Highway network を用いた差分スペクトル法に基づく敵対的 DNN 音声変換 *

◎齋藤 佑樹，高道 慎之介，猿渡 洋（東大院・情報理工）

1 はじめに

統計的パラメトリック音声合成により生成される合成音声の音質は，声のなりすましを防ぐ anti-spoofing に敵対することで改善される．これまでに我々は，anti-spoofing を詐称するように音響モデルを学習する Deep Neural Network（DNN）音声合成のための学習アルゴリズム（敵対的 DNN 音声合成）[1, 2] を提案し，テキスト音声合成への適用においてその音質改善効果を確認している．

本稿では，この敵対的 DNN 音声合成を音声変換に導入するとともに，入出力ユニット間を結ぶ highway network [3] を用いた差分スペクトル法に基づく音声変換を提案する．入出力音声間の差分スペクトル及びその差分重みは，anti-spoofing に敵対するように推定される．実験的評価より，提案アルゴリズムによる品質改善効果が得られることを示す．

2 従来の枠組み

2.1 Minimum Generation Error（MGE）学習

DNN 音声変換の音響モデル学習部では，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 MGE 学習 [4] の損失関数は，自然音声の特徴量系列 $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ と，最尤パラメータ生成 [5] 後の合成音声の特徴量系列 $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ の二乗誤差として次式で与えられる．

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

ここで，t はフレームインデックス，T はフレーム数である．

2.2 差分スペクトル法に基づく音声変換

差分スペクトル法に基づく音声変換 [6] 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．音源特徴量の変換は困難だが，ボコーダ処理による音質劣化を回避できる．本稿では，入力音声特徴量系列 $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_T^\top]^\top$ に対し，$\hat{\boldsymbol{y}} - \boldsymbol{x}$ を差分スペクトル特徴量系列として利用する．

3 敵対的 DNN 音声変換

3.1 損失関数の定式化

3.1.1 Anti-spoofing 学習の損失関数

Anti-spoofing [7] 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 [8]）では，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，各フレーム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．

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

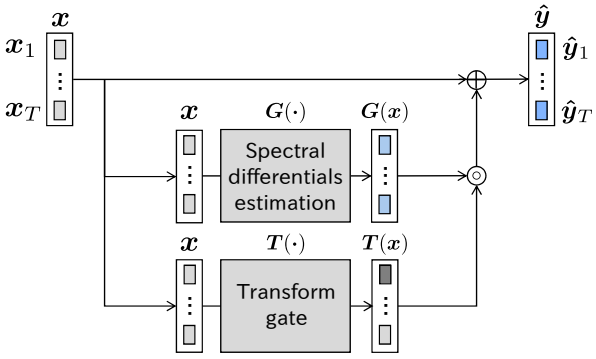

Fig. 1 Highway network を用いた音声変換

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

敵対的 DNN 音声変換では，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．

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

ここで，$L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ は合成音声を自然音声と識別させるための損失であり，$\boldsymbol{y}$ と $\hat{\boldsymbol{y}}$ の分布の差異を最小化する役割を持つ [9]．ω_{D} は，anti-spoofing の損失に対する重みである．また，$E_{L_{\mathrm{G}}}$ と $E_{L_{\mathrm{D}}}$ はそれぞれ $L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ と $L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ の期待値であり，$L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ と $L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ のスケールを調整する役割を持つ．

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．

3.2 Highway network を用いた差分スペクトル法

本稿で提案する，入出力間の highway network を用いた差分スペクトル法（Fig. 1）では，次式に基づいて $\hat{\boldsymbol{y}}$ を推定する．

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

ここで，$\circ$ はアダマール積である．$\boldsymbol{T}(\cdot)$ と $\boldsymbol{G}(\cdot)$ は highway network の transform gate 及び差分スペクトル推定器であり，$\boldsymbol{T}(\cdot)$ は Feed-Forward neural network を用いて記述される．$\boldsymbol{T}(\boldsymbol{x})$ 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 $\boldsymbol{T}(\boldsymbol{x}) \circ \boldsymbol{G}(\boldsymbol{x})$ を用いて入力音声波形をフィルタリングする．$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{0}$ 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，提案アーキテクチャは residual network [10] もしくは明示的な差分推定 [11] に一致する．

---

* Adversarial DNN-Based Voice Conversion Based on Spectral Differentials Using Highway Networks, by SAITO, Yuki, TAKAMICHI, Shinnosuke, and SARUWATARI, Hiroshi (The University of Tokyo).

