DNNテキスト音声合成のための Anti-spoofingに敵対する学習アルゴリズム

齋藤 佑樹[1,a)]　高道 慎之介[1,b)]　猿渡 洋[1,c)]

概要：統計的パラメトリック音声合成方式では，生成される合成音声特徴量系列の過剰な平滑化による音質劣化が問題となる．これまでに，系列内変動や変調スペクトルなどの，自然音声と合成音声を識別できる解析的特徴量を補償する手法が提案され，その音質改善効果が確認されている．本稿では，この枠組みを拡張し，合成音声のさらなる音質改善を目的として，合成音声による声のなりすましを防ぐ anti-spoofing に敵対する deep neural network 音響モデルの学習アルゴリズムを提案する．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．実験的評価により，(1) 提案アルゴリズムによる音質改善効果が得られること，(2) 提案アルゴリズムはハイパーパラメータの設定に対して比較的頑健に動作することを示す．

キーワード：統計的パラメトリック音声合成，DNN 音声合成，anti-spoofing，学習アルゴリズム，敵対的学習，過剰な平滑化

Training Algorithm to Deceive Anti-spoofing Verification for DNN-based Text-To-Speech Synthesis

Yuki Saito[1,a)]　Shinnosuke Takamichi[1,b)]　Hiroshi Saruwatari[1,c)]

Abstract: This paper proposes a novel training algorithm for high-quality deep neural network-based speech synthesis. The parameters of synthetic speech tend to be over-smoothed, and this causes significant quality degradation in synthetic speech. The proposed algorithm takes into account an Anti-Spoofing Verification (ASV) as an additional constraint in the acoustic model training. The ASV is a discriminator trained to distinguish natural and synthetic speech. Since acoustic models for speech synthesis are trained so that the ASV recognizes the synthetic speech parameters as natural speech, the synthetic speech parameters are distributed in the same manner as natural speech parameters. The experimental results demonstrate that 1) the algorithm outperforms the conventional training algorithm in terms of speech quality, and 2) it is robust against the hyper-parameter settings.

Keywords: statistical parametric speech synthesis, DNN-based speech synthesis, anti-spoofing verification, training algorithm, generative adversarial training, over-smoothing effect

1. はじめに

統計的パラメトリック音声合成 [1] は，人間のように自然な合成音声を生成するための技術の一つである．入力情報から音声特徴量を生成する音響モデルは hidden Markov model [2] や Deep Neural Network（DNN） [3] により構築され，尤度最大化 [2] や二乗誤差最小化 [4], [5] などの規範に基づいて学習される．高品質な合成音声を生成するための音響モデル学習の技術は，テキスト音声合成，音声変換，そしてマルチモーダル音声合成で共有できるため，広く研究されている．しかしながら，この音響モデルから生

[1] 東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan.
[a)] yuuki_saito@ipc.i.u-tokyo.ac.jp
[b)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
[c)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

成された音声特徴量は，自然音声と比較して過剰に平滑化される傾向にあり [1], [6]，合成音声の音質は自然音声と比較すると未だに低い．

合成音質の音質を改善させる方法として，自然音声と合成音声の違いを補償することが考えられる．音声特徴量の分布は自然音声と合成音声で異なる [7] ため，合成音声特徴量の分布を自然音声に近づけることで合成音声の音質改善が期待できる．これを実現するための例として，音声特徴量の従う確率分布をパラメトリック [8] またはノンパラメトリック [9] にモデル化し，合成音声の特徴量を生成または変形する手法がある．さらに有効な手法として，合成音声の音質劣化に関連する解析的特徴量を用いる手法が考えられる．代表的な例として，系列内変動（Global Variance: GV） [8] や変調スペクトル（Modulation Spectrum: MS） [10] がある．これらの解析的特徴量は，音響モデル学習時もしくは音声特徴量生成時の制約として機能する [11], [12]．能勢ら [13] や高道ら [11] によって，生成された GV または MS の正規分布を自然音声の GV または MS の正規分布に近づける方法が提案されているが，合成音声の音質劣化は未だに深刻な問題となっている．

