Vol.2017-SLP-119 No.9
2017/12/21

CycleGANを用いた高品質なノンパラレル声質変換

房 福明[1]　山岸 順一[1,2]　越前 功[1]

概要：近年，機械学習の進展により声質変換の性能が大幅に向上した．しかし，学習データが対とならないノンパラレルの場合，ソース話者とターゲット話者の特徴を精密にマッチすることが難しい．ノンパラレル声質変換モデルの学習はまだ困難であり，変換性能はまだ低い問題がある．一方，画像変換分野ではペアなしの画像データベースから変換モデルを学習する方法としてCycleGANが注目されている．CycleGANはGANの一種であり，複数個のgeneratorとdiscriminatorを持つ．また，generatorは入力データの一部の情報を維持しながら，discriminatorとの競争学習によりターゲットドメインへの変換ができる特徴がある．そこで，本研究はこのアイディアに基づいてCycleGANをノンパラレル声質変換に適用する方法を提案する．提案手法では，ソース話者とターゲット話者の類似特徴を直接マッチするのではなく，ソース話者の一部の言語情報を維持しながら話者特徴をターゲット話者にできるだけ近付けるように変換モデルを学習する．被験者評価実験より，提案手法は標準のGANに基づいたパラレル声質変換を上回ったことを示す．

High-quality nonparallel voice conversion using CycleGAN

Fuming Fang[1]　Junichi Yamagishi[1,2]　Isao Echizen[1]

Abstract: Recently, voice conversion (VC) based on deep learning has achieved remarkable performance. However, it is still difficult to train a mapping model using nonparallel training samples. In this work, we propose a high-quality nonparallel VC training method based on CycleGAN. A CycleGAN is a kind of generative adversarial network (GAN) originally developed for unpaired image-to-image translation. This model can be learned by an approach that a part of input information is kept while the corresponding distribution of the input data can be converted into a target distribution without paired training samples. Experimental results show that the proposed method outperforms a standard GAN-based parallel VC system.

1. はじめに

声質変換は言語情報を保ちながらある話者の声が別話者に聞こえるように音声信号を編集する技術である [1]．この技術は幅広く応用されている．例えば，音声合成システムの出力音声カスタマイズ化 [2]，映画吹き替えの制作 [3]，外国語勉強のサポート [4] と歌声の変換 [5] などを挙げられる．また，発声障碍者の発話内容を聞き易くするためのような福祉応用もある [6]．

声質変換システムを利用するため，変換元のソース話者と変換先のターゲット話者の音声データを用いて変換モデルを事前に学習しておく必要がある．ソース話者とターゲット話者の音声データは対であるか対ではないかによって，声質変換システムは「パラレル」と「ノンパラレル」の2種類に分けられる．パラレルの場合，ソース話者とターゲット話者は同じ発話内容で収録したデータを利用する．ノンパラレルの場合ではこのような制限を用いずに任意の発話内容でも構わない．従って，ノンパラレルシステムはパラレルより実用性が高く，構築も安価で実現できる利点がある．しかし，パラレルシステムでは対となる学習データを用いるため，両話者の音響特徴の対応関係を容易にマッチでき，学習したモデルは常に高い性能を持つ．ノンパラレルの場合では音響特徴量の対応関係を高精度にマッチすることが困難であり，モデルの性能が低い問題がある．そこで，本研究はノンパラレル声質変換システムの性能を高めることを目指している．

