様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ４ 月 ２７ 日現在

機関番号：６２６１５
研究種目：研究活動スタート支援
研究期間：2019～2020
課題番号：１９Ｋ２４３７１

研究課題名（和文）One model for all sounds: fast and high-quality neural source-filter model for speech and non-speech waveform modeling

研究課題名（英文）One model for all sounds: fast and high-quality neural source-filter model for speech and non-speech waveform modeling

研究代表者
Ｗａｎｇ　Ｘｉｎ（Wang, Xin）

国立情報学研究所・コンテンツ科学研究系・特任助教

研究者番号：６０８４３１４１

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：デジタルシステムで自然な音声波形を生成する方法は音声科学の分野において基本的な研究テーマの一つである。本研究では、古典的な信号処理方法と最新の深層学習技術を組み合わせることにより、ニューラルソースフィルター波形モデル（NSF）と呼ばれるモデルを提案した。NSFモデルは、グーグルに提案されたWaveNetモデルよりもはるかに高速で高品質の波形を生成できることが実証された。 また、NSFモデルを拡張して、harmonic-plus-noiseという古典的な音声モデルを組み込むことができることも実証された。最後に、NSFモデルを音楽オーディオに適用できることも実証された。

研究成果の学術的意義や社会的意義
Deep learningにより音声波形モデリング技術は近年盛んに研究されている。深層学習手法だけを使用して多くのモデルが提案されている一方で、本研究は深層学習と古典的な信号処理技術の組み合わせることにとりニューラルソースフィルター波形モデル（NSF）と呼ばれるモデルを提案した。 提案されたモデルは、深層学習と信号処理の方法を組み合わせるの方法を示しています。 そして、提案されたモデルは実際のアプリケーションで使用されています。

研究成果の概要（英文）：How to generate natural-sounding speech waveform from a digital system is a fundamental question in speech science. By combining classical speech science, signal processing methods, and recent deep-learning techniques, this research project proposes a family of neural waveform model called neural source-filter waveform (NSF) models. It was demonstrated that the proposed NSF models can produce high-quality waveforms at a much faster speed than the commonly used WaveNet models. It was also demonstrated that the NSF models can be extended to incorporate other classical methods from the speech modeling field, including harmonic-plus-noise speech model. Finally, it was demonstrated that the NSF model can be applied to music instrumental audios, showing its flexibility and potential in modeling speech and non-speech sounds.

研究分野：知覚情報処理

キーワード：Speech synthesis　Waveform modeling　Deep learning　Neural network

１．研究開始当初の背景 Research background

Although humans can easily utter natural speech, **how to enable a machine to produce natural speech waveforms is a fundamental research topic in speech science.**

When this project started in 2019, there had been two research directions on the issue above. The first one is expert-knowledge-based. For example, the source-filter speech production theory [1] argues that the speech waveform is produced by generating a source signal with a specific fundamental frequency (F0) and then filtering it into the output waveform using a filter (top panel of the Fig.1). Based on this theory, many classical speech vocoders have been proposed. Although they can generate understandable waveforms, the quality is far from being good.

The second stream of research mainly relies on recent deep learning methods (middle panel of Fig.1). With a large amount of data, some neural networks such as the WaveNet [2] can generate speech waveforms with a quality close to that of human speech. However, these neural models discard the classical speech knowledge and directly use one deep network to model the waveform in the time domain. They sample one output waveform value, feedback it to the network's input, and generate the next waveform value. However, it is extremely slow to generate the waveform point by point in such an auto-regressive (AR) process.

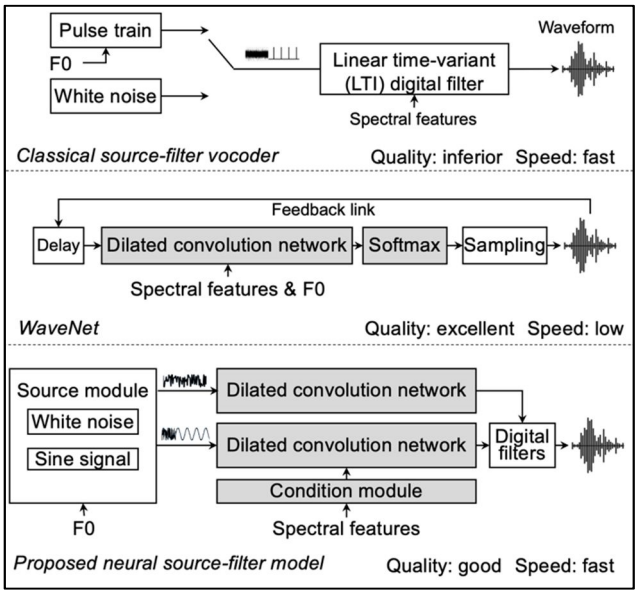

Fig.1 Existing and proposed approaches. White and grey blocks denote classical signal processing and recent deep learning-based techniques, respectively.

