モーメントマッチングに基づく DNN合成歌声のランダム変調ポストフィルタと ニューラルダブルトラッキングへの応用

田丸 浩気[1,a)]　齋藤 佑樹[1]　高道 慎之介[1,b)]　郡山 知樹[2]　猿渡 洋[1]

概要：Deep neural network (DNN) を用いて合成された歌声に，発話間のピッチ変動を与える，generative moment matching network (GMMN) に基づくポストフィルタを提案する．人間の歌唱における発話間ピッチ変動は，自然性や豊かさにつながるほか，同一曲を 2 回録音してミックスすることにより歌声に厚みを持たせるダブルトラッキング (DT) にも活用されている．しかし，従来の決定論的な DNN 歌声合成では，一つの楽譜に対して 1 種類の波形しか生成できない．そこで GMMN を用いて，人間の歌声におけるピッチの変調スペクトルの発話間変動をモデル化することにより，DNN で合成された歌声にランダム性を付与するポストフィルタを提案し，さらに，合成歌声のための DT に応用する．主観評価により，提案手法では，自然性を損なわずに，人間が知覚できる水準で発話間変動が生じること，GMMN を用いた DT の知覚的印象が，従来の信号処理を用いた DT の印象よりも自然な多重録音に近づくことを示す．

Hiroki Tamaru[1,a)]　Yuki Saito[1]　Shinnosuke Takamichi[1,b)]　Tomoki Koriyama[2]
Hiroshi Saruwatari[1]

1. はじめに

近年，合成音声の歌声を用いた音楽制作が盛んに行われている．合成手法としては，素片接続 (Vocaloid [1] など)，hidden Markov model (HMM) [2], [3]，deep neural network (DNN) [4], [5] が挙げられる．合成歌声を用いる目的の一つは，ユーザの性別や歌唱技量に関係なく，表現豊かな歌声や音楽を制作することである．中でも，DNN に基づく手法は，高品質で表現豊かな歌声を合成する手法として期待されている．

しかし，従来の DNN に基づく手法は，図 1 に示すように，発話間変動を欠く．人間は，単一の楽譜を与えられても，歌唱ごとに歌いまわしが異なる．この発話間変動は，豊かな音楽体験につながる．例えば，音楽コンサートにおいて歌手が口パクではなく生で歌うことにより臨場感をもたらすほか，音楽制作において同じフレーズの歌声を複数回録音し，音楽制作者がそれらの中から好みのテイクを選ぶことが可能になる．一方，従来の DNN に基づく手法では，合成過程が決定論的なため，一つの楽譜からは一つの歌声しか生成されない．発話間変動がないため，前述した豊かな音楽体験が存在しないほか，ダブルトラッキング (double-tracking: DT) [6], [7] (図 2) を行うことも不可能になる．DT は，同一フレーズを複数回歌唱および録音し，それらをミックスすることにより，発話間変動を活用して歌声に厚みや豊かさを持たせる手法である．また，1 回分の録音しか存在しない場合でも，アーティフィシャルダブルトラッキング (artificial double-tracking: ADT) と呼ばれる，信号処理的な代替法が存在する．ADT では，1 回の録音を主にコーラスのエフェクトによって変調して原音にミックスする．ADT は発話間変動を持った複数の録音を必要としないため，人間の歌声のみならず，合成歌声にも利用できる．しかし，ADT では不自然な音質の変化が生じることが知られている [8]．

これらの問題に対処するため，本稿では合成歌声に発話間変動を付与するポストフィルタを提案する．歌唱の発話間変動は何らかの複雑な分布に従うと仮定し，複雑な分布のモデル化への有効性が示されている，深層生成モデル

[1] 東京大学 大学院情報理工学系研究科システム情報学専攻, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.
[2] 東京工業大学 工学院 情報通信系, G2-4, 4259 Nagatsuta-cho, Midori-ku, Yokohama, 226-8502, Japan.
[a)] hiroki_tamaru@ipc.i.u-tokyo.ac.jp
[b)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

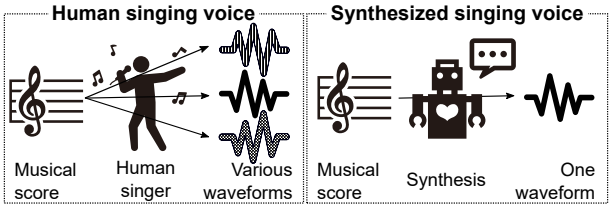

図 1　人間の歌声と合成歌声の比較.