既存の声質変換手法を俯瞰すると，パラレルの場合では

[1] 国立情報学研究所
〒 101-8430，東京都千代田区一ツ橋 2-1-2
[2] エジンバラ大学

まず動的時間伸縮法（dynamic time warping: DTW）[7] を用いてソース話者とターゲット話者の同じ発話に対して，類似の音響特徴量を時間軸上でマッチする．次にマッチした音響特徴量ペアを用いて変換モデルを学習する．変換モデルの学習では，Stylianou ら [8] はペアにした特徴量の分布を混合ガウスモデル（Gaussian mixture model: GMM）により表現し，変換関係をモデル化した．戸田ら [9] はさらに動的特徴量と系列内の変動を考慮して GMM に基づいた手法を改善した．また，Desai ら [10] はニューラルネットワーク（neural network: NN）[11] を用いて，ソースとターゲット話者の音響特徴量をそれぞれ入力と教師信号とし，変換関係を直接モデル化した．Sun ら [12] では BLSTM（bidirectional long short-term memory）[13] を用いてコンテキスト情報も一緒にモデル化し，NN に基づいた手法より大きな改善を得た．近年，GAN（generative adversarial network）[14] は強力な学習方法として声質変換に応用され始めた．この学習方法を用いて，Kaneko ら [15] は彼らの sequence-to-sequence 声質変換システムに適用し，従来の平均二乗誤差（Mean Squared Error: MSE）に基づいた学習方法より高い性能であることを示した．

既存のノンパラレル声質変換手法では，大きく 2 種類に分けられる．特徴量ペアのマッチ方法と，話者情報の置換方法である．特徴量ペアのマッチ方法は音素や類似の特徴量などを自動的にマッチすることでソース話者とターゲット話者のデータをペアにして，パラレル声質変換と同様な方法で変換モデルを学習する．音響特徴量をペアにするため，Ye ら [16] は隠れマルコフモデル（hidden Markov model: HMM）[17] を用いて特徴量に HMM 状態番号を付ける方法を用いた．Erro ら [18] では K 近傍法を利用して，初期の変換モデルより変換したソース話者の特徴量とターゲット話者の特徴量のペアを作成し，モデルの更新と特徴量ペアのマッチを収束まで繰り返し行う方法を用いた．そして，この方法の改良バージョンとして，Benisty ら [19] ではコンテキスト情報，及びマッチ探索時にソース話者からターゲット話者への変換とこれの逆の変換を考慮する方法を用いた．

話者情報の置換方法では，音響特徴量から言語成分と話者成分を分離し，話者成分を置き換えることで声質変換を行う．このアイディアに基づいて，Song ら [20] は MAP（maximum a posteriori）適応手法 [21] を用いて背景モデルから適応した GMM を利用して話者成分を表した．また，Nakashika ら [22] では改善した制限付きボルツマンマシン（restricted Boltzmann machines: RBM）[23] を用いてもっと精密な手法を提案した．この改善した RBM の重みでは話者重みと共通重みから構成し，複数の話者のデータから学習できる．そして，GAN に基づいた学習手法も最近提案されいる．例えば，Hsu ら [24] は VAE（variational autoencoder）[25] と WGAN（Wasserstein GAN）[26] を組み合わせる方法を用いた．VAE のエンコーダは音声特徴量から言語成分の分布を学習しておいて，デコーダは話者情報を表すコードとこの言語分布に基づいてターゲット話者の音声特徴量を生成する．WGAN は変換した特徴量の分布とターゲット話者の特徴量の分布が区別できないように VAE を学習させる．

本研究はノンパラレル声質変換の性能を向上させるために，CycleGAN（cycle-consistent adversarial network）[27] を用いる手法を提案する．CycleGAN は GAN の一種であり，もともとペア画像なしのデータベースに基づいた画像変換手法である．このモデルは入力情報の一部を保ちながら，ターゲットドメインへ変換できる特徴がある．CycleGAN を用いてノンパラレル声質変換モデルを学習する場合では，従来の性質変換手法とは違い，特徴量ペアをマッチする必要がなく，マッチエラーによる悪影響を回避できる．また，従来の話者情報の置換方法に基づいたノンパラレル手法とは違い，提案手法は言語成分と話者成分を分離せず，ソース話者の言語情報を維持しながらターゲット話者の特徴量の分布とできるだけ一致するように変換モデルを学習する．実験では，提案手法は GAN に基づいたパラレル声質変換手法より高い性能であることを示す．

