社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Anti-spoofing に敵対する DNN 音声変換の評価

齋藤　佑樹† 　高道慎之介† 　猿渡　洋†

† 東京大学大学院 情報理工学系研究科　〒 113–8656 東京都文京区本郷 7-3-1

あらまし　統計的パラメトリック音声合成において，生成される合成音声の音質劣化は深刻な問題となる．これまでに我々はテキスト音声合成において，合成音声による声のなりすましを防ぐ技術である anti-spoofing に敵対する音響モデル学習法 (敵対的 DNN 音声合成) を提案し，有効性を示している．本稿では，敵対的 DNN 音声合成の枠組みを音声変換へ適用し，高音質な音声変換を実現するための DNN 音響モデルの学習アルゴリズムを提案する．実験的評価により，(1) Feed-Forward 型ネットワークを用いた特徴量変換に基づく DNN 音声変換，及び，本稿で新たに提案する，(2) highway network を用いた差分スペクトル推定に基づく DNN 音声変換の両方において提案アルゴリズムによる音質改善効果が得られることを示す．

キーワード　DNN 音声変換， anti-spoofing， 敵対的 DNN 音声合成， highway network， 差分スペクトル， 過剰な平滑化

Evaluation of DNN-Based Voice Conversion Deceiving Anti-spoofing Verification

Yuki SAITO†, Shinnosuke TAKAMICHI†, and Hiroshi SARUWATARI†

† Graduate School of Information Science and Technology, The University of Tokyo, Hongo 7–3–1, Bunkyo-ku, Tokyo, 113–8656 Japan

Abstract　This paper proposes a novel training algorithm for high-quality Deep Neural Network (DNN)-based voice conversion. To improve speech quality in DNN-based text-to-speech synthesis, we have proposed a training algorithm to deceive anti-spoofing verification, called adversarial DNN-based speech synthesis. The anti-spoofing is a discriminator to distinguish natural and synthetic speech. This paper extends this idea to DNN-based voice conversion, and we build the acoustic models that can deceive the anti-spoofing verification. To evaluate the proposed algorithm, we conduct evaluations using two conversion frameworks: speech feature conversion using Feed-Forward neural networks and spectral differentials estimation using highway networks from input to output, which is proposed in this paper. The evaluation results successfully demonstrate the speech-quality improvements for both frameworks.

Key words　DNN-based voice conversion, anti-spoofing verification, adversarial DNN-based speech synthesis, highway networks, spectral differentials, over-smoothing

1. はじめに

入力特徴量と出力音声特徴量の対応関係を統計的な音響モデルで表現する統計的パラメトリック音声合成方式 [1] は，その汎用性の高さから広く研究されている．特に，音声認識 [2] における成功を受け提案された Deep Neural Network (DNN) 音声合成方式 [3] は, 従来の hidden Markov model を用いたテキスト音声合成 [4] や，Gaussian mixture model を用いた音声変換 [5] と比較して高音質な合成音声を生成できる．しかしながら，DNN 音声合成により生成される合成音声の音質も，自然音声と比較すると著しく劣化する傾向にある．この音質劣化の要因のひとつとして，統計的な音響モデルにより生成される合成音声特徴量系列の過剰な平滑化が挙げられる．

合成音声の音質は，自然音声の統計量を復元することで改善される．例えば，周波数分解された音声特徴量時系列の二次モーメント (変調スペクトル) [6] を復元することは，テキスト音声合成のみならず，音声変換においても有効である．近年，我々はこの枠組みを拡張し，anti-spoofing に敵対するように DNN 音響モデルを学習するテキスト音声合成 (敵対的 DNN 音声合成) を提案した [7]．Anti-spoofing は合成音声による声

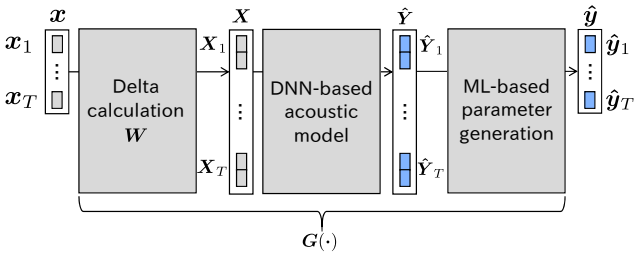

図 1　Feed-Forward 型ネットワークを用いた音声変換

Fig. 1　Voice conversion using Feed-Forward neural networks.

