New method for high-speed synthesis of natural voices

February 6 2019

A research team at the National Institute of Informatics (NII/Tokyo, Japan) including Xin Wang, Shinji Takaki and Junichi Yamagishi has developed a neural source-filter (NSF) model for high-speed, high-quality voice synthesis. This technique, which combines 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.

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 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, which

was developed in the 1960s and remains in widespread use today. The NII research team infused the conventional 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 one 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.

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.

More information: Neural source-filter-based waveform model for statistical parametric speech synthesis. arxiv.org/abs/1810.11946

Source code: github.com/nii-yamagishilab/pr … ject-CURRENNT-public

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

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

Provided by Research Organization of Information and Systems