2. Generative adversarial networks

本章は GAN について紹介する．また，以下に Goodfellow らが最初に提案した GAN [14] は「標準 GAN」と表記する．

2.1 標準 GAN

標準 GAN は discriminator（D）と generator（G）より構成される．D と G はニューラルネットワークを用いる．**図 1** は標準 GAN の構成を示す．標準 GAN の目的はノイズ $\mathbf{z}$ から G を通してイメージを生成することである．通常の学習方法は G の出力を教師信号と比較して誤差を最小化するように G のパラメータを推定するが，標準 GAN を用いた学習では，D は入力データが学習データ $\mathbf{x}$（イメージ）であるか生成したデータであるかを判断すると同時に G はできるだけ D を騙すように生成したデータを学習データの分布に近づけることでパラメータを推定する．D と G の繰り返しの更新により最終的に G はノイズからイメージを生成できるようになる．

標準 GAN の目的関数は式 1 で表現できる．式中の $\mathbb{E}$ は期待値を表す．$D(\cdot)$ は入力データが学習データであるか生成データであるかの確率を表す．

$$\begin{aligned}\mathcal{L}_{GAN}(G,D) = {} & \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] \\ & + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \end{aligned} \quad (1)$$

学習時，D と G のパラメータを交互に更新する．D を更新するときは式 1 を最大化し，G を更新するときは式 1 を

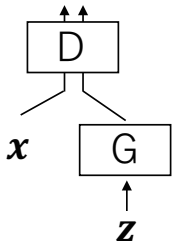

図 1　標準 GAN の構成のイメージ．「D」と「G」はそれぞれ discriminator と generator を表す．$\mathbf{x}$ は学習データで，$\mathbf{z}$ はノイズである．

最小化する．バックプロパゲーション（backpropagation）[28] アルゴリズムを用いて式 2 を解くことでパラメータを学習する．

$$G^* = \arg\min_G \max_D \mathcal{L}_{GAN}(G, D) \tag{2}$$

標準 GAN をパラレル声質変換に応用する場合，ソース話者の音響特徴量 $\mathbf{x}_c$ をコンディションとしてモデルに入力し，式 1 を下式になる．$\mathbf{x}$ はターゲット話者の音響特徴量であり，$\mathbf{z}$ は潜在変数である．

$$\begin{aligned}\mathcal{L}_{GAN}(G, D) = &\mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x},\mathbf{x}_c)}[\log D(\mathbf{x}, \mathbf{x}_c)] \\ &+ \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\mathbf{x}_c\sim p_{data}(\mathbf{x}_c)}[\log(1 - D(G(\mathbf{z}, \mathbf{x}_c)))]\end{aligned} \tag{3}$$

また，学習を安定させるため，上式に MSE を組み合わせることで G を更新する．

2.2 CycleGAN

図 2 に CycleGAN の構成のイメージを示す．CycleGAN は標準 GAN の拡張であり，2 つずつの discriminator（$D_{\mathbf{x}}$ と $D_{\mathbf{y}}$）と generator（G と F）より構成される．図に示しているように，CycleGAN は 2 つの変換方向がある．「forward」の $\mathbf{x} \to \hat{\mathbf{y}} \to \hat{\mathbf{x}}$ と「backward」の $\mathbf{y} \to \hat{\mathbf{x}} \to \hat{\mathbf{y}}$ である．この仕組みにより変換元から変換先及び変換先から変換元への変換モデル（G と F）を同時に学習できる特徴がある．つまり，異なるドメインのデータサンプル $\mathbf{x}$ と $\mathbf{y}$ は変換モデル G と F を通じて対応するそれぞれの変換結果 $\hat{\mathbf{y}}$（$= G(\mathbf{x})$）と $\hat{\mathbf{x}}$（$= F(\mathbf{y})$）を得ることが可能である．

