多様な雑音に対して耐性のある声質変換システム

佐藤邦彦[†1]　暦本純一[†1†2]

概要：声質変換とは，ある人物の話した音声を，まるで別の人が話したかのように変換することである．声質変換はさまざまなアプリケーションへの応用可能性を秘めており，音声チャットや発話障害者支援，テーマパークの着ぐるみ，カラオケ，VR など多岐にわたる．しかし，既存の声質変換手法の変換速度はリアルタイム性に欠ける．また，入力音源に雑音が含まれていると期待通りに変換できないという問題がある．本研究では，Deep Neural Network (DNN) を駆使することで，雑音耐性があり，リアルタイム性のより高い声質変換システムを提案する．評価実験の結果，声質変換の DNN を並列処理可能なモデルに改良し，精度を落とさずに学習速度と変換速度を向上させた．また，多様な種類の雑音が含まれた話者音声に対して，雑音を除去せずに声質変換した場合よりも高い精度で声質変換できることがわかった．

A Noise-Tolerant Voice Conversion System for Imitating Anyone's Voice

KUNIHIKO SATO[†1]　JUN REKIMOTO[†1†2]

Abstract: Voice Conversion (VC) is to modify one speaker's speech to make it sound like another specific speaker's. VC can be widely applied to many fields including voice chat, theme parks, karaoke, virtual reality and speech-impaired patient assistance. However, state-of-the-art VC methods have the lack of real-time performance and suffer from environmental noise. We propose a novel VC system using Deep Neural Network, which allows for parallel computation and filters background noise automatically. Our experiments are performed to evaluate the proposed VC system under noisy environments, and validate the feasibility.

1. はじめに

声質変換とは，ある人物の話した音声を，まるで別の人が話したかのように変換する技術である（図 1）．

この技術によってあらゆる人の声真似が可能になったとき，どのようなことに応用できるだろうか．

例えば，Snapchat や Facebook Messenger などの音声コミュニケーションチャットや Instagram などの Social Networking Service へ応用可能である．上記のアプリケーションでは，動画や写真に映っている顔に対してスタンプを重畳したり顔を変形させたりするイメージフィルター機能が人気である．さらに，イメージフィルター機能に加え，音声インタラクションが次の新しい機能として注目を集めている．事実，Snapchat はユーザーの音声のピッチを変えたりロボットが喋ったようにしたりするボイスチェンジャー機能を 2016 年に新しい機能として追加した．声質変換を音声チャットに応用すれば，ユーザーは自身の声を有名な俳優やアニメキャラクターの声色に真似て楽しむことができる．

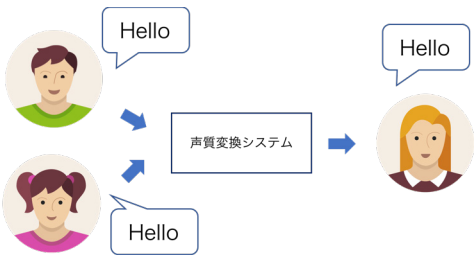

図 1：声質変換のイメージ図．本研究で提案する声質変換システムは，図に示すように，誰が音声を入力しても特定の人物の声質に変換できる．

他の応用例として，喉頭がんなどによって声帯を摘出した人の声質復元がある．声帯摘出者の一部はユアトーン[30]などのバイブレーターを喉頭部にあて，人工的に振動を作り出し音声を発する．しかし，ユアトーンによって発せられた音声は，機械音のような音でありその人自身の声質ではない．声質変換を用いることで，ユーザーは，発せられた音声をその人自身の声質に変えることができる．

さらに，テーマパークの着ぐるみに声質変換機能を搭載すれば，着ぐるみの中の演者がそのキャラクターの声真似をすることができる．それによって，例えばディズニーランドなどのテーマパークで，来場者と着ぐるみが会話することが可能になる．また，カラオケに声質変換機能を搭載すれば，ユーザーはさまざまな歌手のモノマネができる．

このように，声質変換は日常生活の中でさまざまな状況において発展性を有する．

既存の声質変換手法[23]では，Deep Neural Network (DNN) を用いることで，だれが音声を入力しても特定人物

†1 東京大学大学院情報学環
†2 ソニーコンピュータサイエンス研究所

の声質に変換できる．しかし，既存の声質変換手法は，実用面において以下の 2 つの課題がある．

1）既存の声質変換手法は，声質変換のためにDNNを用いているが，採用しているDNNのモデル構造がLSTMである．LSTMは各ユニットの計算が時間的に前のユニットの計算に依存するため，並列計算に適していない．つまり，既存手法の変換処理はリアルタイム性に欠ける．この制約は，リアルタイム性が要求されるようなアプリケーションに声質変換を利用した場合，大きな問題となると考えられる．

2)声質変換は,入力音源の波形データを変形させるため,入力音源に入力話者の発話以外の雑音が含まれていると,期待通りに声質変換ができない．既存の声質変換システムの多くは，話者の音声のみが含まれる音源を入力音源として想定している．しかし，声質変換システムを日常の環境下で使用した場合，車の走行音や音楽，雑踏の騒音など音声以外の雑音が入力音源に含まれる．これらの雑音は，声質変換が期待通り行われず，生成された音声が元の言語情報を保持できない原因となる．そのため，音声チャットやユアトーンのように日常のあらゆる場面で声質変換を行う場合や，テーマパークやカラオケのように周りに騒音がある場合は,音声と雑音を分離するための機能が必要になる．

本研究では，背景雑音を自動で除去し，リアルタイム性の高い声質変換システムを提案する．提案システムの声質変換手法は,既存の声質変換手法と比べ,変換処理が速い．また，本システムは，リビングや公園，カフェなどの多様な環境における雑音に対して，雑音を除去せずに声質変換した場合と比べ，精度の高い声質変換ができる．本システムは，音声を入力する人が誰であっても，特定の人物の声質に変換可能である．これを多対一への声質変換という．本システムは，声質変換モジュールと音源分離モジュールから構成される．それぞれのモジュールは，DNN によって実現されている．

我々は，提案したシステムに対して，3 つの評価実験を行なった.実験の結果,本研究の貢献は以下の通りである．

1. 声質変換手法において，並列処理可能な DNN 構造を提案した．その結果，DNN の学習速度および音声の変換処理の速度が早くなった.また,処理速度の向上を達成しながら，変換の精度は既存の手法と同程度を実現した．
2. DNN 以外の手法も含む既存の声質変換の研究において，雑音耐性を備えた多対一の声質変換を対象としたものはなかった.本研究では,多対一の声質変換に音源分離機能を追加し，雑音のある音声を入力した場合の変換精度を主観評価実験によって確認した.