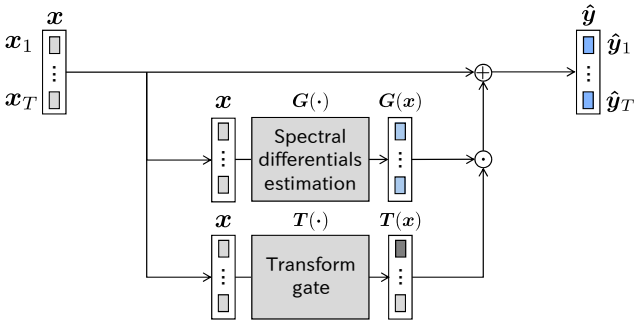

図 2　Highway network を用いた音声変換

Fig. 2　Voice conversion using highway networks.

のなりすましを検出する識別器であり，テキスト音声合成の DNN 音響モデルは，生成する音声特徴量が anti-spoofing を詐称するように学習される．直感的に表現すると，このアルゴリズムは，声のなりすましを成功させるように音声合成の音響モデルを更新する方法である．Anti-spoofing に敵対する学習は，自然音声と合成音声の従う確率分布間の距離を最小化させるため，自然音声の統計量を復元する音響モデルを構築できる．更に，DNN を用いた anti-spoofing により，非常に複雑な確率分布を仮定できる．

本稿では，この敵対的 DNN 音声合成を音声変換に導入する．スペクトル特徴量を変換する音響モデルは，anti-spoofing に敵対するように学習される．実験的評価では，Feed-Forward 型ネットワークを用いた特徴量変換法 (図 1) に加え，本稿で新たに提案する，入出力ユニット間を結ぶ highway network [8–10] を用いた差分スペクトル法 (図 2) において，提案アルゴリズムを評価する．評価結果より，提案アルゴリズムによる音質改善効果が得られることを示す．

2.　従来の DNN 音声変換の枠組み

DNN を用いた統計的パラメトリック音声変換における従来の枠組みについて論述する．

2.1　Minimum Generation Error (MGE) 学習

DNN 音声変換の学習部では，自然音声の特徴量系列と合成音声の特徴量系列から計算される損失関数を最小化する．本稿で採用する Minimum Generation Error (MGE) 学習 [11] の損失関数は，自然音声の特徴量系列 $\boldsymbol{y} = \left[\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top\right]^\top$ と，合成音声の特徴量系列 $\hat{\boldsymbol{y}} = \left[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top\right]^\top$ の二乗誤差として次式で与えられる．

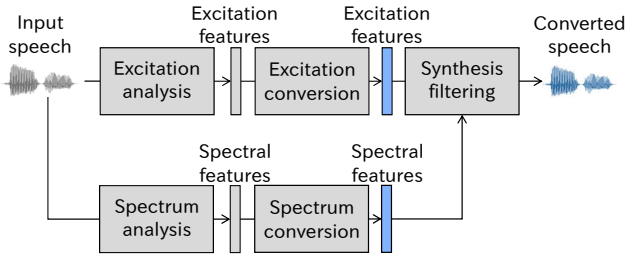

図 3　特徴量変換に基づく音声変換

Fig. 3　Voice conversion based on speech feature conversion.

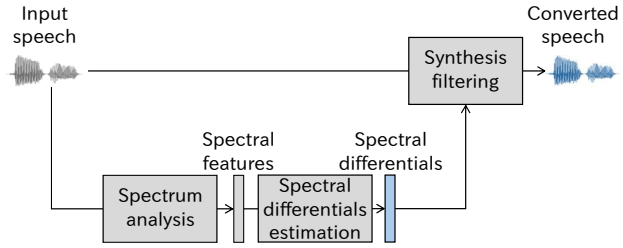

図 4　差分スペクトル推定に基づく音声変換

Fig. 4　Voice conversion based on spectral differentials estimation.

$$L_{\mathrm{G}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = \frac{1}{T}\left(\hat{\boldsymbol{y}} - \boldsymbol{y}\right)^\top \left(\hat{\boldsymbol{y}} - \boldsymbol{y}\right) \tag{1}$$

