NEWS RELEASE

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

December 25, 2018

New method for high-speed synthesis of natural voices

Neural source-filter model uses neural networks to update classical speech-synthesis methods

The research team in the Digital Content and Media Sciences Research Division, the National Institute of Informatics (NII, Director General: Dr. Masaru Kitsuregawa, Chiyoda-ku, Tokyo, Japan) - Researcher by Special Appointment Xin Wang, Assistant Professor by Special Appointment Shinji Takaki, and Associate Professor Junichi Yamagishi - has developed the method of *neural source-filter* (NSF) models for high-speed, high-quality voice synthesis. This new technique, which combines the recent deep-learning algorithms and a classical speech production model dated back to the 1960s, is capable not only of generating high-quality voice waveforms—closely resembling the human voice—but also of conducting stable learning via neural networks.

Background

To date, many speech synthesis systems have adopted the *vocoder* approach, a method for synthesizing speech waveforms that is widely used in cellular-phone networks and other applications. However, the quality of the speech waveforms synthesized by these methods has remained inferior to that of the human voice. In 2016, an influential overseas technology company(*1) proposed WaveNet—a speech-synthesis method based on deep-learning algorithms—and demonstrated the ability to synthesize high-quality speech waveforms resembling the human voice. However, one drawback of WaveNet is the extremely complex structure of its neural networks, which demand large quantities of voice data for machine learning and require parameter tuning and various other laborious trial-and-error procedures to be repeated many times before accurate predictions can be obtained.

Overview and achievements of the research

One of the most well-known vocoders is the *source-filter vocoder*(*2), which was developed in the 1960s and remains in widespread use today. The NII research team infused the conventional

National Institute of Informatics

Research Organization of Information and Systems
National Institute for Informatics
Communications
1-2 Hitotsubashi, 2-chome
Chiyoda-ku, Tokyo 101-8430 JAPAN
Direct: +81(0)3-4212-2164 FAX : +81(0)3-4212-2150
E-Mail: media@nii.ac.jp

Web: https://www.nii.ac.jp
Twitter: @jouhouken
facebook: https://www.facebook.com/jouhouken

source-filter vocoder method with modern neural-network algorithms to develop a new technique for synthesizing high-quality speech waveforms resembling the human voice. Among the advantages of this *neural source-filter* (NSF) method is the simple structure of its neural networks, which require only about 1 hour of voice data for machine learning and can obtain correct predictive results without extensive parameter tuning. Moreover, large-scale listening tests have demonstrated that speech waveforms produced by NSF techniques are comparable in quality to those generated by WaveNet.

Future outlook

Because the theoretical basis of NSF differs from the patented technologies used by influential overseas ICT companies, the adoption of NSF techniques is likely to spur new technological advances in speech synthesis. For this reason, the source code implementing the NSF method has been made available to the public at no cost, allowing it to be widely used.

Source code, trained NSF models, and the actual NSF-synthesized speech samples (both Japanese and English) are available at the following sites:

Source code:
https://github.com/nii-yamagishilab/project-CURRENNT-public

Trained models (may be executed to generate English-language voices):
https://github.com/nii-yamagishilab/project-CURRENNT-scripts

Voice samples (Japanese or English):
https://nii-yamagishilab.github.io/samples-nsf/index.html

Associate Professor Junichi Yamagishi makes the following comment:

“We hope that our NSF method will create new business opportunities for Japanese AI firms that use voice-based interfaces. For future work, we will work to make the method available for use as a real-time voice-synthesis engine in a wide variety of systems. We are also planning to add speaker adaption and other related features to the NSF methods.”

Please visit the following page for comparisons of actual human voices to voice waveforms produced by source-filter vocoder methods, by WaveNet, and by NSF.

https://youtu.be/yr_xMq1gxKY

*It is explained only in Japanese in this movie.

About this research project

The research described here was supported by the Japan Science and Technology Agency under CREST JPMJCR18A6 and by the Japan Society for the Promotion of Science under Grants-in-Aid

for Scientific Research “KAKENHI” 16H06302, 16K16096, 17H04687, 18H04120, 18H04112, and 18KT0051.

Paper title and authors

Title: Neural source-filter-based waveform model for statistical parametric speech synthesis

Authors: Xin Wang, Shinji Takaki, Junichi Yamagishi

Publication for: International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2019 (Submitted)

Date announced: October 30, 2018 (ArXiV: https://arxiv.org/abs/1810.11946)

Media Contact:

Research Organization of Information and Systems

National Institute for Informatics

Publicity Team

TEL: +81(0)3-4212-2164; E-mail: media@nii.ac.jp

(*1) The Google subsidiary DeepMind, also known for developing the AlphaGo artificial-intelligence system for the game of Go.

(*2) A speech production model reported in 1960 by Dr. Gunnar Fant. The model approximates the speech production process by considering both *sources*, such as the human glottis, and linear acoustic *filters*, such as the human vocal tract.