平成30年度 情報工学コース卒業研究報告要旨

戸田 研究室	氏　　名	安　原　和　輝
卒業研究題目	End-to-End 型テキスト音声合成における WaveNet ボコーダの学習に関する調査	

テキスト音声合成 (Text-to-Speech: TTS) とは，入力されたテキストを音声へと変換する技術である．この技術は，カーナビにおける音声案内，スマートスピーカーの対話インターフェース，さらには文章が読めない幼児等に向けたテキストの読み上げといった幅広い分野で利用されており，我々の生活に大きく貢献をしている．

TTS システムは，テキストから音響特徴量を生成する特徴量生成部と，音響特徴量から波形を生成する波形生成部で構成される．従来の TTS システムの特徴量生成部は，入力テキストから言語特徴量を抽出する構文解析器や，音素の継続長モデル，そして音響特徴量生成モデルなどの様々なモジュールで構成されており，構築のコストが極めて高い．また，波形生成部では，音声の生成過程を単純な数理モデルで近似した信号処理ベースのボコーダが用いられるため，合成音声の品質が劣化してしまう．

これらの問題を解決する手法として，ニューラルネットワークのみで処理が完結する End-to-End (E2E) 方式が注目を集めている．中でも，E2E-TTS システムの一つである Tacotron2 は，特徴量生成部に入力テキストの文字系列から音響特徴量系列を直接推定する Sequence-to-Sequence 型のネットワークを利用しており，従来は必須であった言語特徴量の抽出や明示的な継続長のモデル化を必要としない．また，波形生成部に音響特徴量から波形を直接推定する WaveNet ボコーダを利用しており，信号処理ベースのボコーダの利用による音質の劣化を回避している．一方で，これら 2 つのネットワークの接続法や統合法については，未だ不明な点が多い．

本研究では，E2E-TTS システムの品質のさらなる改善に向けて，Tacotron2 における WaveNet ボコーダの学習に関する調査を行う．第一に，学習時と合成時における音響特徴量のミスマッチを解消するため，WaveNet ボコーダに対して，Tacotron2 の推定特徴量を用いたファインチューニングを適用する (図 1 左側)．第二に，ミスマッチに対する頑健性を高めるため，音響特徴量系列に対して時間方向の畳み込み層を適用する (図 1 右側)．主観評価に基づく実験的評価により，各手法が合成音声の自然性に与える影響を明らかにする．実験結果 (図 2) から，1) ファインチューニングの適用のみでは改善が見られないこと，2) 畳み込み層の適用は有効であること，3) ファインチューニングと畳み込み層の適用を組み合わせることで，さらに自然性を改善できることを示す．

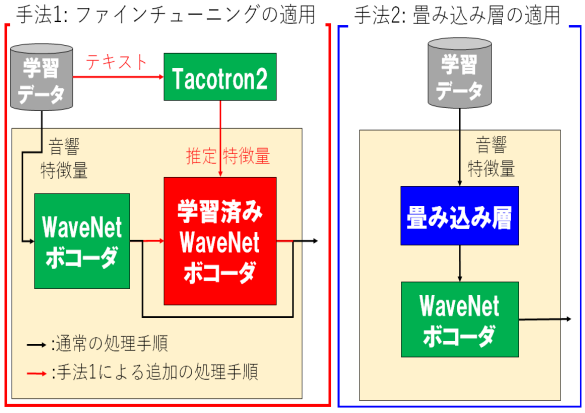

図 1: 各手法による学習過程

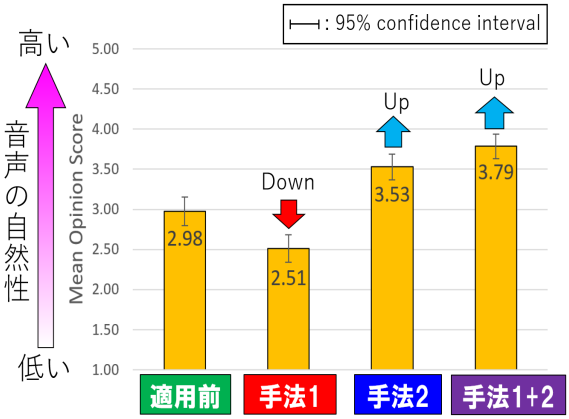

図 2: 主観評価実験の結果