Facing the two streams of research work, **our key scientific question is: how we can combine the speech science knowledge and deep learning for a neural waveform model that generates high-quality waveforms at a fast speed.** Another scientific question is **whether we can apply the same model to sounds that are not human speech**.

The above background motivated us to propose a family of **neural source-filter waveform models (NSF)** that combines the neural network (grey blocks in Fig.1) with classical digital filters (white blocks in Fig.1). The proposed family of NSF models is the core contribution of this research project, and it will be detailed in the rest of this report.

２．研究の目的 Research Motivation and goals

While the general goal is to define waveform models that combine classical speech production theory and recent deep learning techniques, we defined three specific goals:

Goal 1: Accelerate generation speed
How can we revise the slow neural waveform model using the knowledge of speech science (e.g., the source-filter theory) and accelerate the waveform generation speed? We also need to ensure that the quality of the generated waveform will not severely degrade.

Goal 2: Improve the quality of generated waveform
If Goal 1 was achieved, can we improve the quality of the generated waveforms from our neural waveform model by using more advanced speech modeling theory?

Goal 3: Beyond speech waveform
If Goal 2 was achieved, can we use the new model for non-speech waveforms, such as music signals? Since deep learning is data-driven, it may be possible to model non-speech waveforms using the proposed models.

3．研究の方法 Research methods

For Goal 1, we combined classical signal processing techniques with deep learning models in the proposed NSF models. This is possible because many signal processing techniques can be interpreted from the deep-learning point of view. For example, short-time Fourier transform (STFT) can be interpreted as matrix transformation in the complex-valued domain, and it can be plugged into deep learning models. This is used in the proposed NSF models.

For Goal 2, we combined classical speech modeling methods, including the harmonic-plus-noise speech model [3] and glottal excitation theory [4]. These theories improve the NSF models and allow them to better model speech sounds for different applications: multi-speaker, speech sound in reverberation condition, and so on.

For Goal 3, it is straightforward to apply the proposed NSF models to non-speech sounds. In this project, we were particularly interested in applying them to music instrumental audios, including audios from strings, woodwinds, and brass instruments.

4．研究成果 Research outcome

✔ **Goal 1: NSF can generate high-quality speech waveforms with more than real-time speed (cf. papers published in ICASSP 2019 and IEEE Trans. ASLP 2020):**

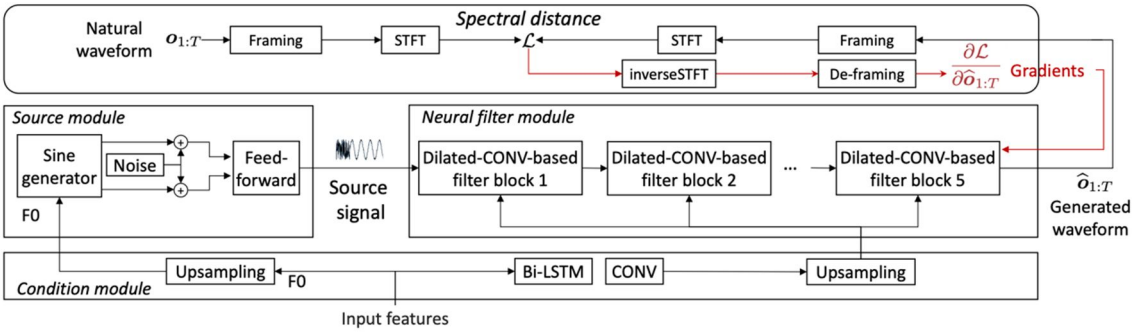

Fig.2 Diagram of the 1st neural source-filter-waveform model. White and grey blocks denote classical signal processing and recent deep learning-based techniques, respectively.

Based on the research methods for Goal 1, we combined the advanced dilated convolution networks with the STFT-based training criterion (i.e., spectral distance), which lead to **the first version of our proposed neural source-filter waveform model** in Fig.2**.** We call it **baseline NSF (b-NSF)** in our research papers.