CycleGAN の目的は対とならないデータから変換モデルを学習することである．これを実現するため，1 つのアイディアは forward と backward 変換はもとのデータ $\mathbf{x}$ と $\mathbf{y}$ を再現する制限を用いる．つまり，$\hat{\mathbf{x}} \approx \mathbf{x}$ と $\hat{\mathbf{y}} \approx \mathbf{y}$ を満すことで，入力データの一部の情報を維持する．次の式で表すサイクル損失を最小化することで制限できる．$\|\cdot\|_1$ は L1 ノルムである．

$$\begin{aligned}\mathcal{L}_{cyc}(G, F) = &\mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\| F(G(\mathbf{x})) - \mathbf{x} \|_1] \\ &+ \mathbb{E}_{\mathbf{y}\sim p_{data}(\mathbf{y})}[\| G(F(\mathbf{y})) - \mathbf{y} \|_1]\end{aligned} \tag{4}$$

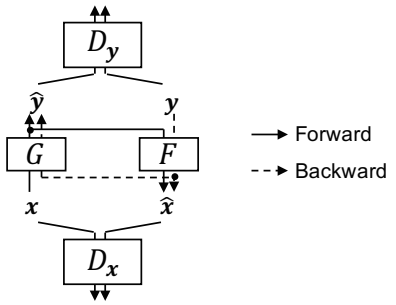

図 2　CycleGAN の構成のイメージ．「$D_{\mathbf{x}}$」と「$D_{\mathbf{y}}$」は discriminator であり，「G」と「F」は generator を表す．$\mathbf{x}$ と $\mathbf{y}$ は異なるドメインからの学習データで，$\hat{\mathbf{x}}$ と $\hat{\mathbf{y}}$ は生成したデータである．

次に，入力データの一部の情報を維持しながらターゲットデータの分布に近づけるようにモデルのパラメータを調節すれば，対とならないデータから変換モデルを学習できると考えられる．そこで，標準 GAN と同じアイディアを利用し，式 1 のような目的関数を用いる．$\mathbf{x}$ から $\mathbf{y}$ へ変換する場合の目的関数は式 5 より定義される．同様に，$\mathbf{y}$ から $\mathbf{x}$ への変換は $\mathcal{L}_{GAN}(G, D_{\mathbf{x}}, \mathbf{y}, \mathbf{x})$ を用いる．

$$\begin{aligned}\mathcal{L}_{GAN}(G, D_{\mathbf{y}}, \mathbf{x}, \mathbf{y}) = &\mathbb{E}_{\mathbf{y}\sim p_{data}(\mathbf{y})}[\log D_Y(\mathbf{y})] \\ &+ \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log(1 - D_Y(G(\mathbf{x})))]\end{aligned} \tag{5}$$

最終的に式 4 と 5 などを組み合わせることで，対とならないデータセットから正しい変換を学習できる可能性がある．全体の目的関数は式 6 より定義し，λ はサイクル損失の重要度を表すハイパーパラメータである．モデル全体のパラメータは $\arg\min_{G,F} \max_{D_{\mathbf{x}},D_{\mathbf{y}}} \mathcal{L}(G, F, D_{\mathbf{x}}, D_{\mathbf{y}})$ より学習する．

$$\begin{aligned}\mathcal{L}(G, F, D_{\mathbf{x}}, D_{\mathbf{y}}) = &\mathcal{L}_{GAN}(G, D_{\mathbf{y}}, \mathbf{x}, \mathbf{y}) \\ &+ \mathcal{L}_{GAN}(F, D_{\mathbf{x}}, \mathbf{y}, \mathbf{x}) \\ &+ \lambda\mathcal{L}_{cyc}(G, F)\end{aligned} \tag{6}$$

3. CycleGAN を用いたノンパラレル声質変換

本研究ではソース話者の音響特徴量を抽出し，ターゲット話者の音響特徴量に変換して合成することで声質変換を行う．メルケプストラム，基本周波数（F_0）と非周期成分を音響特徴量として使用する．図 3 に提案する声質変換手法の概略を示す．本研究では，各音響特徴量を独立的に変換を行う．そして，メルケプストラムはさらに高次元成分と低次元成分に分けて変換を行う．高次元成分はスペクトル微細構造に対応しており，あまり言語情報や話者情報を含めないため，この成分を直接コピーしてターゲット話者の特徴量として利用する．また，低次元成分はスペクトル包絡に対応しており，明らかに大量な言語情報と話者情報を含めるため，本研究は CycleGAN を用いてこの成分を中