この音質劣化問題に対して，本研究では，合成音声による声のなりすましを防ぐ anti-spoofing を用いた統計的パラメトリック音声合成の音響モデル学習アルゴリズムを提案・開発している [14], [15]．Anti-spoofing は合成音声と自然音声を識別するように学習される識別器である．音声合成の音響モデルは anti-spoofing に敵対するように学習されるため，合成音声特徴量の分布は自然音声特徴量の分布に近くなる．提案アルゴリズムにおける音響モデル学習の規範は，従来の学習規範と anti-spoofing に敵対するための規範の重み付き和として表現され，音響モデルと anti-spoofing の両方を DNN で記述することで，この学習は back-propagation により簡潔に行われる．さらに，提案アルゴリズムは従来の GV や MS などの補償手法の拡張としてみなされ，GV や MS などの解析的特徴量だけでなく，DNN により自動的に設計された特徴量も補償できる．また，DNN を用いた anti-spoofing により，自然音声特徴量の従う分布として，従来のガウス分布よりも複雑な分布を仮定できる．本稿では，提案アルゴリズムの挙動を詳細に調査する実験的評価により，(1) 従来の学習規範を用いた音響モデル学習と比較して，提案アルゴリズムによる音質改善効果が得られること，また，(2) 提案アルゴリズムはハイパーパラメータの設定に対して比較的頑健に動作することを示す．

2. 従来の学習アルゴリズム

2.1 音響モデルとしての DNN

DNN 音声合成 [16] において，テキスト特徴量と音声特徴量の対応関係を表す音響モデルは階層的なネットワークにより表現される．音響モデルの学習では，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．以降，自然音声の特徴量系列を $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$，合成音声の特徴量系列を $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ と表記する．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．ここでは，[16] と同様に，音響モデルは各フレーム t における音声特徴量の静的・動的特徴量 $\hat{\boldsymbol{Y}}_t = [\hat{\boldsymbol{y}}_t^\top, \Delta\hat{\boldsymbol{y}}_t^\top, \Delta\Delta\hat{\boldsymbol{y}}_t^\top]^\top$ を予測する．

2.2 Minimum Generation Error（MGE）学習 [5]

従来の DNN 音声合成において，音響モデルは MGE 学習アルゴリズム [5] により学習される．MGE 学習の損失関数 $L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は，$\boldsymbol{y}$ と $\hat{\boldsymbol{y}}$ の二乗誤差として次式で計算される．

$$\begin{aligned} L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \frac{1}{T}(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top (\hat{\boldsymbol{y}} - \boldsymbol{y}) \\ &= \frac{1}{T}\left(\boldsymbol{R}\hat{\boldsymbol{Y}} - \boldsymbol{y}\right)^\top \left(\boldsymbol{R}\hat{\boldsymbol{Y}} - \boldsymbol{y}\right) \qquad (1) \\ \boldsymbol{R} &= \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \qquad (2) \end{aligned}$$

ここで，$\hat{\boldsymbol{Y}} = [\hat{\boldsymbol{Y}}_1^\top, \cdots, \hat{\boldsymbol{Y}}_t^\top, \cdots, \hat{\boldsymbol{Y}}_T^\top]^\top$ は音響モデルにより予測された合成音声特徴量の静的・動的特徴量系列である．$\boldsymbol{W}$ は静的・動的特徴量系列を計算するための行列 [2] であり，$\boldsymbol{\Sigma} = \mathrm{diag}\left[\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_t, \cdots, \boldsymbol{\Sigma}_T\right]$ は学習データを用いて別途推定される共分散行列である．ここで，$\boldsymbol{\Sigma}_t$ はフレーム t における音声特徴量の共分散行列である．

DNN のモデルパラメータ（ネットワークの結合重み及びバイアス項）は，損失関数の $\boldsymbol{Y}$ における勾配を用いた back-propagation により更新される．

3. 提案する学習アルゴリズム

3.1 Anti-spoofing [17]

Anti-spoofing では，合成音声による声のなりすましを防ぐために，当該音声特徴量（もしくは音声波形）を用いて自然音声と合成音声を識別する．DNN に基づく anti-spoofing（例えば [18]）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直接的に識別に用いる．Anti-spoofing 学習の損失関数 $L_{\mathrm{D}}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は，次式に示す cross-entropy 関数として与えられる．

$$L_{\mathrm{D}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{D},1}(\boldsymbol{y}) + L_{\mathrm{D},0}(\hat{\boldsymbol{y}}) \qquad (3)$$

$$L_{\mathrm{D},1}(\boldsymbol{y}) = -\frac{1}{T}\sum_{t=1}^{T} \log D(\boldsymbol{y}_t) \qquad (4)$$

$$L_{\mathrm{D},0}(\hat{\boldsymbol{y}}) = -\frac{1}{T}\sum_{t=1}^{T} \log\left(1 - D(\hat{\boldsymbol{y}}_t)\right) \qquad (5)$$

