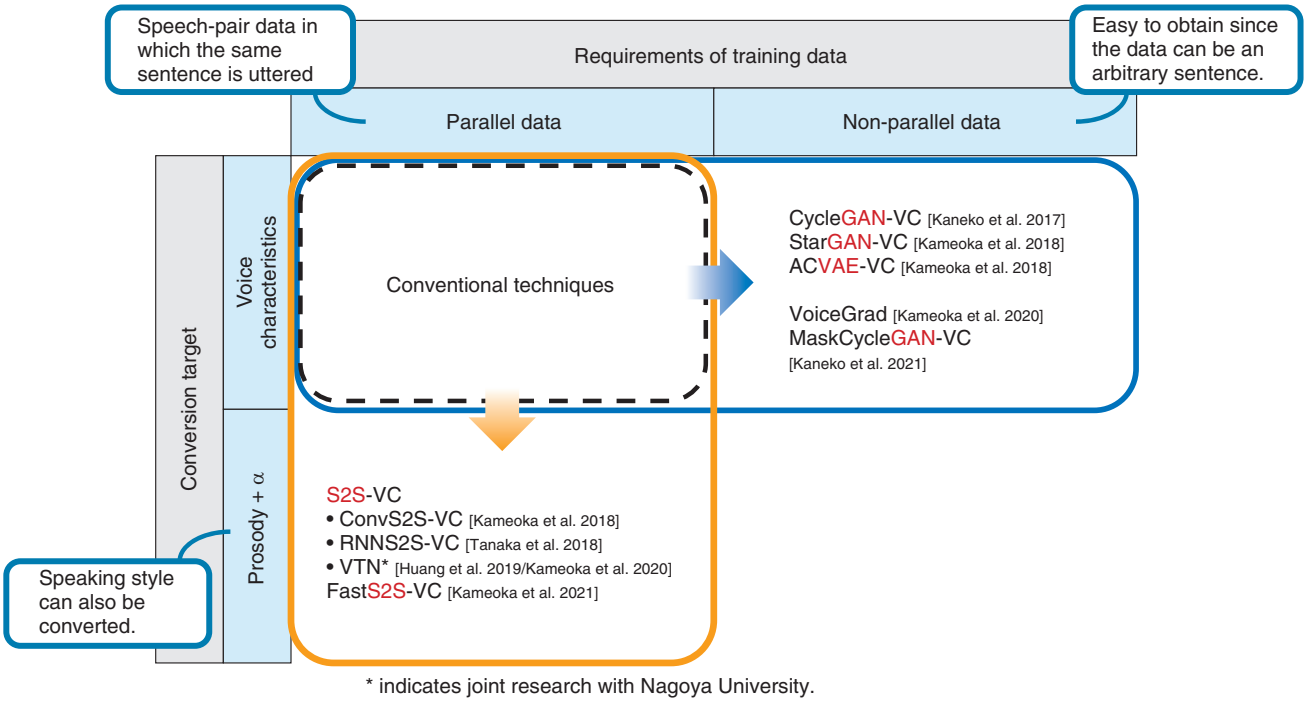


Fig. 1. Flexible voice conversion by using deep generative models.

voice-conversion technique, which complements the vocal function of the sender, are currently considered to be the core of such a system. Sound-source separation involves the decomposition of acoustic signals. Its purpose is to enhance the target sound source by extracting and separating multiple sound sources and removing reverberations and noise from an observed audio signal. The purpose of voice conversion is to change the features of speech to desired ones while preserving the speech content.

I'm also exploring the possibility of new communication methods that effectively use not only audio but also image, video, text, and many other types of media. For example, I'm considering enhancing communication by generating speech that matches a face and a face image that matches speech.

—How is your research on voice conversion with high quality and naturalness you mentioned last time going?

As I mentioned briefly in the previous interview, my research colleagues and I have developed many basic voice-conversion techniques and related peripheral techniques. We started researching voice conversion around 2016. At that time, the mainstream approach was to prepare two pieces of speech data uttering the same sentence, adjust the duration of one piece of the data so that the timing of each phoneme matched, and use the speech-pair data to train a *voice converter* to learn a conversion rule by which the features of the source speech are converted into the features of the target speech.

Such speech-pair data in which the same sentence is uttered by different speakers are called *parallel data*. This approach is effective when a large amount of parallel data can be collected. In many situations, however, parallel data cannot be easily obtained, for example, when the target speech is that of a particular celebrity. To address this issue, we turned our attention to deep generative models such as variational autoencoders (VAEs) and generative adversarial networks (GANs), which were attracting attention in fields such as machine learning and computer vision at the time. Using such deep generative models, we devised a non-parallel voice-conversion technique that can train a voice converter even from samples of source and target speech uttering arbitrary sentences (**Fig. 1**). Since this technique does not require parallel data for training, it is expected to greatly expand the use scenarios of voice conversion.