Fig. 1　Comparison of human and synthesized singing voices.

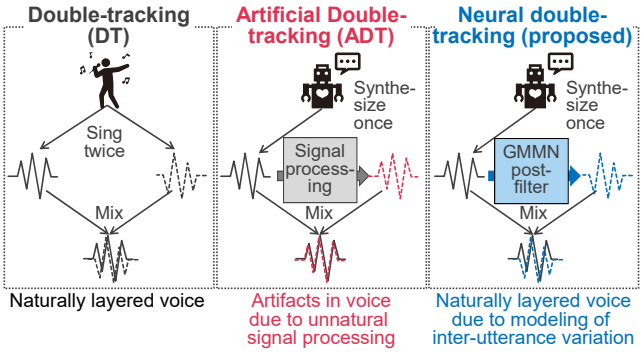

図 2　ダブルトラッキング (DT), アーティフィシャルダブルトラッキング (ADT) と提案するニューラルダブルトラッキング (NDT). この図では，ADT を合成歌声に対して適用しているが，同技術は人間の歌声に対しても適用可能である.

Fig. 2　Double-tracking (DT), artificial double-tracking (ADT), and proposed neural double-tracking. This figure shows ADT performed on synthesized voice, but it can also be performed on natural voice.

を用いる．我々の以前の研究 [9] と同様に，深層生成モデルとして generative moment matching network (GMMN) [10], [11] を用いる．このモデルは，複雑なミニマックス問題を解く必要がある generative adversarial network (GAN) [12] や，過度な正則化により生成の性能が低い variational auto-encoder [13] とは異なり，容易な学習と安定した生成が可能である [14]．提案手法では，合成されたピッチ系列とノイズベクトルを入力とし，自然音声のピッチの発話間変動を表現できるように条件付き GMMN を学習させる．ピッチの超分節的構造を捕捉するため，ピッチ系列の変調スペクトル (modulation spectrum: MS) [15] を用い，GMMN は自然音声の MS の変動をモデル化する．合成時には，ノイズベクトルを入力すると，GMMN はランダムに発話間変動を持った MS を生成する．入力のピッチ系列を，ランダムに生成された MS を用いて変調することにより，自然な範囲内で原音と異なった歌声をランダムに生成することができる．本稿では，さらにこの枠組みを ADT に応用したニューラルダブルトラッキング (neural double-tracking: NDT) を提案する．GMMN に基づくポストフィルタは自然な発話間変動を生成するため，NDT は自然な厚みを持った歌声を生成できる．実験的評価により，提案するポストフィルタは合成歌声の自然性を損なわずに，人間の知覚できる水準の発話間変動を生成できること，さらに NDT は信号処理的な ADT よりも，自然な DT に知覚的に近いことが示される.

2. 従来手法

2.1 DNN に基づく歌声合成

DNN に基づく歌声合成 [4] では，楽譜と対応する歌声の音声パラメータとの関係が DNN によりモデル化される．まず，楽譜は言語的・音楽的コンテキストを表現するベクトル系列に変換される．DNN はコンテキスト系列を入力されると歌声の音声パラメータを出力すべく，以下の平均二乗誤差 (mean square error: MSE) を最小化 [4] するように学習を行う.

$$L_{\mathrm{MSE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = ||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2 \qquad (1)$$

ただし $\boldsymbol{y}$ は自然音声の，$\hat{\boldsymbol{y}}$ は合成音声の音声パラメータ系列であり，それぞれ T フレームであるとする．本稿では 1 次元の連続対数基本周波数 (F_0) 系列 [16] を考えるため，$\boldsymbol{y}$ をスカラー値の系列 $[y(1), \cdots, y(t), \cdots, y(T)]^\top$ と定義する．ただし，$y(t)$ は連続対数 F_0，$\top$ は転置記号である．合成の際は，推定された音声パラメータ $\hat{\boldsymbol{y}} = [\hat{y}(1), \ldots, \hat{y}(t), \ldots, \hat{y}(T)]^\top$ を用いて歌声を合成する．合成過程は決定論的であるため，合成される歌声は発話間変動を持たない.