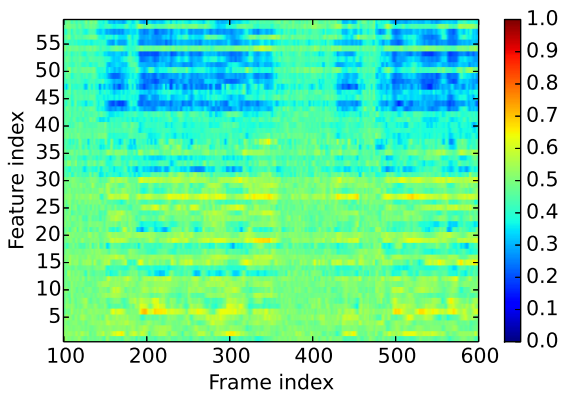

Fig. 2　Transform gate の値の例

3.3　Highway network に関する考察

提案アーキテクチャは，入力音声特徴量 $\boldsymbol{x}$ に応じて，$\boldsymbol{G}(\cdot)$ に学習・推定させる差分スペクトルの重みを各時刻及び各特徴量次元毎に変化させる構造であり，音声強調におけるソフトマスク推定 [12] と類似した枠組みであると解釈できる．故に，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．

$\boldsymbol{T}(\boldsymbol{x})$ の値の例を Fig. 2 に示す．Figure 2 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，ランダム性の大き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．

4　実験的評価

4.1　実験条件

実験に用いるデータとして，ATR 音素バランス 503 文 [13] 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．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 [14] 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 [15] を用いる．変換音声のメルケプストラム係数の 0 次成分は，入力音声のものを使用する．DNN 学習時には，スペクトル特徴量を平均 0，分散 1 に正規化する．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，隠れ層及び出力層の活性化関数はそれぞれ Rectified Linear Unit（ReLU）[16] 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，隠れ層及び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．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．最適化アルゴリズムとして，学習率 0.01 の AdaGrad [17] を用いる．

まず，$\omega_{\mathrm{D}} = 0.0$ 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，$\omega_{\mathrm{D}} = 1.0$ 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 $E_{L_{\mathrm{G}}}$ と $E_{L_{\mathrm{D}}}$ 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．被験者数は各評価に対して 8 名である．

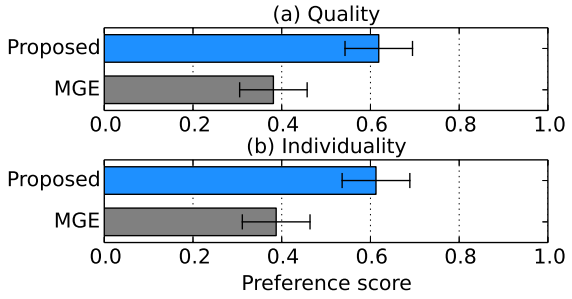

Fig. 3　主観評価結果（エラーバーは 95%信頼区間）

4.2　評価結果

主観評価結果を Fig. 3 に示す．提案アルゴリズムは従来の MGE 学習よりも高いスコアを獲得しているため，提案アルゴリズムによる音質及び話者性の改善が確認された．

5　おわりに

本稿では，DNN 音声変換のための anti-spoofing を考慮した学習アルゴリズムと，highway network を用いた差分スペクトル法による音声変換を提案し，実験的評価によりその有効性を示した．今後は，highway network を用いた際の話者対への依存性について調査する．

謝辞　本研究は，総合科学技術･イノベーション会議による革新的研究開発推進プログラム（ImPACT），セコム科学技術振興財団，及び JSPS 科研費 16H06681 の支援を受けた．

参考文献

[1] 齋藤 他, 音講論（秋）, 3-5-1, 2016.
[2] Saito et al., *Proc. ICASSP*, 2017. (to appear)
[3] Srivastava et al., *Proc. ICML Deep Learning Workshop*, 2015.
[4] Wu et al., *IEEE/ACM Trans. on ASLP*, Vol. 24, No. 7, pp. 1255–1265, 2016.
[5] Tokuda et al., *Proc. ICASSP*, pp. 1315–1318, 2000.
[6] Kobayashi et al., *Proc.INTERSPEECH*, pp. 2514–2518, 2014.
[7] Wu et al., *IEEE/ACM Trans. on ASLP*, Vol. 24, No. 4, pp. 768–783, 2016.
[8] Chen et al., *Proc. INTERSPEECH*, pp. 2097–2101, 2015.
[9] Goodfellow et al., *Proc. NIPS*, pp. 2672–2680, 2014.
[10] He et al., *Proc. CVPR*, pp. 770–778, 2016.
[11] 金子 他, 信学技報, SP2016-12, pp. 89–94, 2016.
[12] Hout et al., *Proc. ICASSP*, pp. 4105–4108, 2012.
[13] 阿部 他, ATR テクニカルレポート, TR-I-0166, 1990.
[14] Kawahara et al., *Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207, 1999.
[15] Ohtani et al., *Proc. INTERSPEECH*, pp. 2266–2269, 2006.
[16] Glorot et al., *Proc. AISTATS*, pp. 315–323, 2011.
[17] Duchi et al., *The Journal of Machine Learning Research*, Vol. 12, pp. 2121–2159, 2011.