ここで，T はフレーム数であり，$\boldsymbol{y}_t$ は時刻 t における音声特徴量である．本稿では，図 1 に示すように，入力音声特徴量系列 $\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_T^\top\right]^\top$ に対して静的・動的特徴量系列間の制約行列 $\boldsymbol{W}$ [4] を適用した後，DNN により合成音声の静的・動的特徴量系列を推定する．その後，最尤パラメータ生成 [12] により $\hat{\boldsymbol{y}}$ を生成する．以降では，このモジュールを $\boldsymbol{G}(\cdot)$ とし，音声特徴量変換を $\hat{\boldsymbol{y}} = \boldsymbol{G}(\boldsymbol{x})$ と表す．DNN 音響モデルの更新における backpropagation は，[7] と同様に行われる．

2.2　音声変換方式

本稿では，(1) 特徴量変換に基づく音声変換と (2) 差分スペクトル推定に基づく音声変換を用いる．

特徴量変換に基づく音声変換 (図 3) [5] では，2.1 節の方法で推定したスペクトル特徴量系列 $\hat{\boldsymbol{y}}$ と，別途推定した音源特徴量系列を用いる．生成した混合励振源波形に対して $\hat{\boldsymbol{y}}$ を用いた合成フィルタを適用することで，最終的な合成音声を得る．一方，差分スペクトル推定に基づく音声変換 (図 4) [13] では，入力音声波形に対して差分スペクトル特徴量系列を用いた合成フィルタを適用することで，最終的な合成音声を得る．本稿では，$\hat{\boldsymbol{y}} - \boldsymbol{x}$ を差分スペクトル特徴量系列として利用する．(1) は，ボコーダ処理による音質劣化を含むが，柔軟な音源特徴量変換を可能にする方法であり，(2) は，音源特徴量変換を困難にするが，ボコーダ処理による音質劣化を回避できる方法である．

3.　敵対的 DNN 音声変換

本稿で提案する，anti-spoofing に敵対する DNN 音声変換のための音響モデル学習について論述する．また，DNN 音声変換を実現するための DNN アーキテクチャについても述べる．

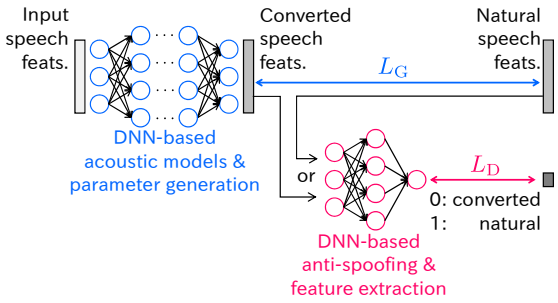

図 5　Anti-spoofing に敵対する DNN 音響モデル学習

Fig. 5　Acoustic model training to deceive anti-spoofing verification.

3.1　損失関数の定式化

3.1.1　Anti-spoofing 学習の損失関数

Anti-spoofing [14] では，自然音声と合成音声を識別する識別器を学習する．DNN に基づく anti-spoofing (例えば [15]) では，入力として与えられた音声特徴量に対して素性関数 $\boldsymbol{\phi}(\cdot)$ を適用した後に，当該音声特徴量が自然音声である事後確率 $D(\boldsymbol{\phi}(\cdot))$ を出力する．本稿では，素性関数を $\boldsymbol{\phi}(\boldsymbol{y}_t) = \boldsymbol{y}_t$ と定義し，各時刻 t における音声特徴量を識別に用いる．学習時に最小化される損失関数 $L_{\mathrm{D}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は，次式で与えられる．

$$L_{\mathrm{D}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{D},1}(\boldsymbol{y}) + L_{\mathrm{D},0}(\hat{\boldsymbol{y}}) \tag{2}$$

ここで，$L_{\mathrm{D},1}(\boldsymbol{y})$ 及び $L_{\mathrm{D},0}(\hat{\boldsymbol{y}})$ は自然音声，合成音声に対する損失であり，それぞれ次式で計算される．

$$L_{\mathrm{D},1}(\boldsymbol{y}) = -\frac{1}{T}\sum_{t=1}^{T} \log D(\boldsymbol{y}_t) \tag{3}$$

$$L_{\mathrm{D},0}(\hat{\boldsymbol{y}}) = -\frac{1}{T}\sum_{t=1}^{T} \log\left(1 - D(\hat{\boldsymbol{y}}_t)\right) \tag{4}$$