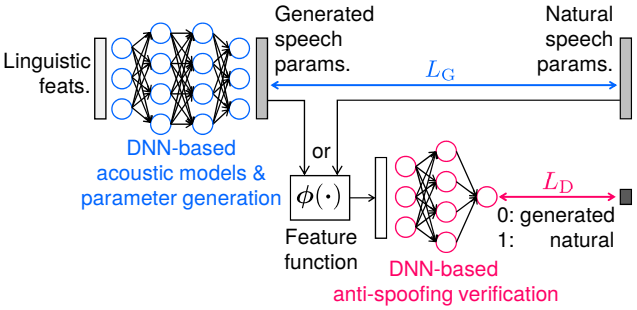

図 1　提案アルゴリズムにおける損失関数の計算手順
Fig. 1　Calculation of loss function in proposed algorithm.

ここで，$L_{D,1}(\boldsymbol{y})$ と $L_{D,0}(\hat{\boldsymbol{y}})$ はそれぞれ自然音声と合成音声に対する損失である．学習時には，back-propagation により，自然音声に対して 1 を，合成音声に対して 0 を出力するように anti-spoofing のモデルパラメータを更新する．

3.2　Anti-spoofing に敵対する音響モデル学習

Anti-spoofing に敵対する音響モデルの学習アルゴリズムを提案する．提案アルゴリズムにおける損失関数の計算手順を図 1 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．

$$L\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = L_{\mathrm{G}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) + \omega_{\mathrm{D}} \frac{E_{L_{\mathrm{G}}}}{E_{L_{\mathrm{D}}}} L_{\mathrm{D},1}\left(\hat{\boldsymbol{y}}\right) \qquad (6)$$

ここで，$E_{L_{\mathrm{G}}}$ 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 (6) 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 MGE 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．$L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ は anti-spoofing に合成音声特徴量を自然音声と識別させるための損失であり，自然音声特徴量と合成音声特徴量の従う確率分布間の距離（厳密には Jensen-Shannon divergence）を最小化させる [19]．故に，提案アルゴリズムにおける損失関数は，生成誤差を最小化させ，かつ，合成音声特徴量の従う確率分布を自然音声特徴量の従う確率分布と等しくさせる効果を持つ．

Anti-spoofing に敵対するように音響モデルを学習した後には anti-spoofing を再学習し，以降，これらの処理を繰り返して最終的な音響モデルを構築する．

3.3　他手法との関連及び提案アルゴリズムの分析

提案アルゴリズムでは, 素性関数 $\phi(\cdot)$ として，GV や MS などの既知の解析的特徴量のみならず, DNN により自動設計された特徴量（例えば [20]）も利用可能である．また，提案アルゴリズムにおける音響モデルと anti-spoofing の学習は back-propagation の枠組みで完結するため，任意の DNN アーキテクチャ（例えば long-short-term memory [21]）を

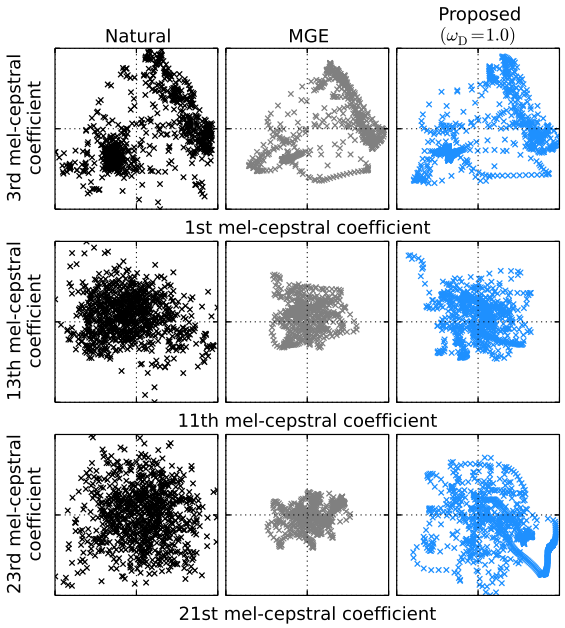

図 2　メルケプストラム係数の各次元の散布図．左から，自然音声，従来の MGE 学習，そして提案アルゴリズムである．これらの値は，評価データの一文から抽出したものである．
Fig. 2　Scatter plots of mel-cepstral coefficients with several pairs of dimensions. From the left, the figures correspond to natural speech, the conventional MGE algorithm, and the proposed algorithm, respectively. These mel-cepstral coefficients were extracted from one utterance of the evaluation data.

