TSUBAME 共同利用　平成 30 年度　学術利用　成果報告書

利用課題名： ニューラルネットワークに基づく音声合成
英文： Neural network-based speech synthesis

利用課題責任者： 山岸　順一
First name Surname: Junichi Yamagishi

所属： 国立情報学研究所
Affiliation: National Institute of Informatics
URL: https://www.nii.ac.jp/

邦文抄録 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・モデルを開発しました。ニューラルネットワークの機械学習も容易に安定して行える新手法となります。この新たな方式は、単純な構造によるニューラルネットワークを利用しているのが特徴です。単純構造でありながら、現在主流になりつつある複雑なニューラルネットワーク構造の「Wavenet 法」と同等の、非常に肉声感の高い音声波形を生成できることを確認しました。

英文抄録 We have developed the method of neural source-filter (NSF) models for high-speed, high-quality speech 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.

Keywords: speech information processing, speech synthesis, deep learning, Wavenet, source filter

背景と目的

現在の音声合成システムは、2017 年に海外の有力 ITC 企業が発表した、人間の音声とほぼ区別がつかないほどの高音質な音声波形生成法が使用されつつあります。「Wavenet」としても知られるこの手法は 2016 年に発表された当初、数秒の音声の合成に GPU サーバ上で数時間の計算が必要など、計算量・計算時間に関する大きな問題がありました。ところが、音質はほぼそのままに約 1,000 倍速く動作するように改良された方法（以下高速 Wavenet 法）が 2017 年に発表され、現在急速に市販製品にも活用されるようになっています。

しかし、高速 Wavenet 法は現在もっとも良い音が生成可能とされているものの、非常に複雑な構造によるニューラルネットワークを複数の基準により同時に学習させるため、色々な試行錯誤や調整を何度も繰り返すことが必要でした。

現在活用されつつある Wavenet 法と同等の品質による音声波形を短時間で生成可能なニューラルネットワークを安定的に学習できる新手法の NSF 法を開発しました。これは高速 Wavenet 法とは全く異なる理論で実現しており、古典的なソースフィルター・ボコーダー法にニューラルネットワークを導入した新手法となっています。また、高速 Wavenet 法と同程度の音声品質でありながら、1秒間に 14 秒分の音声信号を合成できるほど高速に動作します。ニューラルネットワークの学習も容易に安定して行うことができます

概要

今回開発した NSF 法は、条件付けモジュール・音源生成モジュール・ニューラルフィルターモジュールの 3 モジュールで構成されています（図 1）。音源生成モジュールでは、声の高さに相当する基本周波数およびその調波構造からなる音源信号が生成されます。それに続くニューラルフィルター・モジュールでは、音源モジュールで生成された音源信号を受け取り音声波形へと変換されます。

この構造は、人間の発声構造を模している古典的なソースフィルター・ボコーダと同じ構成ですが、フィルター・モジュールにニューラルネットワークを組み込んだ新手法となっています。このニューラルネットワークの機械学習では、時間軸に沿って予測した出力波形を音

声の主要な特徴を示す周波数へ変換し、合成される音声の周波数スペクトルと位相の誤差を学習の際に直接考慮させる処理となっています。

これに対して、高速 Wavenet 法は、教師ネットワークと生徒ネットワークによる複雑な再帰計算となっています（図 2）。そのため、今回開発した新手法の NFS 法はその学習を容易に行えます。

結果および考察

開発した新手法である NSF 法で合成した音声の品質は、現在もっとも良い音を生成可能とされる Wavenet 法と比べても全く遜色がない結果を示しました。今回の評価では、この NSF 法で実装したプログラムを、女性が日本語で発話した 15 時間分の音声を使って機械学習させました。そして、学習後のプログラムが合成した音声による 480 発話を実際に 245 人に聞いてもらい、合成音声の品質を5段階の評点数値（5：とても良い～1：とても悪い）で示す「平均オピニオン評点（MOS）法」で評価しました。

「テキスト情報から予測された周波数等の特徴」を入力して合成した音声の評価は図 3 赤となり、「人間による実際の音声波形から直に抽出した周波数等の特徴」を入力として合成した音声の評価は図 3 青となります。この実験結果から、新開発の NSF 法で合成した音声の品質は入力の方法によらずに、現在もっとも音質が良いとされるオリジナル Wavenet 法および高速 Wavenet 法での合成音質と比べても全く遜色がないことが示されました。

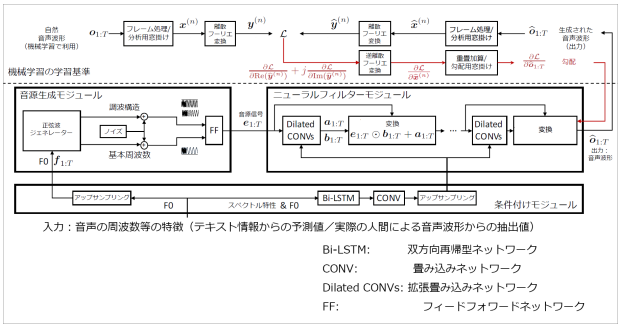

図 1 図 1:今回新開発した単純な構造の「ニューラル・ソースフィルター・モデル（NSF 法）」

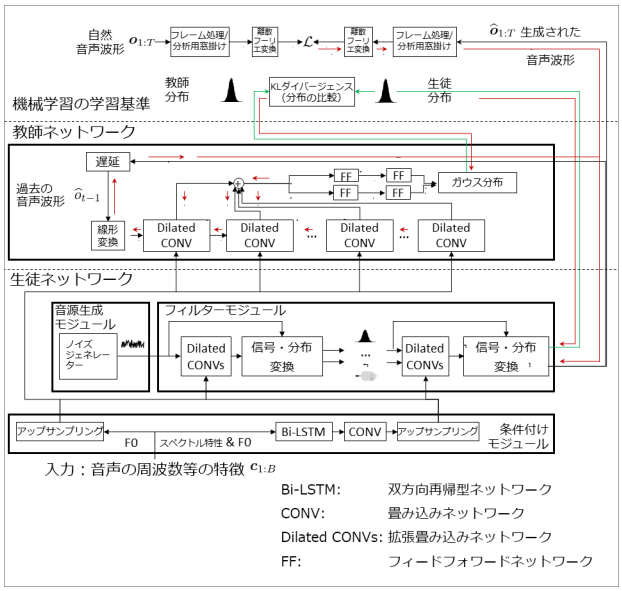

図 2 図 2: 複雑な構造ながら、高速で良い音が生成可能とされている「高速 Wavenet 法」

まとめ、今後の課題

この NSF の開発のほか、日本語 end-to-end TTS システムを構築する作業等も行った。今後は、ニューラルボコーダの改良のほか、バイリンガル音声合成等の開発も行う予定である。