3.1.2　音響モデル学習の損失関数

提案アルゴリズムの枠組みを図 5 に示す．提案アルゴリズムでは，次式の損失関数 $L(\boldsymbol{y}, \hat{\boldsymbol{y}})$ を最小化するように音響モデルを更新する．

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}} \frac{E_{L_{\mathrm{G}}}}{E_{L_{\mathrm{D}}}} L_{\mathrm{D},1}(\hat{\boldsymbol{y}}) \tag{5}$$

$L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ は合成音声を自然音声と識別させるための損失であり，自然音声の特徴量と合成音声の特徴量が従う確率分布間の距離を最小化させる [16]．故に，提案アルゴリズムにおける音響モデルの損失関数は，生成誤差を最小化させ，かつ，合成音声の特徴量が従う確率分布を自然音声の特徴量が従う確率分布と等しくさせる効果を持つ．$E_{L_{\mathrm{G}}}$ と $E_{L_{\mathrm{D}}}$ はそれぞれ $L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ と $L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ の期待値を表す．式 (5) の第 2 項にこれらの比の値をかけることで，$L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ と $L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ のスケールを調整する．また，ω_{D} は anti-spoofing の損失に対する重みを表す．$\omega_{\mathrm{D}} = 0$ のときに，この損失関数は従来の音響モデル学習と等価になり，$\omega_{\mathrm{D}} = 1$ のときに，$L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ と $L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ は等重みをもつ．

音響モデルの学習後には anti-spoofing を再学習し，以降，これらの処理を反復して最終的な音響モデルを構築する．音響モデルは，通常の MGE 学習により初期化する．

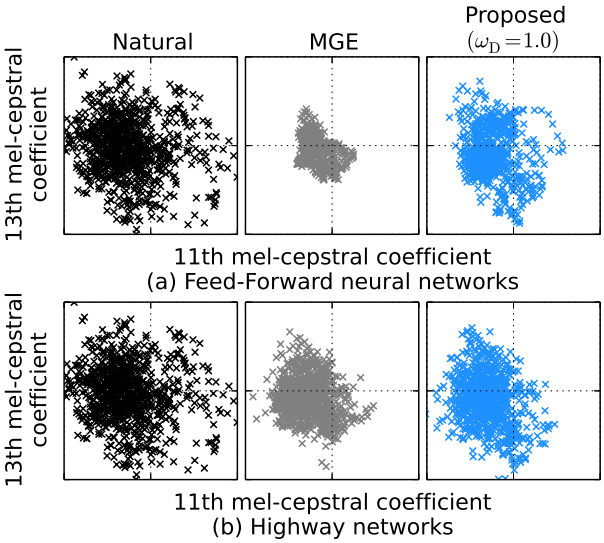

図 6　音声特徴量の散布図

Fig. 6　Scatter plots of speech parameters.

3.2　DNN 音声変換のための DNN アーキテクチャ

本稿では，(1) Feed-Forward 型ネットワークを用いた特徴量変換 (図 1) に加え，本稿で新たに提案する (2) 入出力ユニット間を結ぶ highway network [8–10] を用いた差分スペクトル推定 (図 2) を用いて提案アルゴリズムを評価する．(1) では，Feed-Forward neural network を含むモジュール $\boldsymbol{G}(\cdot)$ により，$\hat{\boldsymbol{y}} = \boldsymbol{G}(\boldsymbol{x})$ としてスペクトル特徴量系列を推定する．(2) では，次式に基づいて $\hat{\boldsymbol{y}}$ を推定する．

$$\hat{\boldsymbol{y}} = \boldsymbol{x} + \boldsymbol{T}(\boldsymbol{x}) \odot \boldsymbol{G}(\boldsymbol{x}) \tag{6}$$