利用できる．ただし，back-propagation による学習を行うため，解析的特徴量を利用する場合，素性関数 $\phi(\cdot)$ は微分可能である必要がある．

提案アルゴリズムにおける損失関数（式 (6)）は，敵対的学習 [19] と，識別器を含むマルチタスク学習 [22] の組合せとみなすことができる．$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{D},1}(\hat{\boldsymbol{y}})$ と定義すると，式 (6) の損失関数は敵対的学習と一致する [19]．すなわち，提案アルゴリズムの学習は，所望の入出力間対応関係を持った敵対的学習 [23] とみなすことができる．

3.2 節で述べたように，提案アルゴリズムによる学習は，合成音声特徴量の分布を自然音声に近づける効果を持つ．DNN を用いた敵対的学習により，自然音声特徴量の従う確率分布として，従来の正規分布よりも複雑な分布を利用できる．自然音声と合成音声のメルケプストラム係数の散布図を図 2 に示す．この図より，従来の MGE 学習による音声特徴量の分布は縮小しているが，提案アルゴリズムによる音声特徴量の分布は自然音声と同様に広がりをもっていることが確認できる．さらに，より高次のメルケプストラムになるほど，提案アルゴリズムによる分布補償効果の影響が大きくなることも確認できる．

本稿では，敵対的学習による分布補償を音響モデル学習時に行う．故に，生成時には通常の生成処理を利用可能で

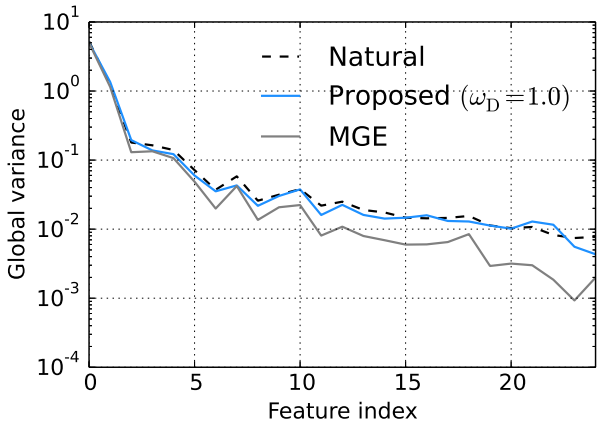

図 3　自然音声及び合成音声のメルケプストラム係数の平均 GV
Fig. 3　Averaged GVs of natural and generated mel-cepstral coefficients.

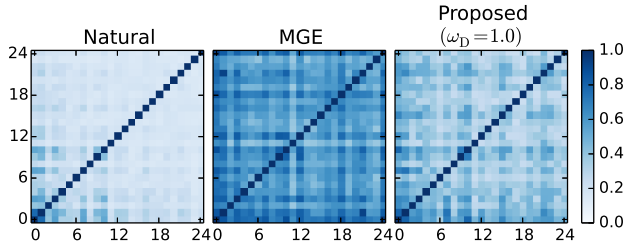

図 4　自然音声及び合成音声のメルケプストラム係数の MIC．MIC は 0.0 から 1.0 の間の値をとり，2 つの変量間に強い相関がある場合に 1.0 に近づく．左から，自然音声，従来の MGE 学習，そして提案アルゴリズムである．これらの値は，評価データの一文を用いて計算されたものである．
Fig. 4　MICs of natural and generated mel-cepstral coefficients. The MIC ranges from 0.0 to 1.0, and the two valuables with a strong correlation have a value closer to 1.0. From the left, the figures correspond to natural speech, the conventional MGE algorithm, and the proposed algorithm, respectively. These MICs were calculated from one utterance of the evaluation data.