We evaluated the quality of the generated waveforms from the b-NSF on a large-scale Japanese Female voice database, with classical speech vocoders and the WaveNet as reference. As the box plot in Fig.3 shows, the mean of the quality scores (white dot) of the b-NSF is better than the classical vocoder (i.e, WORLD) while being close to the WaveNet.

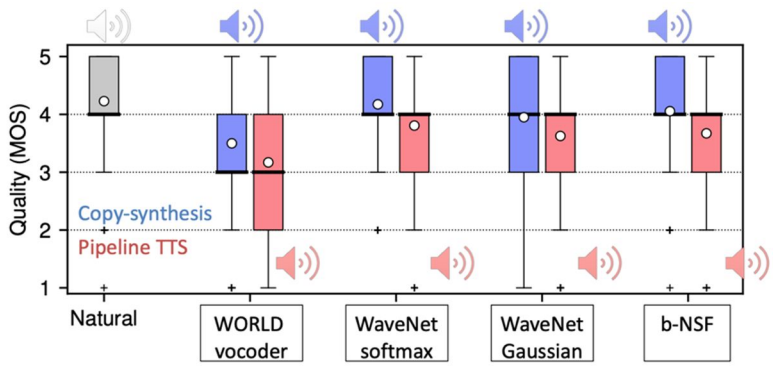

Fig.3 Evaluation of b-NSF, WaveNet, and WORLD vocoder. White dot denotes the mean value of quality scores, and a higher mean score indicates a better performance.

In terms of generation speed, as Fig.4 shows, the b-NSF is much faster than WaveNet, especially in normal mode where the GPU memory is sufficient for the model.

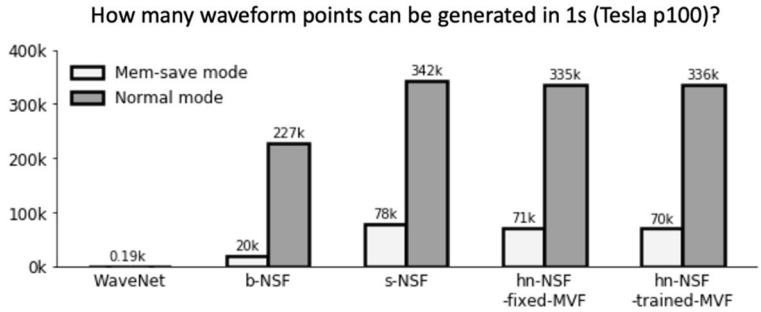

Fig.4 Waveform generation speed

Goal 2: NSF can be improved for waveform generation in different applications (cf. paper published in ssw2019 and Interspeech 2020)

Following the methods for goal 2, we combined the harmonic-plus-noise model [3] and glottal excitation theory [4] and proposed a few variants of the NSF models. One latest NSF model is plotted in Fig.4, where the harmonic branch (blocks 1-5) and the noise branch separately process the cyclic-noise and pure noise-based excitation signals, respectively.

This model was demonstrated to be more suitable than the b-NSF on multi-speaker speech data modeling. Its generation speed is still fast, as Fig.4 shows.

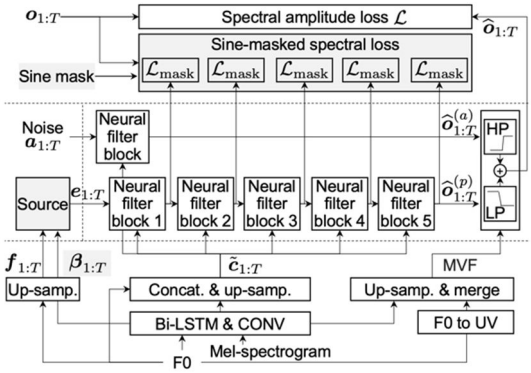

Fig.4 Latest hn-NSF with trainable maximum-voiced-frequency (MVF) and cyclic-noise-based source module

Goal 3: NSF can be applied to music instrumental audios (cf. paper published in ICASSP 2020)

Given the proposed NSF models, we used the latest NSF to model monophonic instruments. Results on a database with more than 10 instruments are illustrated in Fig.5. The NSF (in this case, the hn-NSF) can be trained from scratch (NSF-TS) or fine-tuned after training on speech data (NSF-FT). Its overall performance is better than WaveNet (WN-*) and another neural waveform model called WaveGlow (WG-*). Detailed results each instrument can be found in the paper published at ICASSP 2020.

