Communication with Desired Voice

Kou Tanaka, Takuhiro Kaneko, Nobukatsu Hojo, and Hirokazu Kameoka

Abstract

We convey/understand our intentions/feelings through speech. We also change the impression we want others to have of us by controlling our voice, including intonation, speaking characteristics, and rhythm. Unfortunately, the types of voice and voice controllability an individual can generate are limited. In this article, we introduce the challenges with conventional voice-conversion techniques and introduce improved voice conversion techniques with the following question in mind, "What can be done when the voice is combined with deep learning, which has been advancing in recent years?" Finally, we look at the future of deep learning and speech generation and conversion.

Keywords: deep learning, signal processing, voice conversion

1. Non/para-linguistic information conversion

Speech is one of the fundamental methods for conveying not only linguistic information but also non/para-linguistic information. We can control the impression we want to give to others of us by changing our voice properties, including intonation, speaking characteristics, and rhythm. However, the expression of the speech generated by an individual is limited due to physical or psychological constraints. Voice-conversion techniques help us go beyond this limitation and express ourselves as we wish by converting our voice into the desired voice (**Fig. 1**). There are a wide variety of applications of voice conversion, i.e., speaker identity conversion, assistance for people with vocal disabilities, emotion conversion, and pronunciation/accent conversion for language learning. These applications have several requirements according to the usage scenarios. We particularly focus on the quality of generated speech, the amount of training data, non-parallel data* training, real-time conversion, and not only voice timbre but also suprasegmental features conversion.

2. Research on speech × deep generative model

A typical voice-conversion technique is statistical voice conversion based on Gaussian mixture models [1]. By using the time-aligned acoustic parameters as the training data, we train a model describing the joint probability of acoustic parameters of the source and target speech to obtain a mapping function from the source acoustic parameters to target acoustic parameters. Within the framework that requires the parallel data described above, several methods using a neural network and an example-based method, such as non-negative matrix factorization, have recently been proposed. Although these methods improve conversion quality and controllability, there are still disadvantages, i.e., 1) speech-sample pairs have the same content as the training data, 2) convertible acoustic parameters are limited to the voice timbre, and 3) it is too easy to distinguish the converted speech from normal speech because classical vocoders are used for generating speech waveforms from the given acoustic parameters.

Variational auto-encoder (VAE), generative adversarial networks (GANs), and sequence-to-sequence model (Seq2Seq) have been proposed for tasks such as image recognition and natural-language processing. The auto-regressive model, which is a part of Seq2Seq, VAE, and GANs, is treated as the main deep generative model and has been proven to be

* Non-parallel data: Non-parallel data does not restrict utterance content of input speech and target speech.