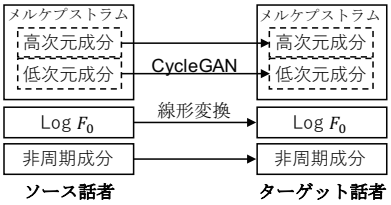

図 3　提案する CycleGAN を用いたノンパラレル声質変換.

心に変換を行う．F_0 の変換では，ソース話者とターゲット話者の $logF_0$ の平均と標準分散が一致するように線形変換を行う．そして，非周期成分は話者性に関して大きな影響がない [29] と報告されているため，本研究では非周期成分を変換しないで，直接ターゲット話者の特徴量として使用し音声を合成する．

CycleGAN を用いた変換では，$\mathbf{x}$ はソース話者のメルケプストラムの低次元成分に対応し，$\mathbf{y}$ はターゲット話者のメルケプストラムの低次元成分に対応する．学習する際にフレームごとの特徴量をランダムに選択しミニバッチを作り，バックプロパゲーションを用いて式 6 よりパラメータを学習する．

4. 実験

4.1 実験条件

本研究では提案する CycleGAN に基づいたノンパラレル声質変換手法を次の 2 種類のベースラインパラレル手法と比較した．1 つ目はエジンバラ大学が開発したオープンソフトウエア Merlin を用いたパラレル声質変換手法である．Merlin はディープニューラルネットワークを利用している．2 つ目は標準 GAN に基づいたパラレル声質変換手法である．提案手法とベースライン手法は男性から女性及び女性から男性への声質変換を行った．また，有意水準を 0.05 に設定し，t-test を用いて有意差の検定を行った．他の設定は具体的に以下に説明する．

4.1.1 データベース

学習データとテストデータは ALAGIN セット B を利用し，男性話者と女性話者を 1 名ずつ選択した．ベースラインのパラレル声質変換手法について，モデルを学習するため，同じ発声内容の音声データを話者ごとに 200 発話ずつ用いた．また，提案するノンパラレル声質変換手法のモデルを学習するため，異なる発声内容の音声データを話者ごとに 200 発話用いた．テストデータについては，全ての手法で共通のデータを使用し，学習データに含めないそれぞれの話者の音声データを 50 発話ずつ用いた．

4.1.2 音響特徴量の抽出設定

音響特徴量は WORLD [30] と SPTK [31] を用いて抽出した．メルケプストラムは 49 次元ベクトルであり，最初の 25 次元を低次元成分とし，残りの 24 次元を高次元成分とした．そして，コンテキストを考慮するため，メルケプストラムの 1 次と 2 次微分特徴を利用した．従って，モデルを学習するための特徴量ベクトルの次元数は 75 次元（低次元の 25 次元及びその 1 次と 2 次微分）であった．そして，ベースラインのパラレル声質変換手法については，モデルを学習する前に DTW を用いて両話者の類似データペアをマッチするプロセスを行った．

表 1　年齢帯別の被験者の分布.

年齢	18-29	30-39	40-49	50-59	60+
人数	10	42	34	18	6

表 2　性別ごとの被験者数.