あり，ポストフィルタによる分布補償 [11], [13], [24] を必要としない．

ここで，生成された音声特徴量の性質（例えば解析的特徴量や直感的理由 [25]）が提案アルゴリズムによりどう変化するかを分析する．自然音声特徴量と合成音声特徴量の平均 GV を図 3 に示す．この図より，従来の MGE 学習と比較して，提案アルゴリズムの音声特徴量の GV は，自然音声の GV に近づいていることが確認できる．次に，メルケプストラム係数の各次元の相関を定量化するために，Maximal Information Coefficient（MIC） [26] を計算した（図 4）．この図より，[7] で報告されているように，自然音声の特徴量は弱い相関を持つが，MGE 学習による合成音声の特徴量は強い相関を持つことが確認できる．一方，提案アルゴリズムによる合成音声の特徴量は比較的弱い相関を持つ．これらの結果から，提案アルゴリズムは GV のみならず，各次元の相関も補償することが分かる．

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 [27] 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 [28] による 0 次から 24 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 [29], [30] 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 [31] を利用する．コンテキストラベルは，音素，モーラ位置，アクセント型，音素内フレーム位置などから成る 274 次元ベクトルである．本稿ではスペクトル特徴量のみを予測するため，韻律に関連するコンテキストラベルを用いていない．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．また，学習データにおける無音フレームの 80%を除去する．

音声合成の音響モデルと anti-spoofing のための DNN は，Feed-Forward 型とする．音声合成の音響モデルの隠れ層数は 3，隠れ層の素子数は 400，隠れ層及び出力層の活性化関数は，それぞれ Rectified Linear Unit（ReLU）[32] 及び線形関数である．Anti-spoofing の隠れ層数は 2，隠れ層の素子数は 200，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．音声合成の音響モデルは各フレーム毎のスペクトル特徴量の静的・動的特徴量（75 次元）を出力し，anti-spoofing はその静的特徴量（25 次元）のみを入力とし，自然音声と合成音声を識別する．最適化アルゴリズムとして，学習率 0.01 の AdaGrad [33] を用いる．F_0，非周期成分，継続長は，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，ω_D を 0.0 以上に設定し，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．ただし，anti-spoofing は，自然音声特徴量と MGE 学習後の合成音声特徴量をある程度識別するように初期化する．この初期化の反復回数は 5 回とする．提案アルゴリズムにおける期待値 E_{L_G} 及び 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．

従来の学習アルゴリズムと提案アルゴリズムを比較するため，まず，客観評価として，合成音声特徴量の生成誤差（式 (1)）と，anti-spoofing における詐称率を計算する．詐称率は合成音声特徴量を自然音声と誤識別した割合を表す．ただし，詐称率を評価する anti-spoofing は，MGE 学習後の合成音声特徴量を用いて別途構築する．これらの客観評価指標は，提案アルゴリズムにおける重み ω_D を 0.0

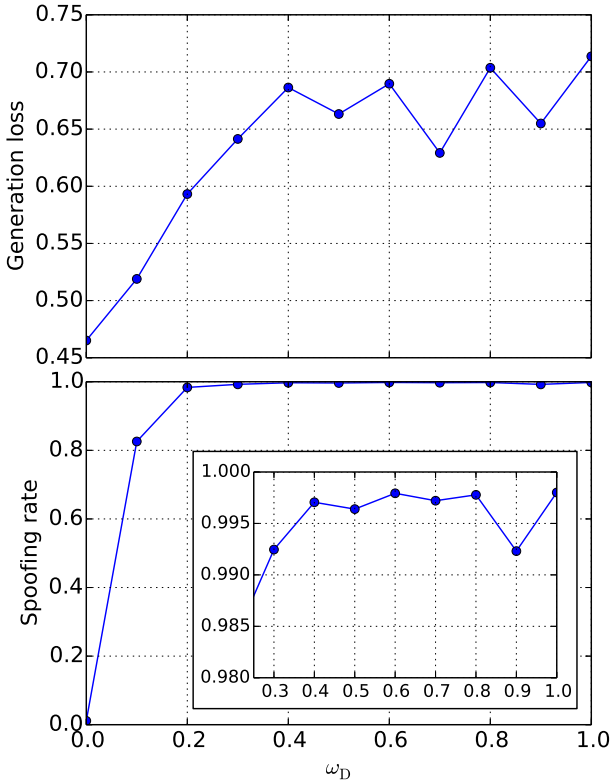

図 5　ω_D の変化に伴う生成誤差（上図）と詐称率（下図）の変化
Fig. 5　Parameter generation loss (above) and spoofing rate (below) for various ω_D.