Most conventional techniques at the time were

limited to converting speech features such as voice characteristics and were not able to convert speaking styles such as intonation and rhythm. We wanted to develop a technique for converting speaking style as well as voice characteristics, so we focused on a framework called sequence-to-sequence (S2S) learning, which had been shown to be very effective in regard to machine translation, speech recognition, and text-to-speech synthesis. This framework is used for training neural-network models that transform one vector sequence into another (with different lengths) while capturing long-term dependencies. The key point regarding this framework lies in the model structure called the attention mechanism, which makes it possible to learn conversion rules as well as association rules between the elements of the source and target speech-feature sequences. As far as we knew at the time, few studies had attempted to apply S2S learning to voice conversion. I remember how excited my colleagues and I were when we tried it out and found through our experiments that, as we had hoped, it could flexibly convert not only voice characteristics but also intonation and speaking rhythm.

Almost without exception, current state-of-the-art voice-conversion techniques consist of two steps: (i) extracting a sequence of speech-feature vectors from the source speech and converting it into a mel-spectrogram and (ii) generating a speech waveform from the sequence of the converted mel-spectrogram. The aforementioned VAE, GAN, and S2S learning are all used for the first step of speech-feature conversion. For the second step, waveform generation, the waveform generator in the context of a neural network is called a *neural vocoder*. As those familiar with speech-related research probably know, a high-quality waveform-generation method called WaveNet was announced by DeepMind in 2016. Since then, many researchers have been actively working on improving its speed, quality, and training efficiency. Although our main focus has been on feature-conversion techniques, we have recently begun to focus on research targeting higher quality and lower delay in regard to waveform generation.

Each of the above-mentioned accomplishments has been reported at international conferences, such as International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and Interspeech, and in academic journals such as IEEE Transactions on Audio, Speech, and Language Processing, and the total number of citations exceeds 1000 times. I think that our recent activities are gradually gaining recognition.

Building a machine-learning infrastructure for improving the accuracy, efficiency, and flexibility of voice conversion and sound-source-separation techniques

—These accomplishments are good news for those who have communication problems. Can you tell us about specific applications?

Applications of voice conversion that we have had good experimental results with include speaker-identity conversion, English-accent conversion, whisper-to-speech conversion, electrolaryngeal (EL)-speech enhancement, emotional-expression conversion, and stuttered-speech conversion.

I believe that English-accent conversion helps facilitate conversation by converting the speaker's English into an accent that is easier for the listener to understand. For example, for Japanese people (depending on the person, naturally), the so-called "Japanese English accent" may be easier to understand than the native speaker's accent, so it may be desirable to add a Japanese English accent to the native speaker's speech.

Whisper-to-speech conversion is a task that aims to convert a whispered voice into a normal speech sound. For example, you want to make a phone call or hold an online meeting in a situation where you are not comfortable speaking, such as on a crowded train or in a coffee shop. If this voice-conversion technique is available, you can speak in a whisper so that people around you cannot hear your voice, but your voice will be transmitted to the other party as normal speech.

EL-speech enhancement is a task for converting EL speech into natural-sounding speech. EL speech is speech produced using an electrolarynx by a person with a speech impediment who has lost their vocal cords due to laryngectomy or other surgery and sounds monotonous and mechanical. This voice-conversion technique makes it possible to convert such EL speech into speech like those of able-bodied people. It has also been found that it is possible to change the expression of emotions by changing the speaking style, and to some extent, it is possible to automatically omit stutter and filler words such as "Ah…" and "Um…" to make the entire speech fluent. Audio samples of these voice conversions are available on our demo sites [1–3].