	男性	女性
人数	55	55

4.1.3 モデル構造及び学習と変換の設定

Merlin を用いた手法のネットワーク，標準 GAN と CycleGAN の discriminator と generator は 6 層のフィードフォワード全結合ネットワークを用いた．隠れ層のユニット数はそれぞれ 128，256，256，128 であり，隠れ層の活性化関数に sigmoid を用いた．また，標準 GAN と CycleGAN の discriminator へ入力するコンディション $\mathbf{x}_c$ 及び generator へ入力する潜在変数 $\mathbf{z}$ を省略した．標準 GAN と CycleGAN は TensorFlow [32] を用いて実装した．学習率は常に 0.001 を設定したが，discriminator を更新する際は 0.0001 を設定した．そして，ミニバッチはランダムに選択した 128 フレームの音響特徴量ベクトルより構築した．エポックサイズについて，Merlin を用いた場合は 60 に設定し，標準 GAN と CycleGAN の場合は 400 に設定した．CycleGAN を学習するため，式 6 の λ を 10 に設定した．音声を合成する前に MLPG（maximum likelihood parameter generation）[33] とポストフィルタ [34] を用いて音響特徴量をスムージングした．

4.1.4 被験者評価の設定

上述した 3 種類の変換手法により，合計 300（$= 50 \times 3 \times 2$）発話を変換した．これらの発話を被験者に聞いてもらい，自然のリファレンス音声と比較して，音声品質及び話者類似性について評価した．評価基準は平均オピニオン評点（mean opinion score: MOS）を用いた．MOS は 5 段階の評点があり，5 点の場合は最もよい評価，1 点の場合は最も悪い評価の意味を表す．被験者はクラウドソーシングを利用して集めた．被験者は 1 回につき，ランダムに選択した 12 発話を評価し，最大で 6 回評価できる．最終的に被験者 110 人は合計 600 回評価した．つまり，平均して変換した発話ごとに 24 回ずつ評価を行った．表 1 と表 2 は被験者の年齢帯別及び性別の分布である．

4.2 実験結果

図 4 と**図 5** はそれぞれ音声品質と話者類似性に関する

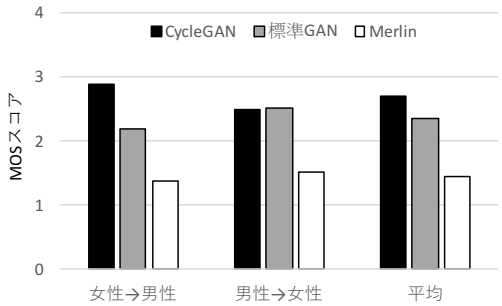

図 4 音声品質に関する被験者評価の結果.

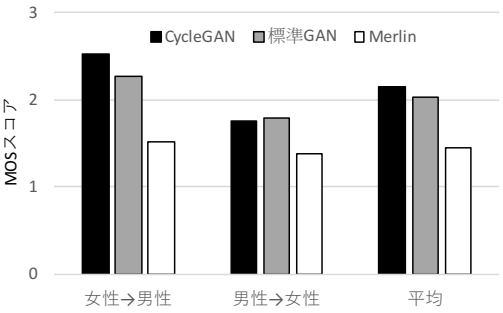

図 5 話者類似性に関する被験者評価の結果.

被験者の評価結果を示す．平均の音声品質と話者類似性について，提案する CycleGAN を用いたノンパラレル声質変換手法は，標準 GAN と Merlin に基づいたパラレル声質変換手法より有意な改善が見られた．CycleGAN を用いたノンパラレル声質変換がパラレル声質変換を上回った理由として，DTW を使わず入力データの言語情報を一部保ちながらターゲット話者のデータ分布に近づくような仕組みを用いたため，パラレル声質変換のようなミスマッチより発生したエラーを回避できたと考えられる．もう一つの理由として，ベースラインのパラレル声質変換手法を用いた場合では，DTW からマッチしたデータペアしか学習できなかったのに対し，CycleGAN を用いた場合では，任意のデータ対を用いてモデルを学習できたと考えられる．

また，提案手法を標準 GAN に基づいた手法と比べ，男性から女性への変換は改善が見られなかった．これは 49 次元中の 25 次元のメルケプストラム係数だけを用いたため，変換のための情報が足りなかったためと考えられる．変換性能をさらに向上させるために適切な次元数を選択する必要がある．