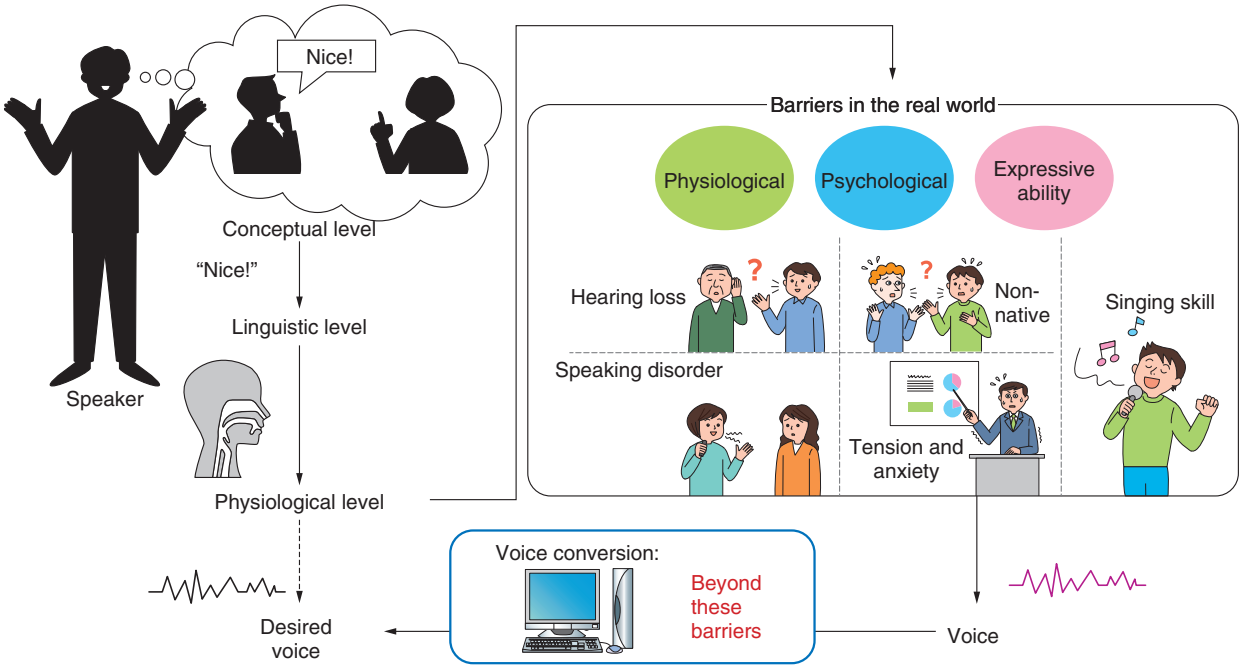


Fig. 1. Speech-generation process and voice conversion.

outstanding in various research tasks. In the research field of the machine translation in mid-2015, an attention mechanism [2] was proposed and has attracted attention due to its effectiveness.

At NTT Communication Science Laboratories, to develop a versatile voice-conversion system that can be flexibly applied to various usage scenarios and overcome the problems with the conventional voice-conversion techniques mentioned above, we proposed the following functions based on the main deep generative models: 1) voice-series conversion for converting not only voice timbre but also prosody and accent, which are long-term dependent features, 2) non-parallel voice conversion trained using non-parallel training data with no restrictions on the utterance content, and 3) waveform post-filter for directly modifying synthetic speech waveforms to real voice waveforms not on the acoustic parameters space but on the waveform space. With these functions, we achieved high-quality, high-efficiency, and real-time voice conversion. In addition, as wider research, we also developed a cross-modal voice-conversion function for converting voice by using the face image of the target speaker.

3. High-quality and stable training-voice conversion based on Seq2Seq

We briefly introduce our voice-series conversion function [3]. Unlike words (symbols) that are treated as discrete values in natural-language processing, speech is observed as a series of continuous values. While conventional voice-conversion techniques using series conversion can be expected to significantly improve the quality of generated voice, the difficulty in learning when input and output are continuous value series is an issue. To address this issue, the mainstream approach with conventional speech-sequence conversion is to combine automatic speech recognition (ASR) and text-to-speech synthesis. This approach uses ASR to recognize symbols such as words from given speech, converts the recognized symbol sequence into a symbol sequence of the target speaker, and synthesizes the desired speech from the converted symbol sequence. Since conversion is executed using symbols that are discrete values, the training is relatively stable. However, not only speech but also text labels are required to train the model. In addition, it is not easy to convert laughter, which makes it difficult to convert speech into text.

NTT Communication Science Laboratories has achieved voice-sequence conversion that enables