また，我々は，DNN の技術的改善以外で，音声の聞こえ方に基づいた精度向上を試みた．声質変換モジュールによって変換された音声に対して残響を加えることで聞こえ方が変わるか，主観評価実験を行なった．さらに，変換した音声に対して，入力音声の背景雑音を再合成することで音声の聞こえ方が変わるかどうかを検証した．

第 7 章では，本システムの技術が他の音声タスクに対して大きな発展性を有していることを述べる．

2. 関連研究

2.1 声質変換

既存の主要な声質変換手法として，統計的音声合成がある．当手法では，まず，変換される音声を入力する話者と声質の変換先となる出力話者が，同一内容を発話した音声データ（パラレルデータ）を用意する．パラレルデータとは，入力話者が「こんにちは」と言った音声データと，変換先の人物が「こんにちは」と言った音声データのように，同じ言葉をなるべく同じ早さで話した対の音声データである．当手法では，用意したパラレルデータを統計的手法に基づいてコンピュータに学習させ，入力音声の特徴量と出力音声の特徴量をマッピングするための変換規則をモデル化する.この特徴量のマッピングを最適化するために，さまざまな手法が提案されてきた．Stylianou ら[22]や Toda ら[27]は，入力音声の特徴量が与えられた際の，出力話者の特徴量の出現確率を混合ガウス分布によってモデル化した．Wu ら[28]は，非負値行列因子分解を用いた特徴量のマッピングを提案した．Nakashika ら[17]は DNN を用いてマッピングの最適化を試みた．これらの先行研究における提案は良好な結果をもたらしているが，パラレルデータを用いた手法である．パラレルデータは通常，容易に入手できるものではない．パラレルデータを作成するためには，入力話者と出力話者が同一内容を発話した音声を収録する必要がある．ひとりあたり，数十分から数時間分の音声データを収録する．パラレルデータの作成には，多大な制約と労力が必要である．

そのため，近年では，パラレルデータを用いない手法が提案されてきた[7][25][10][2][26]. ここで意味するパラレルデータを用いない手法とは，入力話者と出力話者の発話内容が同一でない音声データ，あるいは，出力話者のみの音声データで声質変換を実現することである．これらの手法には DNN が使われておらず，パラレルデータを利用した手法よりも性能が劣っていた．パラレルデータを用いない手法の困難な点は，入力音声の発話内容を保持しながら声質変換を行うことである．

パラレルデータを用いない手法のうち，パラレルデータを利用した手法と同程度以上の性能を出したのが，Sun ら[23]の手法である．Sun らの手法の特徴は，音素認識モデルと声質変換モデルの 2 つの DNN を組み合わせたことである．この 2 つの DNN によって，パラレルデータを用いない手法の困難な点を克服した．Sun らの手法は，だれが音声入力した場合でも,特定の人物の声質に変換可能である．

2.2 音源分離

本研究における音源分離とは，入力音源から音声とそれ以外の音を分離することを指す．

入力波形から特定の音源を分離するための古典的な手法としてウィーナフィルタリング[21]がある．ウィーナフィルタリングの手法は，パラメータをヒューリスティックに決定するため，さまざまな音源に対してパラメータを最適化することはできない．

近年では，深層学習を用いることで音源分離を試みる研究が出てきた[15][14][4][5][11][29]．これらの手法では，雑音を含む話者音声の周波数領域の特徴量と，話者音声のみの周波数領域の特徴量を DNN によってマッピングする．Rethage ら[20]は，周波数領域の特徴量を用いずに音声波形をそのまま入出力する音源分離 DNN を提案し，良好な結果を示した．

2.3 雑音環境下における声質変換

雑音環境化において声質変換が期待通り行われないことを確かめた研究として Monisankha ら[19]の研究がある．Monisankha らは，5 つの異なった声質変換手法に対して実験を行い，それらの性能は雑音環境下では著しく低下することを報告した．

高島ら[24]や相原ら[1]は，雑音環境に耐性のある声質変換手法を提案し，その有用性を示した．しかし，それらの手法は，声質変換にパラレルデータを利用している．

真坂ら[16]は，相原らの手法に加え，発話時の口唇画像情報を用いたマルチモーダルな声質変換手法によって精度を向上させた．真坂らの手法はパラレルデータが必要なことに加え，画像取得のために，発話中は口唇部分をビデオカメラで撮影する必要がある．

このように，雑音耐性があり，かつ，パラレルデータを用いない声質変換手法については先行事例がない．

3. 提案手法

3.1 システム全体の概要

提案システムは，図 2 に示すように，2 つのソフトウェアモジュールを有し，アレイマイクのような特殊なハードウェアを必要としない．モジュールのひとつは，あらゆる人物が音声を入力した場合でも任意の人物の声質に変換することができる声質変換モジュールである．もうひとつは，入力話者の音声と雑音を分離するための音源分離モジュールである．我々は，それぞれのモジュールを，DNN を用いて実装した．

3.2 声質変換モジュール

3.2.1 概要

我々が提案する声質変換モジュールは，Sun ら[23]の手法を元にしている．以後，Sun らの手法をベースラインと呼ぶ．我々は，ベースラインを改良し，声質変換の処理を高速化した．以下，ベースラインについて説明をした後，提案した手法の説明をする．

3.2.2 ベースライン：Deep bidirectional long short-term memory

概要

ベースラインは，パラレルデータを用いない手法であり，だれが音声を入力しても，任意の人物の声質に変換できる．これを多対一への声質変換という．多対一への声質変換を実現するために，ベースラインは 2 つの DNN を提案した（図 3）．ひとつは音素認識モデルで，もうひとつは声質変換モデルである．Sun らのアイディアは，音素認識モデルから得られた音素分布 (phoneticposterior-grams，PPG)を利用して話者間の声質を橋渡しすることである．PPG は，音声の内容を保持しながら，話者固有の特徴を取り除いたベクトル空間を表現することができると仮定されている[9][12]．ベースラインのアプローチは，まず，入力音声の PPG を得る．次に，deep bidirectional long short-term memory (DBLTSM)に基づく声質変換モデルを使用して，PPG と変換先の音声の音響特性の関係をモデル化する．ベースラインは，任意の入力音声を変換するために，同一の音素認識モデルから PPG を取得し，それを変換先の人物の音響特性を学習した声質変換モデルに入力する．声質変換モデルは，変換先の人物の音響特性を出力する．