ここで，$\odot$ はベクトルの要素積を表す演算子である．$\boldsymbol{T}(\boldsymbol{x})$ は，highway network の transform gate であり，本稿では Feed-Forward 型ネットワークで記述される．$\boldsymbol{T}(\boldsymbol{x})$ の各要素は 0 から 1 の値をとり，$\boldsymbol{x}$ の各時刻・各特徴量次元毎に $\boldsymbol{G}(\boldsymbol{x})$ を重み付けして出力する役割を持つ．出力音声波形生成時には，差分スペクトル特徴量系列 $\boldsymbol{T}(\boldsymbol{x}) \odot \boldsymbol{G}(\boldsymbol{x})$ を用いて入力波形をフィルタリングする．$\boldsymbol{G}(\cdot)$ と $\boldsymbol{T}(\cdot)$ はそれぞれ差分スペクトル推定器及び差分スペクトル重みに相当するため，$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{0}$ の場合，入力音声波形は変換されずに出力され，$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{1}$ の場合，residual network [17] や明示的な差分推定 [18] に一致する．

3.3　考　　察

3.3.1　学習アルゴリズムに関する考察

各 DNN アーキテクチャにおける学習アルゴリズムの影響を考察するために，生成されたスペクトル特徴量 (メルケプストラム係数) の散布図を図 6 に示す．Feed-Forward 型ネットワーク (図 6(a)) を用いる場合，MGE 学習後の音響モデルから生成された特徴量の分布は，自然音声特徴量の分布と比較して明らかに縮小するが，提案アルゴリズムによる分布は，その縮小をある程度緩和していることが確認できる．一方，highway network (図 6(b)) を用いる場合，入力音声特徴量を直接的に

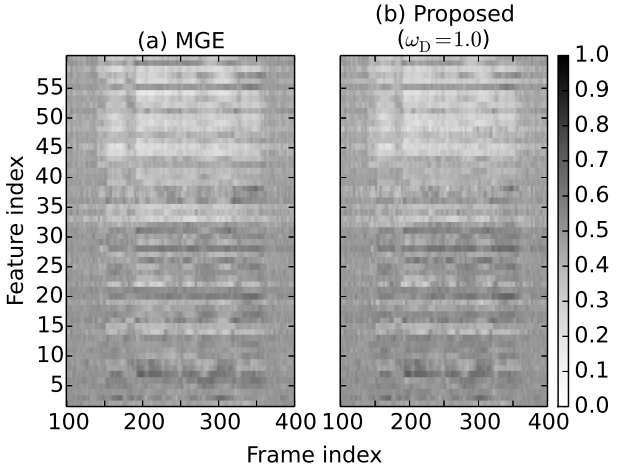

図 7　Transform gate の値の例．各値が大きいほど，入力音声特徴量は差分スペクトル推定により大きく変形される．

Fig. 7　An example of activation of transform gates. The higher activation means that the input speech parameters are strongly transformed by spectral differentials estimation.

利用できるため，MGE 学習においても図 6(a) のような縮小は見られない．しかしながらその分布は，入出力話者間の差異により，出力話者の自然音声特徴量の分布とは異なる．提案アルゴリズムはこの分布差異も補償することが，図 6(b) 右から確認できる．

3.3.2　Highway network に関する考察

話者間のスペクトル特徴量の差異は，話者対のみならず，周波数帯域や音韻に強く依存する．例えば，フォルマント構造は男女間で大きく異なるが，単一性別においては低域周波数の話者間分散は小さい．一方で，低域周波数帯域の音韻間分散 (話者内分散) は大きいことが知られている [19]．故に，音声変換における音響モデルは，入出力特徴量間の差異が小さい場合に過度の変換を避け (例えば周波数伸縮法 [20])，差異が大きい場合に柔軟な変換を実現するもの (例えば DNN) が望ましい．

図 2 に示す highway network は，$\boldsymbol{G}(\cdot)$ に学習・推定させる差分スペクトルの重みを $\boldsymbol{x}$ に応じて変化させる構造であると解釈できる．各時刻・各特徴量次元における $\boldsymbol{T}(\boldsymbol{x})$ の値の例を図 7 に示す．ここでは，音声のスペクトル構造を保持するため，4.1 節で述べる特徴量標準化は行っていない．図 7(a)(b) から，話者変換において支配的な特徴量である低次ケプストラムは $\boldsymbol{G}(\cdot)$ によって大きく変形され，比較的影響の小さい高次ケプストラムはほぼ変形されないことが確認できる．また，$\boldsymbol{T}(\boldsymbol{x})$ に対する学習アルゴリズムの影響に着目すると，提案アルゴリズム (図 7(b)) と MGE 学習 (図 7(a)) の間には顕著な差が生じていないことも確認できる．