Regarding sound-source separation, we have

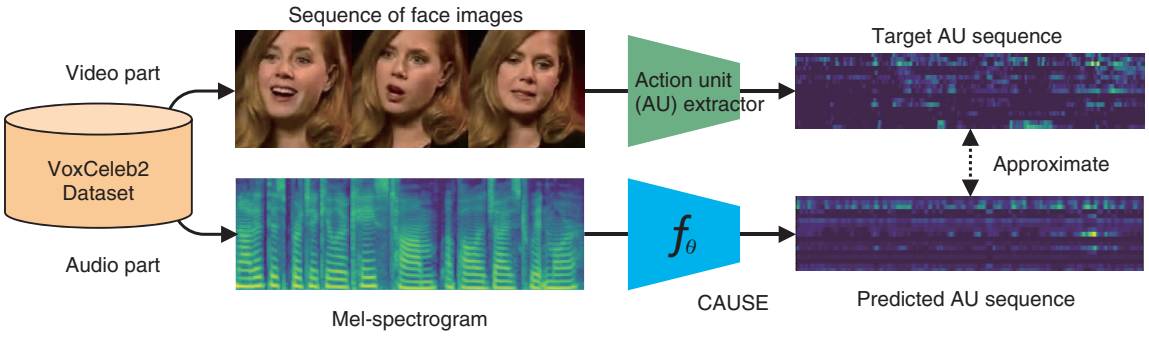


Fig. 2. Training crossmodal action unit sequence estimator (CAUSE).

previously proposed a multi-channel source-separation technique that uses the aforementioned VAE to model the source signals and have been studying ways to increase the speed and accuracy of this technique. The research field of multi-channel source separation has dealt with mixtures of a maximum of five sound sources. However, we have demonstrated that our technique can separate mixed signals from up to 18 sound sources with high accuracy, achieving unprecedented performance. Audio examples of this multi-channel source separation are also available on our demo site [4].

—Listening to the comparison between conventional techniques and your techniques, the listener can clearly hear the superiority of your techniques.

I'm thankful you think so. In addition to these studies, we are investigating crossmodal signal-generation technology that uses media other than sound to generate and control sound or uses sound to generate and control signals other than sound. For example, this technology can generate voice that matches a face image or generate a face image that matches the voice (**Fig. 2**). We aim not only to enrich voice communication but also enable intuitive control of communication to enhance communication functions. Specifically, we investigated crossmodal face-image generation, which predicts a speaker's face from speech alone and outputs the predicted face as an image, and crossmodal voice-characteristics conversion, by which the target voice characteristics can be specified via a face image (instead of a speaker identity). The demonstration of these techniques was well received at the NTT Communication Science Laboratories Open House 2022 and were covered by various media.

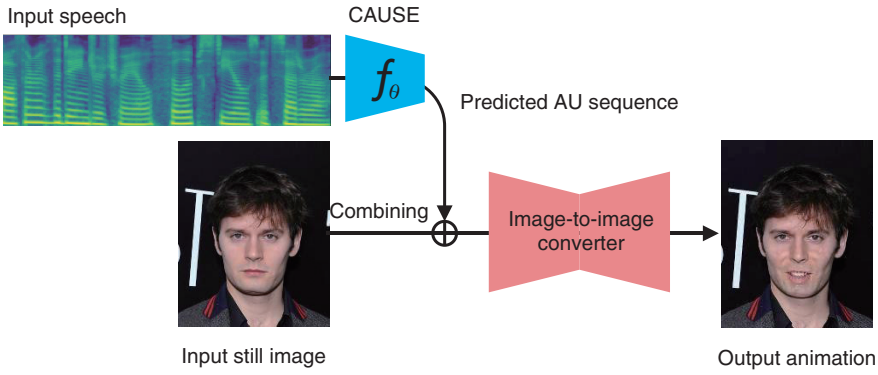
We also attempted to estimate a sequence of action units (AUs) (i.e., facial-muscle motion parameters) of the speaker from speech alone. To the best of our knowledge, no one else had attempted such an estimation, so we had no idea how accurate it would be. Through experimentation, we found that it was possible to estimate the AU sequence to some extent. By using an image-to-image converter and the AU sequence estimated from speech, it is possible to move the facial expression of a still face image in accordance with the speech (**Fig. 3**). If we improve the accuracy of this AU sequence estimation and make good use of it, we will be able to provide visual feedback on how one's speaking style and voice characteristics affect the conversation partner, which will be useful for improving one's presentation and customer-service skills.

Examples of each of these studies on crossmodal signal generation are available online on our demo sites [5, 6].

Keep "Think like an amateur, do as an expert" in mind

—Would you tell us what has been important to you as a researcher?

The title of Dr. Takeo Kanade's book, "Think like an amateur, do as an expert," is one of the mottos I always keep in mind as a researcher. As one's specialized knowledge increases, one tends to fall into the trap of "research for research's sake" and set research themes that seem like nitpicking. There is a chance that such a theme could develop into an important research theme. However, I try to ask myself as calmly as possible whether I find the research theme interesting and whether it is really useful to society.