ベースラインは，訓練ステージ 1，訓練ステージ 2，変換ステージの 3 段階に実装が分かれている．訓練ステージ 1 では，複数話者の英語コーパスを利用して音素認識モデルを学習する．音素認識モデルの役割は，入力音声から音素

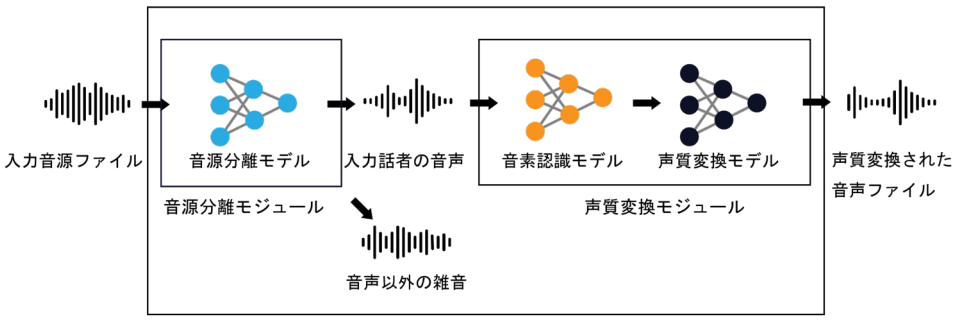

図 2：提案システムの概要図

を予測することである．訓練ステージ 2 では，PPG と変換先の人物のメルケプストラム係数（MCEP）の関係をモデル化する．MCEP は声質の特徴を表すスペクトル包絡のひとつである．変換ステージでは，音素認識モデルから得られた入力話者の PPG と学習済みの DBLSTM モデルを利用し，声質変換を行う．以下，訓練ステージ 1，2 と変換ステージの詳細を説明する．

訓練ステージ 1 と 2

訓練ステージ 1 では，PPG を予測するために，発話の音素分布ラベルが付与されている複数話者の英語音声コーパスを用いた学習を行う．モデルへの入力は，入力音声のフレーム目のメル周波数ケプストラム係数（MFCC）のベクトルであり，ここではX_tと表す．MFCC は，人間の知覚に基づいてフィルタリングされた，声質の特徴を表す周波数領域の特徴量である．出力は，PPG のベクトルであり，ここでは$P_t = (p(s|X_t)|s = 1, 2, \dots, C)$と定義する．$p(s|X_t)$は各音素クラスの$s$の事後確率である．

学習ステージ 2 では，声質変換モデルを学習して，PPG と MCEP とのマッピング関係を得る（図 5）．声質変換モデルへの入力は，学習済みの音素認識モデルによって予測された PPG $[P_1, \dots, P_t, \dots, P_N]$ である（t は入力音声の t フレーム目を表す）．出力は，声質変換モデルによって予測された MCEP のベクトルであり，$[Y_1, \dots, Y_t, \dots, Y_N]$と表す．ここで，正解データの MCEP ベクトルを $[Y'_1, \dots, Y'_t, \dots, Y'_N]$とすると，学習ステージ 2 の損失関数は以下のように定義される．

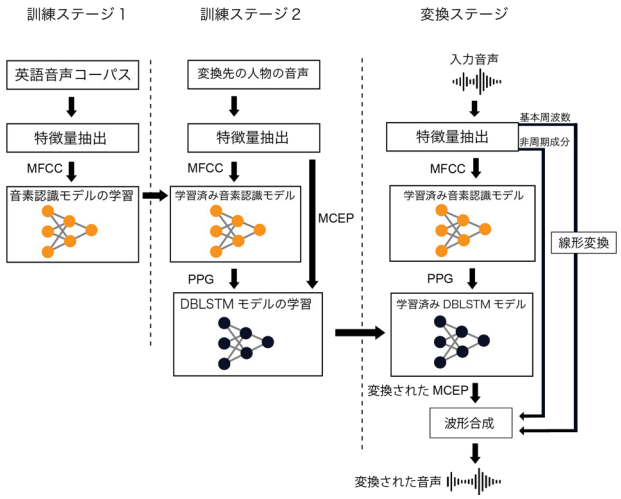

図 3：ベースラインにおける訓練ステージ 1，訓練ステージ 2，変換ステージの概要図．Sun ら[10]の論文の図から再構成．訓練ステージ 1 では，複数の英語話者で構成される音声コーパスを使用して，音素認識モデルを学習させる．訓練ステージ 2 では，訓練ステージ 1 で学習した音素認識モデルを利用して声質変換モデルを学習させる．変換ステージでは，学習済みの声質変換モデルを利用して，入力音声の PPG から対象話者の MCEP を出力する．

$$min \sum_{t=1}^{N} \|Y'_t - Y_t\|^2 \quad (1)$$

声質変換モデルは誤差逆伝搬によって，損失関数を最小化するように学習する．声質変換モデルの学習には，音素ラベルなどの言語情報なしで，変換先の人物の音声ファイルさえあればよい．

変換ステージ

2 つのモデルの学習後，変換ステージは次のように動作する．まず，入力音声から基本周波数，非周期成分，MFCC を抽出する．非周期成分は，息遣いや声のかすれを表す．基本周波数は声の高さを表す．次に，学習済み音素認識モデルに入力音声の MFCC を入力し PPG を得る．その後，学習済み声質変換モデルに PPG を入力し MCEP を得る．最後に，生成された MCEP，変換先の人物の声の高さに線形変換された基本周波数，入力音声の非周期成分を用いて，再び波形データを合成し出力する．

制約

ベースラインの手法は，多対一への声質変換を実現したが，処理速度に制約がある．ベースラインは，声質変換モデルの実装に，DBLTSM を採用している．DBLSTM は，時間方向の前方と後方の双方に学習を行う LSTM である．LSTM は，再帰結合を有するため，GPU による並列処理が難しい（図 4）．そのため，並列処理可能な Convolutional Neural Network (CNN) などのモデル構造と比べ，処理速度が遅いことがわかっている[3].