から 1.0 の範囲で変化させて計算する．次に，提案アルゴリズムによる音質改善効果を確認するための主観評価を実施する．

4.2　客観評価結果

客観評価結果を図 5 に示す．この図より，ω_D が 0.0 から増加するにつれて，生成誤差も単調に増加することが確認できる．しかし，ω_D が 0.4 を超えると，そのような傾向は見られない．一方で，ω_D が 0.0 から 0.2 に増加するとき，詐称率は大幅に増加する．また，ω_D が 0.2 を超えると，詐称率はほぼ変化しないことも確認できる．これらの結果より，提案アルゴリズムを用いることで，生成誤差は悪化するが，anti-spoofing に敵対する特徴量を生成できることが示された．

4.3　主観評価結果

合成音声の音質を比較するために，8 名の評価者によるプリファレンス AB テストを実施する．本稿では以下の手法を比較する．

MGE:　従来の MGE 学習 [5]
Proposed（$\omega_D = 0.3$）:　詐称率が 0.99 以上
Proposed（$\omega_D = 1.0$）:　ω_D の標準設定

主観評価結果を図 6 に示す．提案アルゴリズム（Proposed）は従来の MGE 学習（MGE）と比較して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改善効果が確認できる．さらに，提案アルゴリズム同士を比較すると，Proposed（$\omega_D = 0.3$）と Proposed（$\omega_D = 1.0$）の間には有意な差が見られない．図 5 において，ω_D が 0.3 から 1.0 をとる場合，客観評価値はほぼ変化しないことから，合成音声の音質もこの範囲でほぼ変化しないと予想される．これらの結果より，提案アルゴリズムはハイパーパラメータの設定に対して比較的頑健に動作することが示された．

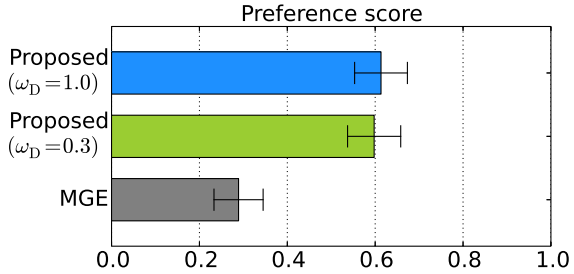

図 6　主観評価結果（エラーバーは 95%信頼区間）
Fig. 6　Preference scores of speech quality with 95% confidence intervals.

5.　おわりに

本稿では，高音質な音声を生成する統計的パラメトリック音声合成の手法として，anti-spoofing に敵対する学習アルゴリズムを提案し，実験的評価によりその有効性を示した．今後は，時間 [34] 及び言語依存 [23] の anti-spoofing について検討する．

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] H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
[2] 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.
[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] Y. J. Wu and R. H. Wang, “Minimum generation error training for HMM-based speech synthesis,” in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 89–92.
[5] 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.

[6] T. Toda, L. H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *Proc. INTERSPEECH*, California, U.S.A., Sep. 2016, pp. 1632–1636.

[7] Y. Ijima, T. Asami, and H. Mizuno, "Objective evaluation using association between dimensions within spectral features for statistical parametric speech synthesis," in *Proc. INTERSPEECH*, California, U.S.A., Sep. 2016, pp. 337–341.

[8] 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.

[9] Y. Ohtani, M. Tamura, M. Morita, T. Kagoshima, and M. Akamine, "Histogram-based spectral equalization for HMM-based speech synthesis using mel-LSP," in *Proc. INTERSPEECH*, Portland, U.S.A., Sep. 2012.

[10] 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.

[11] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4859–4863.

[12] K. Hashimoto, K. Oura, Y. Nankaku, and K.Tokuda, "Trajectory training considering global variance for speech synthesis based on neural networks," in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 5600–5604.

[13] T. Nose and A. Ito, "Analysis of spectral enhancement using global variance in HMM-based speech synthesis," in *Proc. INTERSPEECH*, MAX Atria, Singapore, May 2014, pp. 2917–2921.

[14] 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.

[15] Y. Saito, S. Takamichi, and H. Saruwatari, "Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis," in *Proc. ICASSP*, New Orleans, U.S.A., Mar. 2017.

[16] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 7962–7966.

[17] 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.

[18] 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.

[19] 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.

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

[21] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4470–4474.

[22] B. Huang, D. Ke, H. Zheng, B. Xu, Y. Xu, and K. Su, "Multi-task learning deep neural networks for speech feature denoising," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2464–2468.

[23] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text-to-image synthesis," in *Proc. ICML*, 2016, pp. 1060–1069.

[24] 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.

[25] T. R. Marco, S. Sameer, and G. Carlos, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proc. KDD*, San Francisco, U.S.A., Aug. 2016, pp. 1135–1164.

[26] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," vol. 334, no. 6062, pp. 1518–1524, 2011.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.