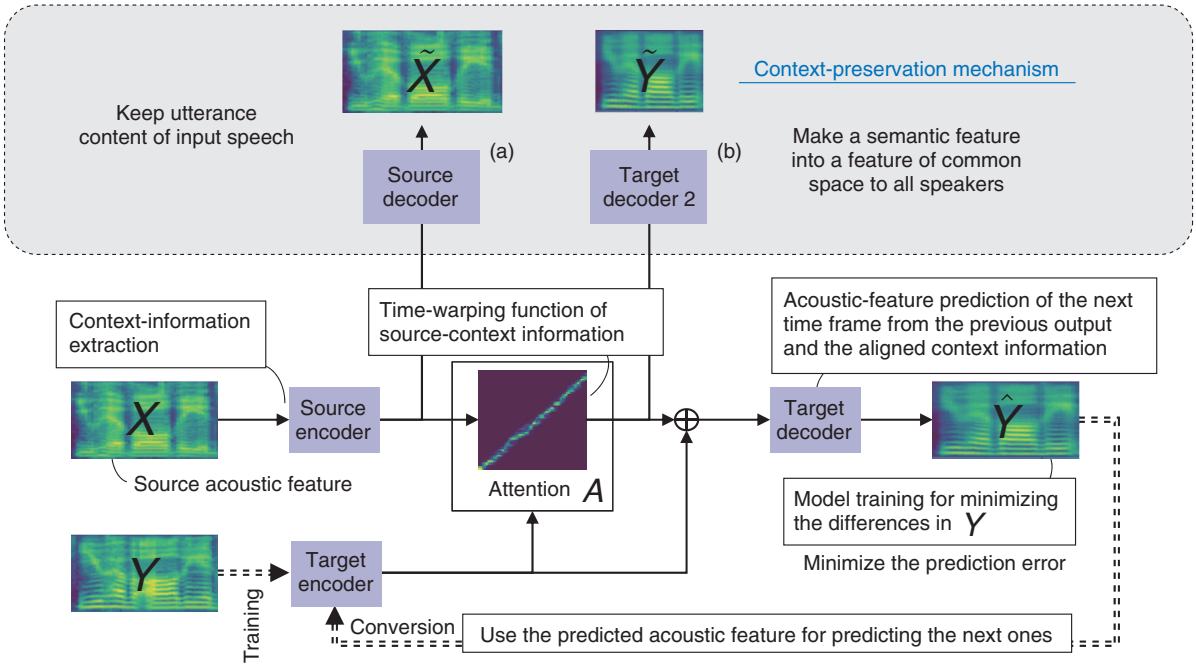


Fig. 2. Seq2Seq voice conversion.

stable training without using text labels. We developed a context-preservation mechanism, shown in **Fig. 2** and described later, to train all modules from only audio data. Instead of text labels, we use this mechanism for adequate model training to convert utterance context. In our context-preservation mechanism, the left module (Fig. 2(a)) is used to restore the input speech. Namely, the utterance content of the input speech is kept in the conversion process. This is called auto-encoding. The right module (Fig. 2(b)) is used to predict the acoustic parameters of the target speech and converts the semantic features into predictable features for both the input voice and the target voice. In other words, it is a constraint that the input speech is mapped to the common space for all speakers. Unlike speech recognition, the important semantic features are automatically determined from the training data. Therefore, richer semantic features rather than the symbols are extracted from the input speech, enabling more accurate conversions.

4. Waveform modification for converting synthetic speech to real speech

Another issue with conventional voice-conversion techniques is that they are affected by the accuracy of the waveform synthesizer when synthesizing waveforms from acoustic parameters. Therefore, we may be able to determine that voices have been artificially synthesized. We aim to eliminate such synthetic-voice likeness and directly modify waveforms to normal voice waveforms with higher sound quality and real-voice likeness.

There are two main difficulties in modifying a speech waveform in deep learning. One involves the sampling rate. For example, 16-kHz sampling audio has 16,000 samples in one second. It is easy to understand the difficulty in alignment, which finds the correspondence point of the synthesized speech waveform and real speech. The other involves phase information of speech. In conventional voice-conversion techniques, it is standard practice to discard phase information and handle only amplitude information.

NTT Communication Science Laboratories is tackling these problems by directly modifying speech waveforms using deep learning [4]. A converter, which is the core module in the model, is trained according to two criteria, as shown in **Fig. 3**. One is minimization of discrimination error, which is a criterion for adversarial training. The discriminator recognizes the difference between processed speech and non-processed speech that is difficult to quantify, and the converter is trained to eliminate as much of

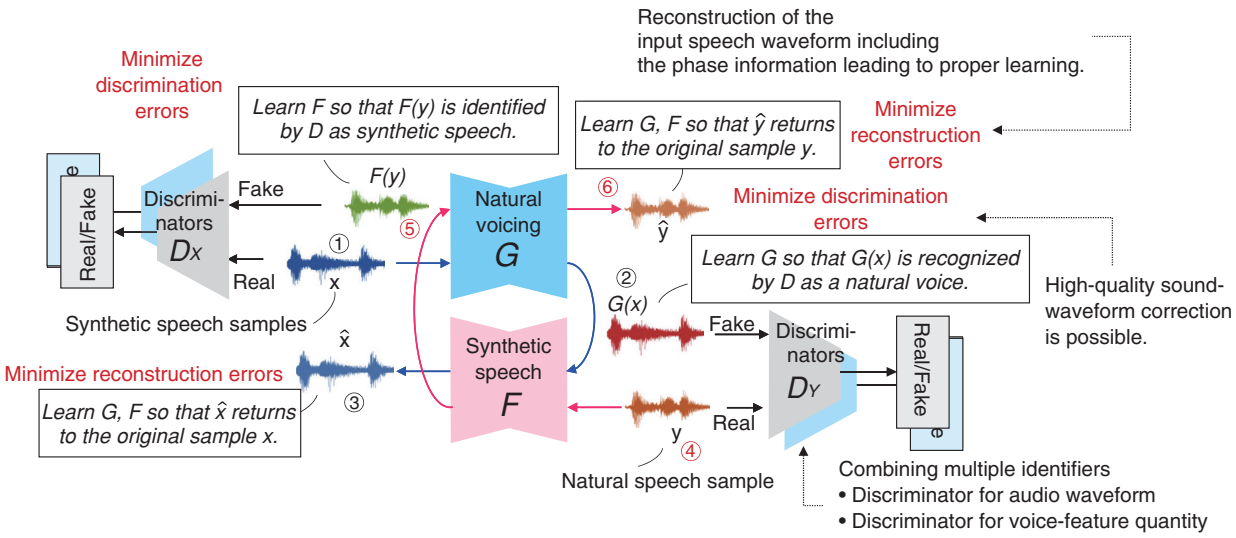


Fig. 3. Synthetic-to-natural speech-waveform modification.