3.2.3 提案手法：Dilated Convolutional Neural Network

概要

ベースラインの制約を解消するために，以下の取り組みをした．声質変換モデルを DBLSTM から，並列処理可能な Dilated CNN に変更した（図 5）．Dilated CNN は CNN のフィルターサイズを層ごとに拡張する DNN モデルである．Dilated CNN は画像処理のために提案された手法であり，そのまま声質変換に適用するのは難しい．そこで，テキストから音声を生成する「テキスト読み上げ」のタスクに対して，Dilated CNN を適用した DNN モデルである WaveNet[18]の手法を参考にした．しかし，WaveNet は

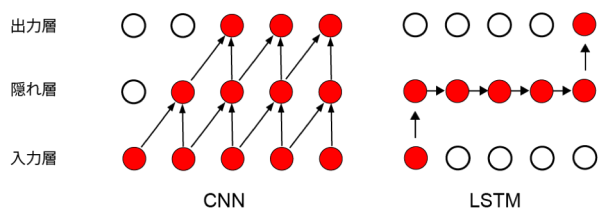

図 4：典型的な LSTM および Dilated CNN アーキテクチャの計算構造の図解．赤色は畳み込みまたは行列乗算を意味する． 各タイムステップでの LSTM の計算は、以前のタイムステップからの結果に依存している． そのため、LSTM が並列処理を実行することは困難である．

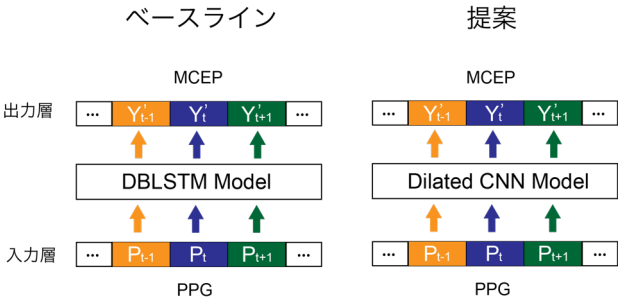

図 5：左：DBLSTM を用いた声質変換モデル．右：提案手法の Dilated CNN を用いた声質変換モデル．

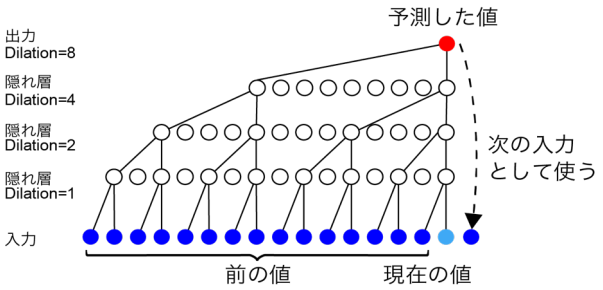

図 6：WaveNet で用いられている Dilated CNN の図解．WaveNet は，前の値と現在の値を入力として予測を行う．予測した値は次の入力値として使う．

Dilated CNN を基本とした DNN モデルであるが，音声波形を逐次的に生成する．そこで提案手法では，WaveNet を声質変換に適用し，変換処理を高速化した．以下，WaveNet の説明をしたあと，提案手法について述べる．

WaveNet

WaveNet は，予測した値をモデルの次の入力値として利用する自己回帰型モデルである（図 6）．以前に生成した値をモデルに逐次的に与えることで，音声波形が持つ時間的な連続性を担保している．WaveNet は画像処理に対して提案された Dilated CNN を音声生成に適用するために，ゲートユニットやスキップコネクション，残差ブロックなどの工夫を提案した．

提案手法

WaveNet は予測した値を次の入力値として利用するため，予測は逐次的に行われる．そのため，WaveNet は並列処理が困難であり，リアルタイム処理に適していない．そこで，提案する声質変換モデルを，入力音声の PPG からすべての出力値（MCEP）を予測するように設計し直した．提案するモデルは，すべての入力を同時に処理することができ，学習時と予測時の両方において並列計算が可能である．

また，我々は当初，WaveNet と同様に，時間的に順方向の畳み込み演算だけを行うように声質変換モデルを設計した．しかし，順方向のみで学習したモデルから生成された音声は雑音を多く含んでいた．そこで，モデルを再設計し，順方向と逆方向の畳み込みを同時に行うようにした（図 7）．再設計されたモデルにおいて，順方向と逆方向の畳み込み演算は同時に行われ，それぞれの演算結果は合算される．各畳み込み演算は並列に処理されるため，ニューラルネットワーク全体の処理時間は順方向のみの畳み込み演算のときと同じである．

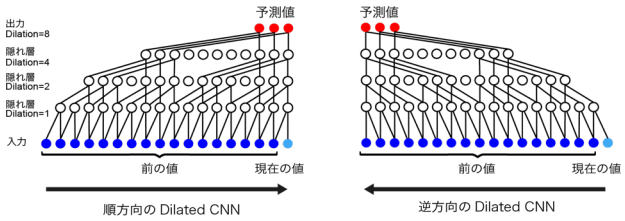

図 7：提案手法の Dilated CNN の図解．WaveNet との違いは，時間的に順方向の畳み込みだけでなく，逆方向の畳み込みを行うことである．

実装

実装した声質変換モデルの詳細を図 8 に示す．各層のフィルターサイズは，{1, 2, … 256, 512}のように，2 のべき乗ごとに拡大させた．この層群を 2 つ重ねた．各層の構造に，残差ブロックとスキップコネクションを採用することで，勾配消失を防ぎ収束がより早くなるように設計した．声質変換モデルの学習には，アメリカ英語の発話データである CMU ARCTIC コーパス[13]を使用した．声質変換先のターゲットとして，当コーパスから男性 1 名，女性 1 名を選んだ．変換対象の話者数は 2 名だが，これはベースラインの論文[23]と同様である．訓練データ量は，男女それぞれ 200 文である．学習におけるバッチサイズは 16，エポック数は 16000 とした．使用した GPU は Nvidia Titan X Pascal，CPU は Intel i7-5930K である．

音素認識モデルは LSTM で実装した．データセットは複数の英語話者が含まれる TIMIT コーパス[8]を利用した．TIMIT コーパスには，発話の音素分布ラベルが付与されている複数の英語話者の音声データがあり，音素認識モデルの学習に適している．

3.3 Deep Neural Network を応用した音源分離

3.3.1 概要

音源分離モジュールは，雑音が含まれる音声データを，音声と雑音に分離する．関連研究の調査から，Rethage ら[20]が提案する DNN は，ホワイトノイズだけでなく多様な環境の雑音に対して，音声と雑音を分離できることがわかった．そこで，Rethage らの提案手法を追実装した．Rethage らの手法は，声質変換モデルと同様に，Dilated CNN を用いることで任意の時系列方向に対する柔軟な畳み込みを可能にしている．一方，声質変換モデルとの大きな違いは，音声データの入出力に，MCEP のような周波数領域の特徴量を使わずに，音声波形をそのまま用いている点である．

