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

2018 年（平成 30 年）12 月 25 日

自然な音声を高速に合成可能な新手法を開発

古典的手法にニューラルネットワークを導入したニューラル・ソースフィルター・モデル

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（NII、所長：喜連川 優、東京都千代田区）のコンテンツ科学研究系の特任研究員 シン ワン、特任助教 高木 信二（たかき・しんじ）、准教授 山岸 順一（やまぎし・じゅんいち）の研究チームは、高品質な音声を高速に合成する手法であるニューラル・ソースフィルター・モデル（以下、NSF 法）を開発しました。この新手法は、1960 年に発表された音声生成モデルを深層学習により発展させた新たな手法で、人間の肉声に近い高品質な音声波形を生成できるだけでなく、ニューラルネットワークの学習を安定して行えるのも特徴です。

【背景】

従来、音声波形を合成する手法として、ボコーダ法と呼ばれる手法が提案され、携帯電話等で広く利用されてきました。しかし、合成された音声の品質は、人間の音声より品質が劣るものでした。2016 年に海外の有力 ICT 企業 (*1) が、深層学習（ディープラーニング）を駆使した音声合成手法 WaveNet 法を提案し、人間の肉声に近い高品質な音声波形が生成できることを示しました。しかし、WaveNet 法は、非常に複雑な構造のニューラルネットワークのため、機械学習に大量の音声データが必要であること、また、正しい予測結果を得るためにはパラメータ調整など色々な試行錯誤を幾度も繰り返さなければならないなどの問題がありました。

【研究概要と成果】

1960 年代に発表されたソースフィルター・ボコーダ法 (*2) は、ボコーダ法の最も有名なモデルとして広く活用されています。NII の研究チームは、このソースフィルター・ボコーダ法にニューラルネットワークを導入することで、人間の肉声に近い高品質な音声波形を生成する新手法を開発しました。NSF 法と名付けたこの手法は、ニューラルネットワークの機械学習のために必要な音声データが 1 時間程度でよいこと、簡易な構造のニューラルネットワークのため、パラメータ調整をしなくても正しい予測結果を得ることができるなどの特徴があります。また、大規模な検証から WaveNet 法から生成された音声と同等に高品質であることが示されました。

National Institute of Informatics

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
総務部企画課 広報チーム
〒101-8430 東京都千代田区一ツ橋 2-1-2
直通：03-4212-2164 FAX：03-4212-2150
E-Mail: media@nii.ac.jp

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

【今後の展望】

NSF 法は、海外の有力 ICT 企業の特許技術とは異なる理論による手法であることから、NSF 法を活用することにより音声合成の新たな技術開発が進むことが期待されます。そこで NSF 法のソースコードを無償で公開し広く利用できるようにしました。

今回の評価に使った機械学習データのサンプル（ソースコード、学習済みのモデル）と、実際に合成された音声データのサンプル（日本語・英語）は、以下のページで公開しています。

＊ソースコード
https://github.com/nii-yamagishilab/project-CURRENNT-public

＊学習済みのモデル（これを実行すると英語の音声を生成することができます。）
https://github.com/nii-yamagishilab/project-CURRENNT-scripts

＊音声サンプル（日本語・英語）
https://nii-yamagishilab.github.io/samples-nsf/index.html

山岸 順一准教授からのコメント：「NSF 法により、音声インターフェースを利用する日本の AI 企業に新たなビジネスチャンスがもたらされるのではと期待しています。この手法をリアルタイム音声合成エンジンとして多様なシステムにおいて利用できるようにしていきます。話者適応技術等も追加する予定です。」

なお、以下のページで、人間の肉声、ソースフィルター・ボコーダ法を用いた音声、WaveNet 法を用いた音声、NSF 法を用いた音声を聞き比べていただくことができます。

https://youtu.be/yr_xMq1gxKY

【研究プロジェクトについて】

本研究成果は、科学技術振興機構 CREST JPMJCR18A6 および日本学術振興会 科研費 16H06302, 16K16096, 17H04687, 18H04120, 18H04112, 18KT0051 の助成を受けたものです。

【論文タイトルと著者】

タイトル：Neural source-filter-based waveform model for statistical parametric speech synthesis

著　　者：シン ワン、高木 信二、山岸 順一

掲 載 誌：International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2019（投稿中）

発 表 日：2018 年 10 月 30 日（Arxiv 掲載：https://arxiv.org/abs/1810.11946）

〈メディアの皆様からのお問い合わせ先〉

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

総務部企画課 広報チーム

TEL:03-4212-2164　E-mail：media@nii.ac.jp

(*1) Google 傘下である DeepMind 社。囲碁人工知能「AlphaGo」を開発したことでも知られる。
(*2) 1960 年に Gunnar Fant 博士が発表した音声生成モデル。音声生成のプロセスを、人間の声門などの音源（ソース）と 声道などの線形音響フィルタで近似している。