Face images are acquired from VoxCeleb2 Dataset*1 and CelebA Dataset*2.
*1 J. S. Chung, A. Nagrani, and A. Zisserman: "VoxCeleb2: Deep Speaker Recognition," Proc. of Interspeech, pp. 1086–1090, 2018.
*2 Z. Liu, P. Luo, X. Wang, and X. Tang: "Deep Learning Face Attributes in the Wild," Proc. of ICCV, pp. 3730–3738, 2015.

Fig. 3. Facial-expression control from speech by using CAUSE and image-to-image converter.

For example, in my research on augmenting communication functions, I'm constantly thinking about whether there are any discomforts or inconveniences in our daily lives that we are not usually aware of and whether there are ways to overcome them. We are now entering an era of rapid development and upheaval in the fields of artificial intelligence (AI) and machine learning, and while it is obviously important to always follow the latest trends and research, I believe it is also important to remain calm and listen to our inner voice.

While researching AI, I've been reminded of the importance of doing as much hands-on work, namely, coding and experimentation, as possible. I have been rather good at research in which I take my time to formulate a hypothesis and a theory for each problem then develop a solution; in contrast, I feel that in research using deep learning and neural networks, it is important to repeat the process of verifying a hypothesis through experiments over and over again at a rapid pace. The behavior of a neural network is not always as intuitively imagined, and it may feel like you are dealing with a living being. I feel that the more I deal with it, the more I understand it, so now I try to code and conduct experiments at least once a day. In deep learning, many training samples are input to a neural network, and the network learns behaviors that match the training data. Through a large amount of coding and experimentation, I get the feeling that I'm also learning the behavior of neural networks, which is very refreshing and interesting.

—What are your future plans and what would you like to say to the younger generation of researchers?

My first plan is to conduct more research on voice conversion to meet the demands of using sensory language. For example, if a "cute voice," "gentle voice," or "stately voice" is requested, the voice conversion will convert speech to such a voice. For the voice conversion that we have been researching, the voice characteristics of the conversion target were easy to uniquely define; however, as the examples I mentioned show, the definition of a sensory language is ambiguous and varies from person to person. The key is how to quantify the sensory language the definitions of which are ambiguous and subjective, so I'm currently working on this issue with my colleagues.

When this voice-conversion system is put into practical use, we cannot deny the possibility that it could be misused to impersonate other people's voices in a malicious way. We thus intend to conduct research to prevent the misuse of voice-conversion systems.

I also think it is necessary to create a white-box model for practical use. In the example of voice conversion, if a system that converts a voice in real time is actually used, it must be guaranteed that no unexpected conversions will occur. This is because some conversions may give the listener an impression that is contrary to the speaker's intention. Neural networks are very good at learning behaviors that match the training data; however, their internal structure is a

black box that makes it difficult to predict their behavior when data that do not exist in the training data are input, so they are not always easy to control. Therefore, I believe that we must pursue research on model structures and control mechanisms to ensure that voice-conversion models can be used with confidence.

Finally, to younger generations of researchers, I believe that the mission of researchers is to make the world a better place. I hope that all researchers will cooperate, exercise their wisdom to meet the hidden human desire for convenience and comfort, and make the world a safer, more secure, and happier place.

You will face a lot of hard times when you are doing research. This may sound cliché, but I think it is important to enjoy your research instead of focusing only on the negative aspects. Many people at NTT laboratories are now working remotely, and I especially encourage those people to have online meetings frequently, which might just be for chatting, and create many opportunities to communicate with their colleagues and seniors. It will be fun and stimulating to talk with them. I also want you to remember to respect each other as researchers. I often work with students and have the opportunity to check their manuscripts, and sometimes I see statements that needlessly downplay conventional technologies so as to assert the superiority of their proposed technology. However, science and technology have been built up little by little by the collective wisdom of our predecessors, and it is our job as researchers to try to make them even better. Therefore, I want you to look at previous research from the viewpoint of finding the positive aspects and further improving them.

References

[1] S2S-VC (sequence-to-sequence voice conversion), https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/s2s-vc/index.html
[2] ACVAE-VC (non-parallel many-to-many voice conversion (VC) method using an auxiliary classifier variational autoencoder), https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/acvae-vc3/index.html
[3] StarGAN-VC (nonparallel many-to-many voice conversion (VC) method using star generative adversarial networks), https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/stargan-vc2/index.html
[4] Audio examples of MVAE and FastMVAE2, https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/mvae-ss/index.html
[5] Crossmodal voice conversion, https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/crossmodal-vc/index.html
[6] CAUSE (crossmodal action unit sequence estimation/estimator), https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/cause/index.html

■ **Interviewee profile**

Hirokazu Kameoka 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 member of the IEEE Audio and Acoustic Signal Processing Technical Committee since 2017.