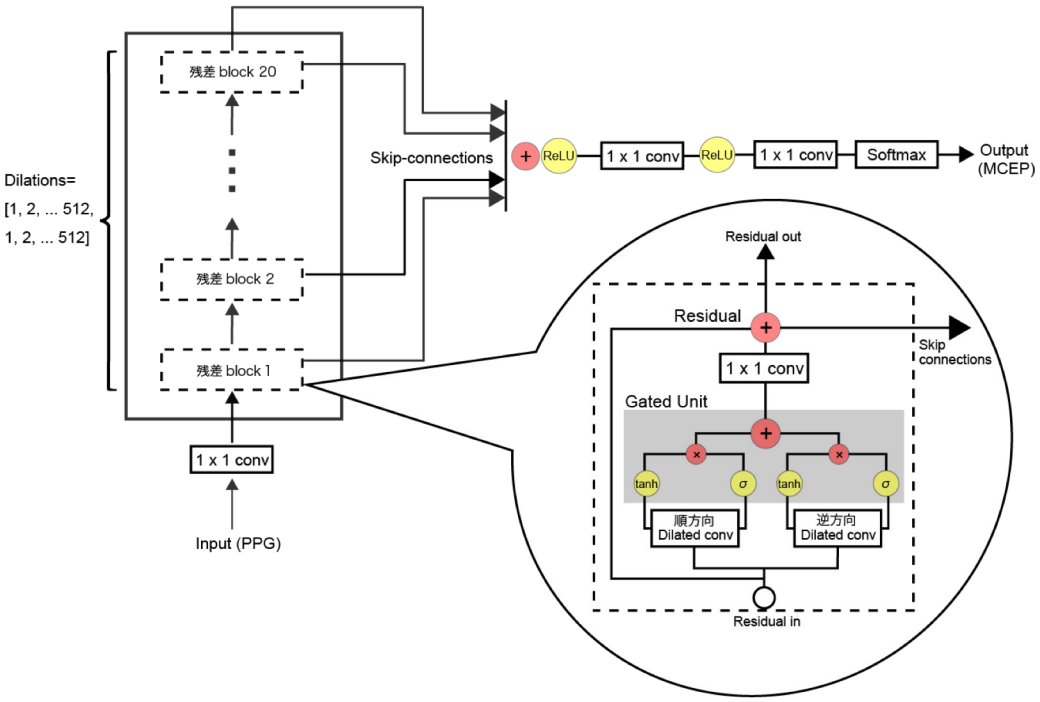

図 8 : 提案した声質変換モデルの全体図．音素認識モデルから生成された PPG を入力とし，MCEP を出力する．WaveNet で提案された残差ブロック，ゲートユニット，スキップコネクションを採用している．一方，WaveNet との違いは，残差ブロック内の畳み込み演算を順方向と逆方向の 2 つ行うことである．それぞれの畳み込みは並列に実行可能である．

3.3.2 学習

音源分離モデルの学習のために，話者音声のみを含む音源に雑音を重畳したデータセットを作成した．雑音のデータは Diverse Environments Multichannel Acoustic Noise DatabaseValeitini (DEMAND)[6]にある音源を利用した．DEMAND から 8 種類の異なった環境下（キッチン，ミーティングルーム，車内，地下鉄，交差点，食堂，レストラン，駅）の音源を利用し，CMU ARCTIC コーパスの音声に重畳した．重畳した雑音の大きさは，各環境につき，SN 比（Signal Noise 比）で 0, 5, 10, 15dB の 4 種類である．SN 比は dB の値が小さいほど，雑音の音が大きいことを表す．話者音声のみの音源を正解データ，雑音が含まれる音源を入力データとして学習を行なった．

学習におけるバッチサイズは 16 であった．テスト用データセットの誤差が小さくならなかった段階で学習を止めた結果, エポック数は 33 であった. 使用した GPU は Nvidia Titan X Pascal，CPU は Intel i7-5930K である．

4. 実験 1：ベースライン vs 提案手法

本実験では，ベースラインと我々が提案した声質変換モジュールの処理速度と変換の精度を比較した．本実験は，声質変換モジュールの手法を対象としているため，音源分離モジュールは使用しない．

処理速度の比較は 2 種類である．ひとつは学習時間の速さであり，もうひとつは学習済みのモデルに音声を入力し出力までにかかる時間である．

精度については，客観的な評価と主観的な評価を行なった．

4.1 セットアップ

実験にあたり, 我々はベースラインの手法を追実装した．ベースラインの論文に基づき，モデル構造は DBLSTM を採用し，隠れ層のユニット数は，{64, 64, 64, 64}である．各隠れ層は，前方向の LSTM と後方向の LSTM の層を含む．学習データは，公正性を保つために，提案手法の実装に用いた男女（男性 1 名，女性 1 名）それぞれ 200 文ずつの音声ファイルを用いた．音素認識モデルは提案システムと同じものを用いた．

4.2 処理速度

4.2.1 学習速度

学習時間の早さは，10 エポックの合計時間を集計し，1 エポックあたりの平均を求めた．エポックは学習データをすべて入力した回数を表す単位である．

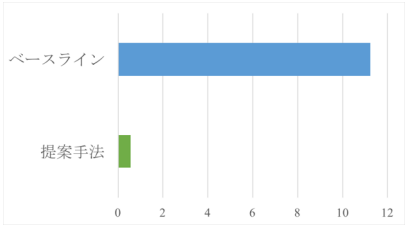

図 9 : 学習時間の結果．1 エポックあたりの平均時間は，ベースラインが 11.24 秒，提案手法が 0.55 秒であった．提案手法はベースラインより 20.44 倍早かった．

結果は，図 9 に示すように，ベースラインが 11.24 秒/エポックであり，提案手法が 0.55 秒/エポックであった．モデル構造から推測した通り，GPU を用いて学習を行なった場合，提案手法のほうがベースラインより 20.44 倍学習速度が早かった．

4.2.2 変換時間

学習済みのモデルに音声を入力した際に，出力までにかかる時間を測定した．音声ファイルの長さは，10 秒と 1 分のものを用意した．それぞれの長さに対して 10 ファイルずつ入力し，1 ファイルあたりの平均時間を求めた．

結果は，図 10 に示すように，10 秒の音声ファイルに対してベースラインは 3.201 秒，提案手法は 1.208 秒であった．1 分の音声ファイルに対しては，ベースラインは 19.113 秒，提案手法は 1.218 秒であった．学習速度同様に，GPU を用いた場合，提案手法のほうが早かった．