2.2 ADT

図 2 に，DT と ADT の違いを示す．DT は，同一フレーズを複数回歌唱および録音してミックスすることにより，歌声に厚み・豊かさを持たせる手法である [6], [7]．しかし，同一フレーズを 2 回ほぼ同じように歌う (タイミングや伸ばす長さを合わせる) のは難しく，時間と手間がかかる．ADT は，1 回の録音しか必要としない，DT の信号処理的な代替法である．ADT は，ボーカルの信号を 1 台のテープマシンから出力し，2 台目のテープマシンで変速して再生し，1 台目に戻して原音に重ねて録音したのが始まりである [6]．近年では，コーラスエフェクト [8] に代表される信号処理的な手法が主に用いられる．コーラスエフェクトでは，波形のコピーに対して，そのピッチを low-frequency oscillator を用いて変調する．すなわち，元のピッチ系列に対して時間変動する関数 (主に正弦波) が加算される．そして，多重録音感 (人間が同一フレーズを 2 回歌唱および録音した感じ) を強調するため，変調歌声に微少な時間遅延を付与したのち，変調歌声を原音にミックスする．ADT は 1 個の歌声波形のみで行えるので，合成歌声に関しても実行可能である．しかし，ADT は位相の類似した 2 個の波形を重ねるので，コムフィルタ効果とそれに起因する不自然な音質の変化が生じてしまう [8].

3. GMMN に基づくポストフィルタと NDT への応用

本節では，GMMN に基づくポストフィルタと，合成歌声

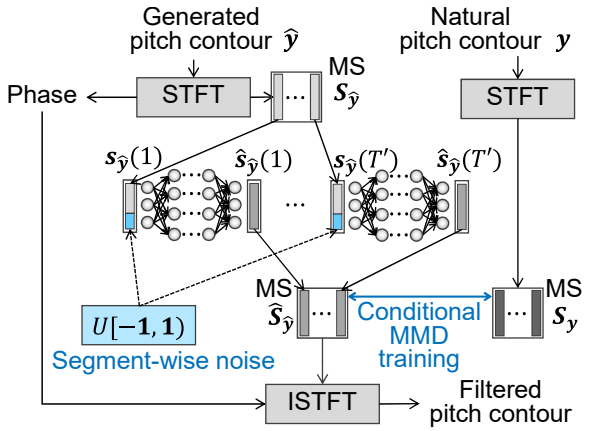

図 3　提案するポストフィルタ．
Fig. 3　Schematic diagram of our post-filter.

のための NDT を提案する．まず，ピッチの超文節的構造を捕捉するため，連続対数 F_0 の MS を自然音声と合成音声のそれぞれについて抽出する．次に，自然な発話間変動を持ったピッチ系列がランダムに生成できるよう，GMMN の学習を行う．最後に，合成歌声のパラメータにポストフィルタをかけてボコーディングにより音声を合成し，NDT はそれと元の音声をミックスして行う．

3.1　MS の抽出

MS は，音声パラメータ系列にフーリエ変換を行って得られるパワースペクトルの対数と定義され [15]，系列の時間構造を捕捉するために用いることができる．$\boldsymbol{y}$ の MS である $\boldsymbol{S_y}$ は，短時間フーリエ変換 (short-time Fourier transform: STFT) を用いて以下のように算出できる．

$$\boldsymbol{S_y} = [\boldsymbol{s_y}(1), \cdots, \boldsymbol{s_y}(\tau), \cdots, \boldsymbol{s_y}(T')] \tag{2}$$

$$\boldsymbol{s_y}(\tau) = [s_{\boldsymbol{y}}(\tau, 0), \cdots, s_{\boldsymbol{y}}(\tau, m), \cdots, s_{\boldsymbol{y}}(\tau, M)]^\top \tag{3}$$