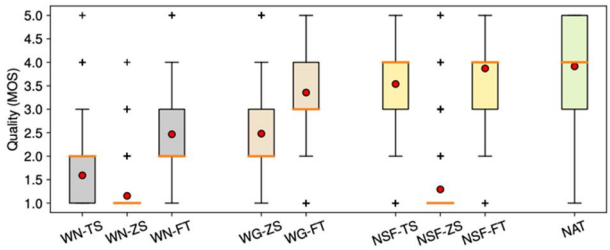

Fig.5 Applying NSF on music instrumental audios

We also tested the NSF models on piano sounds. A key difference between piano sound and human speech is that the former is polyphonic, i.e., a piano sound contains multiple notes at the same time. Interestingly, we found that the NSF model can still produce the piano sounds in a reasonably good quality. This work has been submitted as a paper.

Open source codes and audio samples:

We released the source code of NSF models in both CUDA and Pytorch:
https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts
https://github.com/nii-yamagishilab/project-CURRENNT-scripts

We also publish the audio samples from the NSF models:
https://nii-yamagishilab.github.io/samples-nsf/

Other research outcomes are detailed in the published papers.

[1]G. Fant. Acoustic theory of speech production, Walter de Gruyter, 1970.
[2]A. van den Oord, *et.al.* WaveNet: A generative model for raw audio. arXiv:1609.03499, 2016.
[3]Y. Stylianou. Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. PhD thesis, Ecole Nationale Superieure des Telecommunications, 1996.
[4] T.Drugman, *et.al.* Glottal Source Processing: From Analysis to Applications. *CSL* : 1117–1138. 2014.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Wang Xin、Takaki Shinji、Yamagishi Junichi	４．巻 28
２．論文標題 Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis	５．発行年 2020年
３．雑誌名 IEEE/ACM Transactions on Audio, Speech, and Language Processing	６．最初と最後の頁 402～415
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TASLP.2019.2956145	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計8件（うち招待講演　3件／うち国際学会　7件）

１．発表者名 Wang Xin
２．発表標題 Neural-network-based waveform modeling for text-to-speech synthesis
３．学会等名 Lecture Series on Natural Language Processing（招待講演）
４．発表年 2019年

１．発表者名 Yamagishi Junichi, Wang Xin
２．発表標題 Neural auto-regressive, source-filter and glottal vocoders for speech and music signals
３．学会等名 ISCA 2020 Speech Processing Courses in Crete（招待講演）（国際学会）
４．発表年 2020年

１．発表者名 Wang Xin
２．発表標題 Tutorial on Neural statistical parametric speech synthesis
３．学会等名 The Speaker and Language Recognition Workshop, Odysessy 2020（招待講演）（国際学会）
４．発表年 2020年

１．発表者名
Wang Xin、Yamagishi Junichi

２．発表標題
Neural Harmonic-plus-Noise Waveform Model with Trainable Maximum Voice Frequency for Text-to- Speech Synthesis

３．学会等名
Proceeding of Speech Synthesis Workshop（国際学会）

４．発表年
2019年

１．発表者名
Zhao Yi, Wang Xin, Juvela Lauri, Yamagishi Junichi

２．発表標題
Transferring neural speech waveform synthesizers to musical instrument sounds generation

３．学会等名
IEEE International Conference on Acoustics, Speech and Signal Processing（国際学会）

４．発表年
2019年

１．発表者名
Wang Xin, Yamagishi Junichi

２．発表標題
Using Cyclic Noise as the Source Signal for Neural Source-Filter-Based Speech Waveform Model

３．学会等名
Proc. Interspeech（国際学会）

４．発表年
2020年

１．発表者名
Ai Yang, Li Haoyu, Wang Xin, Yamagishi Junichi, Ling Zhenhua

２．発表標題
Denoising-and-Dereverberation Hierarchical Neural Vocoder for Robust Waveform Generation

３．学会等名
2021 IEEE Spoken Language Technology Workshop (SLT)（国際学会）

４．発表年
2021年

１．発表者名 Reverberation Modeling for Source-Filter-Based Neural Vocoder
２．発表標題 Ai Yang, Wang Xin, Yamagishi Junichi, Ling Zhenhua
３．学会等名 Proc. Interspeech（国際学会）
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Home page of neural source-filter waveform models https://nii-yamagishilab.github.io/samples-nsf/ Neural source-filter waveform model in Pytorch https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts Neural source-filter waveform model in CUDA https://github.com/nii-yamagishilab/project-CURRENNT-public Scripts to use the CUDA implementation https://github.com/nii-yamagishilab/project-CURRENNT-scripts

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
英国	University of Edinburgh			
フィンランド	Aalto University			
中国	中国科学技術大学			