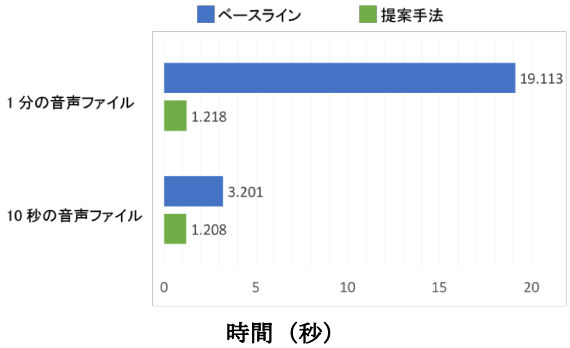

図 10：変換速度の結果．

4.3 精度

4.3.1 客観評価実験

先行研究において，声質変換された音声がどのくらい本物に近いかを定量的に求めるための手法として，メルケプストラム歪み（Mel-cepstral distortion，MCD）を用いることが多い[23]．本実験においても MCD を計算し，客観的な評価を行なった．

MCD は，変換された音声の本物の音声の MCEP のユークリッド距離を計算することで求めることができる．

$$\mathrm{MCD[dB]} = \frac{10}{ln10}\sqrt{2\sum_{d=1}^{N}(C_d - C_d^{converted})^2} \qquad (2)$$

C_d と $C_d^{converted}$ は入力音声と変換された音声の d 次元目の MCEP を表す．本評価では N=40 で計算した．

我々は，テストデータとして，CMU コーパスから学習に使用してない男女それぞれ 1 名ずつ，計 2 名を選んだ．文章の量は，全部で 80 文，男女それぞれ 40 文ずつランダムに選んだ．声質変換の種類は，男性から男性，男性から女性，女性から男性，女性から女性の 4 パターンである．

すべてのパターンの MCD を合計し，平均した結果を表 1 に示す．MCD の値は小さいほうがより本物の声質に近いと仮定されているため，ベースラインのほうがわずかに提案手法よりも精度が高いことがわかる．

測定方法	ベースライン	提案手法
MCD	6.37	6.49

表 1：MCD の結果．数値が小さいほうが，変換先の人物の声質に近い．

4.3.2 主観評価実験

我々は，ベースラインと提案手法の声質変換モジュールによって変換された音声に対して平均オピニオン評点（mean opinion score, MOS）テストと AB テストを行なった．MOS テストは変換された音声の自然さを評価するために行なった．AB テストは変換された音声の声質がどのくらい本物に似ているかを評価するために行なった．

テストのために，入力音声は，CMU コーパスから学習データに含まれていない 20 文（男女 10 文ずつ）をランダムに選んだ．変換パターンは客観評価実験同様，男性から男性，男性から女性，女性から男性，女性から女性である．

実験参加者数は，MOS と AB テストともに 14 名である．

MOS テストでは実験参加者に対して，変換された音声の発話としての自然性について質問した．評価段階は，音声品質の主観評価法について定めた ITU-T 勧告 P.800 に基づいて，5 段階である．1 が「最も悪い」，2 が「悪い」，3 が「どちらでもない」，4 が「良い」，5 が「最も良い」である．

MOS テストの結果は図 11 に示している．それぞれのス

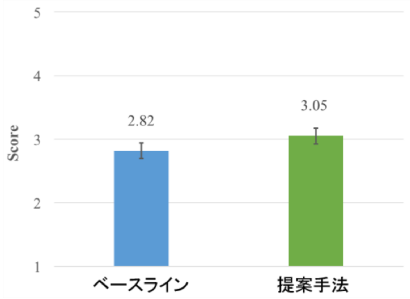

図 11：95%信頼区間での MOS テストの結果．実験参加者は、変換された音声の自然さを 5 段階で評価した．1：最も悪い，2：悪い，3：どちらでもない，4：良い，5：最も良い．

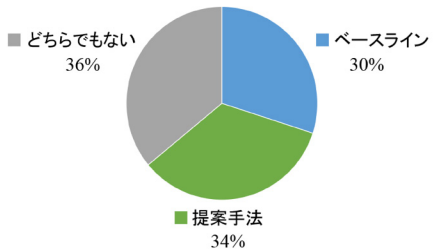

図 12：変換された音声がどのくらい本物と類似しているかについての AB テストの結果．

コアはベースラインが 2.82 であり，提案手法が 3.05 である．

AB テストでは、ベースラインと提案手法で変換した音声に対して，どちらがより本物の声質に似ているかを実験参加者に選んでもらった．もし，優劣がつけられない場合は「どちらでもない」を選んでもらった．

AB テストの結果は，図 12 に示している．ベースラインより，提案手法が選ばれた割合のほうがわずかに高かった．

4.4 結果

実験 1 では提案手法の声質変換モジュールとベースラインを比較した．

その結果，GPU を用いた場合，学習速度および変換時間ともに提案手法のほうが早いことがわかった．

精度については，客観評価実験ではわずかにベースラインのほうが高いが，主観評価実験ではわずかに提案手法のほうが高いことがわかった．両実験において両手法の差はわずかであるので，精度については，両手法ともに同程度と言うことができるだろう．

5. 実験 2：雑音環境化における声質変換

本実験では，以下の 3 パターンの声質変換に対して主観評価実験を行うことで，雑音が含まれている音声を入力した場合の本システム全体の声質変換の精度を確認した．

1) 雑音が含まれた音声を入力し，雑音を除かずに声質変換した場合
2) 雑音が含まれた音声を入力し，音源分離モジュールによって雑音を除去した後，声質変換した場合
3) 話者音声のみのクリーンな音源を入力し，音源分離モジュールを通さずに声質変換した場合

1) は提案システムのうち，音源分離モジュールを使用しない場合の声質変換の精度を確かめるための条件である．2) は提案システム全体の精度を確かめるための条件である．3) は提案システムの精度と比較するための条件である．2) の条件の結果が 3) に近いほど提案システムの精度が高いことがわかる．

5.1 セットアップ

1)，2)に入力するテストデータとして，DEMAND から学習データに含まれない雑音環境を 7 種類（リビング，洗い場，運動場，公園，川，カフェ，バス）を選び，CMU コーパスの音声に重畳した． 重畳する雑音の大きさは，SN 比で 5, 10, 15dB の 3 種類である．テストには， 20 文（男女 10 文ずつ）をランダムに選んだ．変換パターンは実験 1 と同様に，男性から男性，男性から女性，女性から男性，女性から女性である．

5.2 主観評価実験

実験 1 の主観評価実験同様，MOS テストと AB テストを行なった．実験参加者の人数は，MOS テストと AB テストともに 14 名である．

MOS テストの質問内容および回答方法は実験 1 と同様である．結果は，図 13 の左図に示している．

AB テストは，1）と 2) の比較と，2) と 3) の比較において，どちらが変換先の人物の声質に似ているかを選んでもらった．結果は，図 13 に示している．提案システムは，雑音を除去せずに声質変換した場合に比べて，精度高く変換できている．しかし，クリーンな音声を変換した場合と比較すると，まだその精度には達していないことがわかる．