ただし τ はセグメント (1 個のセグメントは 1 個の窓かけされた連続 F_0 系列に対応) のインデックス，m は変調周波数インデックスである．$s_{\boldsymbol{y}}(\tau, m)$ は変調周波数 m，セグメント τ における MS である．T' はセグメントの総数，M は STFT の窓長の半分である．$\hat{\boldsymbol{y}}$ の MS である $\boldsymbol{S}_{\hat{\boldsymbol{y}}}$ も同様に算出される．ゼロパディングに起因する誤差の問題を防ぐため，本研究ではゼロ平均連続 F_0 系列 [15] を用いる．ポストフィルタをかけた MS と，元の位相を用い，それらに逆 STFT を行うことによって，新たな F_0 系列を得ることができる．STFT 分析の条件 (窓長や窓関数) は，STFT と逆 STFT による完全再構成条件を満たすよう設定する．変調周波数の低い成分 (緩やかに時間変化する構造に対応する) のみをフィルタリングの対象とする．変調周波数の高い成分に変動を与えると，F_0 軌跡に高速で不自然な変動が生じてしまうためである．

3.2　GMMN に基づくポストフィルタリング

連続 F_0 系列 $\hat{\boldsymbol{y}}$ をランダムに変調する，GMMN に基づくポストフィルタについて説明する．図 3 に，提案するポストフィルタを示す．GMMN [10] は深層生成モデルの一つであり，GAN [12] よりも安定した学習が可能である．学習の規範は maximum mean discrepancy (MMD) と呼ばれ，二つの分布間の統計量の不一致度を表す．DNN はノイズベクトルを入力とし，これがフィルタリングの際の発話間変動の源となる．ポストフィルタの学習段階では，合成された連続 F_0 の MS とセグメントごとのノイズを入力されると，自然音声の連続 F_0 の MS の条件付き分布が DNN によりモデル化される．$\boldsymbol{n}(\tau) \sim U[-\mathbf{1}, \mathbf{1})$ をセグメント τ におけるノイズベクトル，$G(\cdot)$ をポストフィルタの DNN とする．セグメント τ における DNN の入力は，ジョイントベクトル $[\boldsymbol{s}_{\hat{\boldsymbol{y}}}(\tau)^\top, \boldsymbol{n}(\tau)^\top]^\top$ であり，出力はフィルタリングされた MS: $\hat{\boldsymbol{s}}_{\hat{\boldsymbol{y}}}(\tau)$ である．すなわち，$\hat{\boldsymbol{s}}_{\hat{\boldsymbol{y}}}(\tau) = G([\boldsymbol{s}_{\hat{\boldsymbol{y}}}(\tau)^\top, \boldsymbol{n}(\tau)^\top]^\top)$ である．$\hat{\boldsymbol{S}}_{\hat{\boldsymbol{y}}} = [\hat{\boldsymbol{s}}_{\hat{\boldsymbol{y}}}(1), \cdots, \hat{\boldsymbol{s}}_{\hat{\boldsymbol{y}}}(\tau), \cdots, \hat{\boldsymbol{s}}_{\hat{\boldsymbol{y}}}(T')]$ とおくと，以下に示す条件付き MMD (conditional MMD: CMMD) [11] を最小化するように学習が行われる．

$$\begin{aligned} L_{\mathrm{CMMD}}(\boldsymbol{S}_{\hat{\boldsymbol{y}}}, \boldsymbol{S}_{\boldsymbol{y}}, \hat{\boldsymbol{S}}_{\hat{\boldsymbol{y}}}) = \frac{1}{T'^2}\{&\mathrm{tr}(\boldsymbol{L}_{\boldsymbol{S}_{\hat{\boldsymbol{y}}}} \cdot \boldsymbol{K}_{\boldsymbol{S}_{\boldsymbol{y}}, \boldsymbol{S}_{\boldsymbol{y}}}) \\ &+ \mathrm{tr}(\boldsymbol{L}_{\boldsymbol{S}_{\hat{\boldsymbol{y}}}} \cdot \boldsymbol{K}_{\hat{\boldsymbol{S}}_{\hat{\boldsymbol{y}}}, \hat{\boldsymbol{S}}_{\hat{\boldsymbol{y}}}}) \\ &- 2 \cdot \mathrm{tr}(\boldsymbol{L}_{\boldsymbol{S}_{\hat{\boldsymbol{y}}}} \cdot \boldsymbol{K}_{\boldsymbol{S}_{\boldsymbol{y}}, \hat{\boldsymbol{S}}_{\hat{\boldsymbol{y}}}})\} \end{aligned} \tag{4}$$

$$\boldsymbol{L}_{\boldsymbol{S}_{\hat{\boldsymbol{y}}}} = \tilde{\boldsymbol{H}}^{-1}_{\boldsymbol{S}_{\hat{\boldsymbol{y}}}} \boldsymbol{H}_{\boldsymbol{S}_{\hat{\boldsymbol{y}}}} \tilde{\boldsymbol{H}}^{-1}_{\boldsymbol{S}_{\hat{\boldsymbol{y}}}} \tag{5}$$

$$\tilde{\boldsymbol{H}}_{\boldsymbol{S}_{\hat{\boldsymbol{y}}}} = \boldsymbol{H}_{\boldsymbol{S}_{\hat{\boldsymbol{y}}}} + \lambda \boldsymbol{I}_{T'} \tag{6}$$