the difference as possible. We also proposed using multiple classifiers instead of one classifier for recognizing the difference among various views and enabling high-quality modification.

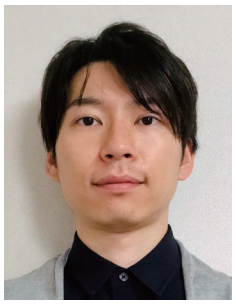
The other criterion is minimization of reconstruction error, which is a criterion for cyclic models. By converting synthetic voice into natural voice then converting it into synthetic voice again, the input synthetic voice has to be restored. The key is that the amplitude information as well as phase information have to be restored. Namely, it is possible to properly process the phase information and make the training process stable. Moreover, this cyclical adversarial model does not require parallel data of synthetic speech and normal speech. By appropriately collecting synthetic speech and natural speech, we can train the desired model.

5. Future work

NTT Communication Science Laboratories will continue to improve and demonstrate technology for converting more diverse voices. To date, we have succeeded in high-quality conversion when we have a specific target speaker, e.g., "I want to speak with a target speaker's voice." However, perceptual score conversion is still challenging, e.g., "I want to speak with a cute voice" and "I want to speak with a sterner voice." To enable such conversion, it is necessary to model the perceptual space of speech and interpolate latent variables in that space. This is why perceptual score conversion is challenging. We will promote research and development toward a versatile voice-conversion system that can flexibly work on various usage scenarios and respond to all user requirements.

References

[1] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory," IEEE Trans. TASLP, Vol. 15, No. 8, pp. 2222–2235, 2007.
[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," Proc. of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, May 2015.
[3] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "AttS2S-VC: Sequence-to-sequence Voice Conversion with Attention and Context Preservation Mechanisms," Proc. of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, May 2019.
[4] K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka, "Synthetic-to-natural Speech Waveform Conversion Using Cycle-consistent Adversarial Networks," 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, Dec. 2018.



Kou Tanaka
Research Scientist, NTT Communication Science Laboratories.

He received a B.E. from Kobe University in 2012 and an M.E. and D.E. from the Nara Institute of Science and Technology in 2014 and 2017. He was a research fellow of the Japan Society for the Promotion of Science from 2015 to 2017. His research interests include speech-signal processing with a strong focus on deep generative models.

Takuhiro Kaneko
Distinguished Researcher, NTT Communication Science Laboratories.

He received a B.E. and M.S. from the University of Tokyo in 2012 and 2014. He has been pursuing his Ph.D. degree at the University of Tokyo from 2017. His research interests include computer vision, signal processing, and machine learning. He is currently working on image generation, speech synthesis, and voice conversion using deep generative models. He received the ICPR2012 Best Student Paper Award at the 21st International Conference on Pattern Recognition in 2012.

Nobukatsu Hojo
Research Scientist, NTT Communication Science Laboratories.

He received a B.E. and M.E. from the University of Tokyo in 2012 and 2014. He joined NTT Media Intelligence Laboratories in 2014, where he engaged in the research and development of speech synthesis. He is a member of the Acoustical Society of Japan, the International Speech Communication Association, and the Institute of Electronics, Information and Communication Engineers of Japan.

Hirokazu Kameoka
Senior Distinguished Researcher, 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 an adjunct associate professor at the National Institute of Informatics. From 2011 to 2016, he was an adjunct associate professor at the University of Tokyo. His research interests include audio, speech, and music signal processing and machine learning. He has been an associate editor of the IEEE/ACM Transactions on Audio, Speech, and Language Processing since 2015, a member of IEEE Audio and Acoustic Signal Processing Technical Committee since 2017, and a member of IEEE Machine Learning for Signal Processing Technical Committee since 2019. He received 17 awards, including the IEEE Signal Processing Society 2008 SPS Young Author Best Paper Award. He is the author or co-author of about 150 articles in journal papers and peer-reviewed conference proceedings.