6. 実験 3：聞こえ方の違いによる主観評価

我々は，変換された音声に対して，エフェクトを加えることで精度向上を図れるか，2 つの主観評価実験を行なった．ひとつは，入力音声から取り除いた背景雑音を変換された音声に再合成する実験である．もうひとつは，変換された音声に残響を加える実験である．

6.1 背景雑音の再合成

実験 1, 2 の実験参加者から，「声質変換された音声の粗さやノイズが目立つ」と言う趣旨のコメントがあった．そこで，我々は，一度取り除いた背景雑音を変換後の音声に再合成することで，変換された音声の粗さを目立たなくすることができるのではないかと考えた．再合成する雑音はホワイトノイズではなく日常生活の背景音なので，変換された音声に重畳しても音声自体の自然さは失われないので

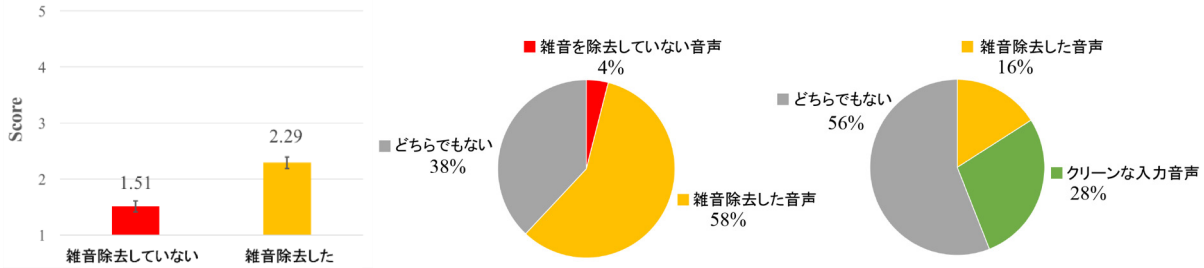

図 13：左図は声質変換された音声の自然さに対する，95%信頼区間での MOS テストの結果．
真ん中の図と右図は声質変換された音声がどのくらい変換先の人物の声質に似ているかについての AB テストの結果．真ん中図は雑音除去せずに声質変換した場合と，雑音除去したあと声質変換した場合の比較．右図は雑音除去したあと声質変換した場合と，クリーンな音声を入力して声質変換した場合の比較．

はないかと予想した．

主観評価実験では，雑音が含まれた音声を入力とし雑音除去したあと声質変換した音声と，同様の処理をしたあと除去した雑音を再合成した音声の 2 つに対して AB テストを行なった．実験参加者には，音声がより自然な発話に聞こえるほうを選択してもらった．テストデータとして雑音が含まれた男女の英語音声を 10 文選んだ．実験参加者の人数は，10 人である．変換パターンは実験 1，2 と同様である．

結果を図 14 に示す．雑音を再合成するより，しなかったほうが高い割合になった．

実験参加者からのコメントとして,「雑音を再合成した音声の中で, 声が二重に聞こえるサンプルがある」とあった．二重に聞こえる原因としては，音源分離モジュールにおいて，音源の分離が完璧に行われず，雑音側に入力話者の声が残ってしまうことが考えられる．そのため，雑音を再合成した音声は，入力話者の音声と変換された音声が重なって聞こえる箇所が生じてしまい，高い評価にならなかったと考えられる．

6.2 残響を加える

実験 1, 2 の実験において，実験参加者から「変換された音声がこもって聞こえる」というコメントがあった．そこで，変換された音声に残響を加えることで，より自然な音声に聞こえるのではないかと我々は考えた．

主観評価実験では，クリーンな音声を入力とし声質変換した音声と，同様の処理をしたあと残響を加えた音声の 2 つに対して AB テストを行なった．実験参加者には，音声がより自然な発話に聞こえるほうを選んでもらった．残響効果の追加には，音声処理のコマンドラインツールの SoX の reverb 処理を利用した．reverb の大きさは 0~100 段階のうち 1 を選択した．テストデータとして雑音が含まれた男女の英語音声を 10 文選んだ．実験参加者の人数は，10 人である．変換パターンは実験 1 と同様である．

結果を図 15 に示す．残響を加えた音声より，残響を加えなかった音声のほうが高い割合になった．

実験参加者からのコメントとして,「変換先の人物が男性の場合は，声がくぐもっているので残響をつけたほうが聞きやすいときがある．一方で女性は残響がないほうが自然

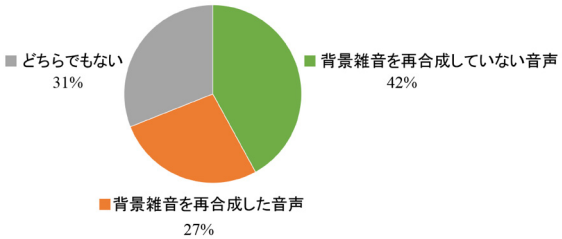

図 14：各条件で声質変換された音声の自然さに対する AB テストの結果．

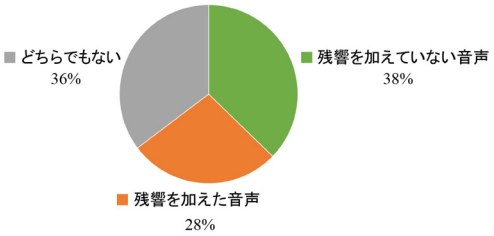

図 15：各条件で声質変換された音声の自然さに対する AB テストの結果．

な発話に聞こえる」とあった．今回の実験では，SoX の reverb 機能を使用したため，残響が大きすぎた可能性がある．もっと自然な大きさで残響を加えることができるアルゴリズムを利用すれば，改善できるかもしれない．

7. 発展性

7.1 アクセント，イントネーション模倣

声質変換システムは，声質だけでなくアクセントやイントネーションも学習すれば，それらも特定の人物のものに変えることができ得る．それが実現すれば，例えば，日本人の英語発音をネイティヴ風に変えることも可能である．あるいは，ネイティヴの発音を日本人が聞き取りやすい発音に変換できるだろう．そうすれば，異言語間でのコミュニケーションがよりスムーズになると予想される．

7.2 カクテルパーティー効果

音源分離モジュールは，音声と雑音の関係だけでなく，さまざまな音を任意にキャンセリングしたり強調したりできる可能性がある． 例えば，カラスの鳴き声だけを消したり, 雑音から車の走行音だけを抜き出したりできるだろう．つまり，あらゆる音に対して，選択的にカクテルパーティー効果を生むことができる．