本アーキテクチャは，音声強調における適応 soft-masking フィルタ [21] と類似した枠組みであるため，音声変換と音声強調間の知見共有を可能にすることが期待される．

4.　実験的評価

4.1　実験条件

実験的評価に用いるデータとして，ATR 音素バランス 503 文 [22] を利用し，A–I セット 450 文を学習に，J セット 53 文を評価に用いる．入力話者と出力話者は，どちらも男性とする．学習データのサンプリング周波数は 16 kHz，フレームシフトは 5 ms とする．スペクトル特徴量として STRAIGHT 分析 [23] による 0 次から 59 次のメルケプストラム係数，音源特徴量として F_0，5 周波数帯域における平均非周期成分 [24, 25] を用いる．スペクトル特徴量に対する前処理として，50 Hz のカットオフ変調周波数による trajectory smoothing [26, 27] を利用する．変換音声のメルケプストラム係数の 0 次の成分は，入力音声のものをそのまま使用する．音声変換の音響モデルと anti-spoofing のための DNN は, Feed-Forward 型とする．音声変換の音響モデルの隠れ層数は 3，隠れ層の素子数は 512，隠れ層・出力層の活性化関数は，それぞれ ReLU [28] 及び線形関数である．anti-spoofing の隠れ層数は 3，隠れ層の素子数は 256，隠れ層・出力層の活性化関数は，それぞれ ReLU 及び sigmoid 関数である．anti-spoofing はスペクトル特徴量 (59 次元)，音声変換の音響モデルはその静的・動的特徴量 (118 次元) のみをそれぞれ DNN の入力・出力特徴量として扱う．DNN の学習時には，スペクトル特徴量を各次元ごとに平均 0，分散 1 に標準化する．Highway network の transform gate は 2 層の Feed-Forward 型ネットワークであり, 差分スペクトル重みを sigmoid 関数により計算する．DNN の最適化手法として，学習率 0.01 の AdaGrad [29] を用いる．F_0，非周期成分については，自然音声の特徴量を使用する．

まず，$\omega_D = 0.0$ として，反復回数 25 回の MGE 学習により音響モデルを初期化する．次に，$\omega_D = 1.0$ として，anti-spoofing 学習及び，提案アルゴリズムによる音響モデル学習を交互に実施する．この際の反復回数は 25 回とする．提案アルゴリズムにおける期待値 E_{L_G} と E_{L_D} は，反復毎に，その時点における anti-spoofing と音響モデルを用いて計算する．

提案アルゴリズムによる音質改善効果を確認するため，DNN 音声変換における音質に関するプリファレンス AB テスト，及び，話者性に関する XAB テストを実施する．評価する手法は従来の MGE 学習と提案法であり，被験者数は各評価に対して 8 名である．

4.2　評価結果

各主観評価結果を図 8 及び図 9 に示す．Feed-Forward 型ネットワークを用いた音声変換に関する結果 (図 8) において，提案アルゴリズムによる音質及び話者性の大幅な改善が確認できる．同様に，highway network を用いた音声変換に関する結果 (図 9) において，改善度は減少しているものの提案法による改善が確認できる．この改善度の減少は，図 6(b) に示すように，入力音声特徴量の直接的な利用により過剰な平滑化が緩和され，MGE 学習後の合成音声も音質が改善されたためであると思われる．

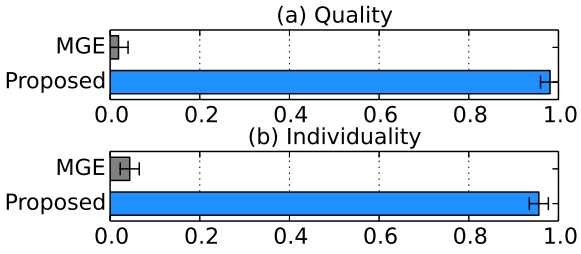

図 8　主観評価結果 (Feed-Forward 型ネットワーク，エラーバーは 95%信頼区間)

Fig. 8　Results of subjective evaluations with 95% confidence intervals (Feed-Forward neural networks).