5. まとめと課題

本研究では CycleGAN を用いたノンパラレル声質変換手法を提案した．提案手法は従来のパラレル声質変換手法とは違い，入力データの一部の言語情報を保ちながらターゲット話者の特徴量の分布に近づけるように変換モデルを学習するため，特徴量ペアをマッチする必要がなく，ミスマッチにより発生するエラーを回避できる．被験者実験では，提案手法は標準 GAN と Merlin に基づいたパラレル声質変換手法より高い性能であることを示した．

しかし，CycleGAN は入力データの言語情報を厳密に制約することが難しいため，変換した音声に音素などが置き換える可能性がある．モデルを正しく学習するためにランダムシードというハイパーパラメータを適切に調節する必要がある．そこで，今後は言語情報を厳密に制約できるように CycleGAN を改善する必要がある．そして，変換の性能を向上させることも今後の課題である．

謝辞

本研究の一部は MEXT 科研費 15H01686，16H06302 と 17H04687 の助成を受けたものです．

参考文献

[1] D.G. Childers, K. Wu, D.M. Hicks, and B. Yegnanarayana, “Voice conversion,” *Speech Communication*, vol. 8, no. 2, pp. 147 – 158, 1989.

[2] S. Mohammadi and A. Kain, “An overview of voice conversion systems,” *Speech Communication*, vol. 88, no. Supplement C, pp. 65 – 82, 2017.

[3] O. Turk and L. Arslan, “Subband based voice conversion,” in *Seventh International Conference on Spoken Language Processing*, 2002.

[4] S. Zhao, S. Koh, S. Yann, and K. Luke, “Feedback utterances for computer-adied language learning using accent reduction and voice conversion method,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8208–8212.

[5] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, “Voice timbre control based on perceived age in singing voice conversion,” *IEICE Transactions on Information and Systems*, vol. E97.D, no. 6, pp. 1419–1428, 2014.

[6] A. Kain, J. Hosom, X. Niu, J. Santen, M. Fried-Oken, and J. Staehely, “Improving the intelligibility of dysarthric speech,” *Speech Communication*, vol. 49, no. 9, pp. 743 – 759, 2007.

[7] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-26, no. 1, pp. 43–49, 1978.

[8] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131–142, 1998.

[9] T. Toda, A. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[10] S. Desai, A. Black, B. Yegnanarayana, and K. Prahallad, “Spectral mapping using artificial neural networks for voice conversion,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[11] Jürgen Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Networks*, vol. 61, pp. 85–117, 2015.

[12] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4869–4873.

[13] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[15] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," *Proc. Interspeech 2017*, pp. 1283–1287, 2017.

[16] H. Ye and S. Young, "Voice conversion for unknown speakers," in *Eighth International Conference on Spoken Language Processing*, 2004.

[17] Lawrence R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, pp. 257–286.

[18] 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.

[19] H. Benisty, D. Malah, and K. Crammer, "Non-parallel voice conversion using joint optimization of alignment by temporal context and spectral distortion," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7909–7913.

[20] P. Song, W. Zheng, and L. Zhao, "Non-parallel training for voice conversion based on adaptation method," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6905–6909.

[21] C. Lee, C. Lin, and B. Juang, "A study on speaker adaptation of parameters of continuous density hidden Markov models," in *Proceedings IEEE Transactions on Signal Processing*, 1990, pp. 145–148.

[22] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.

[23] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[24] C. Hsu, H. Hwang, Y. Wu, Y. Tsao, and H. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," in *Interspeech*. ISCA, 2017.

[25] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[26] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[27] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[28] DRGHR Williams and GE Hinton, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–538, 1986.

[29] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with straight mixed excitation," in *Proc. ICSLP*, 2006, pp. 2266–2269.

[30] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[31] SPTK Working Group et al., "Speech signal processing toolkit (SPTK)," *http://sp-tk.sourceforge.net*, 2009.

[32] "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[33] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 3, pp. 1315–1318.

[34] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporating a mixed excitation model and postfilter into HMM-based text-to-speech synthesis," *Systems and Computers in Japan*, vol. 36, no. 12, pp. 43–50, 2005.