7.3 音声データセットの自動作成

機械学習の技術が発展するにあたって，データセット作成の自動化は重要な問題である．話者識別と雑音除去の自動化ができれば，例えば映画のように，話者が複数人登場し多様な BGM が含まれる映像から特定の人物の音声データのみを自動で抽出することが可能になる．本研究の技術において，声質変換モデルは話者識別に応用可能であり，音源分離モデルは BGM の自動除去に利用できるだろう．

8. 結論

本研究では, DNN を利用して, 背景雑音を自動で除去し，変換速度がより速い声質変換システムを提案した．

評価実験の結果から以下のことを確かめることができた．

1. 提案した声質変換手法は, 既存の手法と比べて, 精度は保ちつつも学習速度と変換速度が早い．
2. 本システムは，多様な環境雑音が含まれる入力音声に対して，雑音を除去しない場合よりも高精度に声

質変換できる．しかし，その精度は，話者音声のみのクリーンな音声を入力した場合ほどではない．

また，音声の聞こえ方に基づいて，変換済み音声の精度向上を 2 つの方法で試みた．ひとつは，変換された音声に対して残響を加える方法である．もうひとつは，変換された音声に対して，入力音声の背景雑音を再合成する方法である．しかし，主観評価の結果，両手法ともに精度の向上には貢献できなかった．今後，人間の知覚に基づいた精度向上の可能性をさらに探っていきたい．

また，本研究の技術は，音声のアクセント模倣や発音矯正，人工的なカクテルパーティー効果，音声データセットの自動作成などのタスクに応用可能であり，今後その実現可能性についても実証していきたい．

参考文献

[1] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, “Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary,” 2014, pp. 7944–7948.
[2] H. Benisty, D. Malah and K. Crammer, "Non-parallel voice conversion using joint optimization of alignment by temporal context and spectral distortion," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 7909-7913.
[3] J. Bradbury, S. Merity, C. Xiong, and R. Socher, “Quasi-Recurrent Neural Networks,” arXiv preprint arXiv:1611.01576, 2016.
[4] J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, and F. J. Rodríguez-Serrano, “Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings,” EURASIP Journal on Advances in Signal Processing, no. 1, pp. 1–16. 2013.
[5] P. Chandna, M. Miron, J. Janer, and E. Gómez, “Monoaural Audio Source Separation Using Deep Convolutional Neural Networks,” 13th International Conference on Latent Variable Analysis and Signal Separation (LVA ICA2017), 2017, pp. 258–266.
[6] DEMAND: Diverse Environments Multichannel Acoustic Noise Database, Parole.loria.fr, http://parole.loria.fr/DEMAND/
[7] D. Erro, A. Moreno, and A. Bonafonte, “INCA algorithm for training voice conversion systems from nonparallel corpora,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 944–953, 2010.
[8] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus. 1993.
[9] T. J. Hazen, W. Shen, and C. White, “Query-by-example spoken term detection using phonetic posteriorgram templates,” Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on, 2009, pp. 421–426.
[10] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, “Voice Conversion Using Dynamic Kernel Partial Least Squares Regression,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 3, pp. 806–817, 2012.
[11] P. Sen Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 23, no. 12, pp. 2136–2147, 2015.
[12] K. Kintzley, A. Jansen, and H. Hermansky, “Event selection from phone posteriorgrams using matched filters,” Twelfth Annual Conference of the International Speech Communication Association, pp. 1905–1908, 2011.
[13] J. Kominek and A. W. Black, “The CMU Arctic speech databases,” Fifth ISCA Workshop on Speech Synthesis, 2004.
[14] A. Kumar and D. Florencio, “Speech Enhancement In Multiple-Noise Conditions using Deep Neural Networks,” arXiv preprint arXiv:1605.02427, 2016.
[15] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech Enhancement Based on Deep Denoising Autoencoder,” Interspeech, pp. 436–440, 2013.
[16] K. Masaka, R. Aihara, T. Takiguchi, and Y. Ariki, “Multimodal Voice Conversion Using Non-Negative matrix Factorization In Noisy Environments,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1561–1565.
[17] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, “Voice conversion in high-order eigen space using deep belief nets,”. Interspeech, pp. 369–372, 2013.
[18] A. van den Oord et al., “WaveNet: A Generative Model for Raw Audio”, arXiv preprint arXiv:1609.0349, pp. 1–15, 2016.
[19] M. Pal, D. Paul, M. Sahidullah, and G. Saha, “Robustness of Voice Conversion Techniques Under Mismatched Conditions,” arXiv preprint arXiv:1612.07523, 2016.
[20] D. Rethage, J. Pons, and X. Serra, “A Wavenet for Speech Denoising,” arXiv preprint arXiv:1706.07162, 2017.
[21] P. Scalart and J. V. Filho, “Speech enhancement based on a priori signal to noise estimation,” 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 2. pp. 629–632, 1996.
[22] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” in IEEE Transactions on Speech and Audio Processing, 1998, vol. 6, no. 2, pp. 131–142.
[23] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” Multimedia and Expo (ICME), 2016 IEEE International Conference on, 2016. p. 1-6.
[24] R. Takashima, R. Aihara, T. Takiguchi, and Y. Ariki, “Noise-Robust Voice Conversion Based on Spectral Mapping on Sparse Space,” in 8th ISCA Workshop on Speech Synthesis, 2013, no. 6, pp. 71–75.
[25] R. Takashima, T. Takiguchi, and Y. Ariki, “Exemplar-based voice conversion in noisy environment,” In: Spoken Language Technology Workshop (SLT), 2012, pp. 313-317.
[26] J. Tao, M. Zhang, J. Nurminen, J. Tian, and X. Wang, “Supervisory data alignment for text-independent voice conversion,” IEEE Trans. Audio, Speech Lang. Process, vol. 18, no. 5, pp. 932–943, 2010.
[27] T. Toda, A. W. Black and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222-2235, Nov. 2007.
[28] S. Wu, S. Zhong, and Y. Liu, “Deep residual learning for image steganalysis,” Multimedia Tools and Applications, pp. 1–17, 2017.
[29] Y. Xu, J. Du, L. Dai, and C. Lee, “A Regression Approach to Speech Enhancement Based on Deep Neural Networks,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 23, no. 1, pp. 7–19, 2015.
[30] YOURTONE | Electric artificial larynx | DENCOM, Dencom.co.jp, http://www.dencom.co.jp/product/yourtone/yt2_eng.html