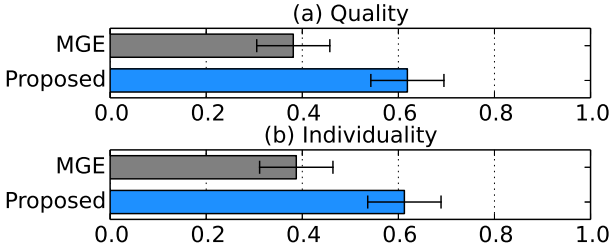

図 9　主観評価結果 (highway network，エラーバーは 95%信頼区間)

Fig. 9　Results of subjective evaluations with 95% intervals (highway networks).

5. おわりに

本稿では，統計的パラメトリック音声変換の音質改善を目的として，anti-spoofing に敵対する DNN 音声変換の学習アルゴリズムを提案した．また，DNN 音声変換を実現するための DNN アーキテクチャとして，Feed-Forward 型ネットワークを用いた特徴量変換法に基づく音声変換と，本稿で新たに提案した，highway network を用いた差分スペクトル法に基づく音声変換を採用し，実験的評価により両方の変換方式において提案アルゴリズムが有効であることを示した．今後は，本稿で提案した入出力間の highway network を用いた多人数話者変換について検討する．

文　献

[1] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] G. Hinton, L. Deng, D. Yu, G. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine of IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[3] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.

[4] K. Tokuda, Y Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[5] T. Toda, A. W. 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.

[6] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, 2016.

[7] 齋藤佑樹, 高道慎之介, and 猿渡洋, "DNN 音声合成のための Anti-spoofing を考慮した学習アルゴリズム," **日本音響学会** *2016* **年秋季研究発表会講演論文集**, pp. 149–150, Sep. 2016.

[8] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," in *Proc. ICML Deep Learning Workshop*, France, Jul. 2015.

[9] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. NIPS*, 2015, pp. 2377–2385.

[10] X. Wang, S. Takaki, and J. Yamagishi, "Investigating very deep highway networks for parametric speech synthesis," in *Proc. 9th ISCA Speech Synthesis Workshop*, California, U.S.A., Sep. 2016.

[11] Z. Wu and S. King, "Improving trajectory modeling for DNN-based speech synthesis by using stacked bottleneck features and minimum trajectory error training," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1255–1265, 2016.

[12] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 1315–1318.

[13] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Proc. INTERSPEECH*, Max Atria, Singapore, Sep. 2014, pp. 2514–2518.

[14] Z. Wu, P. L. D. Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768–783, 2016.

[15] N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu, "Robust deep feature for spoofing detection — the SJTU system for ASVspoof 2015 Challenge," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2097–2101.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, U.S.A., June 2016, pp. 770–778.

[18] 金子卓弘, 亀岡弘和, 北条伸克, 井島勇祐, 平松薫, and 柏野邦夫, "統計的パラメトリック音声合成のための敵対的学習に基づくポストフィルタリング," **電子情報通信学会技術研究報告**, vol. SP2016-12, pp. 89–94, Dec. 2016.

[19] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," *Journal of the Acoustical Society of Japan (E)*, vol. 16, no. 5, pp. 283–289, 1995.

[20] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[21] J. Hout and A. Alwan, "A novel approach to soft-mask estmation and log-spectral enhancement for robust speech recognition," in *Proc. ICASSP*, Kyoto, Japan, Mar 2012,

pp. 4105–4108.

[22] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, "ATR technical repoart," , no. TR-I-0166M, 1990.

[23] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[24] H. Kawahara, Jo Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *MAVEBA 2001*, Firentze, Italy, Sep. 2001, pp. 1–6.

[25] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, Pittsburgh, U.S.A., Sep. 2006, pp. 2266–2269.

[26] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura, "The NAIST text-to-speech system for the Blizzard Challenge 2015," in *Proc. Blizzard Challenge workshop*, Berlin, Germany, Sep. 2015.

[27] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016," in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2007.

[28] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, Lauderdale, U.S.A., Apr. 2011, pp. 315–323.

[29] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.