ただし，$\boldsymbol{I}_{T'}$ は T'-by-T' の単位行列，λ は正則化係数である．$\boldsymbol{K}_{\hat{\boldsymbol{S}}_{\hat{\boldsymbol{y}}}, \boldsymbol{S}_{\boldsymbol{y}}}$ は $\hat{\boldsymbol{S}}_{\hat{\boldsymbol{y}}}$ と $\boldsymbol{S}_{\boldsymbol{y}}$ との間の T'-by-T' のグラム行列である．すなわち，(i, j) 成分は $k(\hat{\boldsymbol{s}}_{\hat{\boldsymbol{y}}}(i), \boldsymbol{s}_{\boldsymbol{y}}(j))$ (ただし $k(\cdot)$ は二つのベクトルに対する任意のカーネル関数) である．同様に，$\boldsymbol{H}_{\boldsymbol{S}_{\hat{\boldsymbol{y}}}}$ は $\boldsymbol{S}_{\hat{\boldsymbol{y}}}$ に関するグラム行列であり，(i, j) 成分は $h(\boldsymbol{s}_{\hat{\boldsymbol{y}}}(i), \boldsymbol{s}_{\hat{\boldsymbol{y}}}(j))$ である (ただし $h(\cdot)$ は二つのベクトルに対する任意のカーネル関数)．$k(\cdot)$ と $h(\cdot)$ は異なる関数でも構わない．学習後には，このモデルは合成されたピッチ系列の MS を入力すると，自然音声の MS の分布 (ピッチ系列の発話間変動) を表現することができる．合成段階では，まず DNN 歌声合成の歌声パラメータ生成により $\hat{\boldsymbol{y}}$ を合成し，その MS と位相を計算する．次に，学習済み GMMN とランダムに生成されたノイズを用いて MS をフィルタリングして自然な発話間変動を持たせる．最終的な F_0 軌跡はフィルタリングされた MS と元の位相に逆 STFT を行うことによって生成する．

3.3　NDT への応用

合成歌声に自然な厚みを持たせる NDT を提案する．DNN 音声合成の歌声パラメータ生成の後，一つの波形を通常のボコーダ処理により合成する．もう一つの波形は，ポストフィルタで変調された F_0 系列を用いて合成する．NDT 後の歌声は，変調した歌声波形に微少な時間遅延を

付与した後，その歌声波形を元の合成歌声にミックスすることで得られる．

3.4 考察

我々の以前の研究 [9] では，テキスト音声合成において，GMMN に基づいて，フレームごとにノイズを与えることにより，スペクトルに発話間変動を持たせる手法を設計した．しかし，この手法には二つの問題があった．第一に，フレームごとにノイズを入力して変動をモデリングすると，F_0 軌跡に関しては速すぎる不自然な変動が発生すること，第二に，言語情報はスパースな特徴量であるため，言語情報で条件づけられた GMMN は知覚できる水準の発話間変動を生じさせることができないことである．本稿の手法は，効果的にこれらの問題を解決する．第一の問題は，F_0 軌跡の MS という，低次元で効果的な表現を用いてセグメントごとの時間構造を捕捉すること，また低変調周波数成分のみフィルタリングすることで，F_0 軌跡の連続性を維持したまま変調を行うことにより解決する．第二の，スパース性の問題は，GMMN をポストフィルタとして用いて，言語情報を用いないことにより解決し，4.2 節に示すように，知覚できる水準の発話間変動を生じさせることができる．

我々の手法は自然音声の MS の分布を考慮するので，フィルタリングされた歌声の変動も自然な範囲に収まっていると考えられる．図 4 に，MSE に基づく決定論的なピッチ系列と，それにポストフィルタをかけて生成したピッチ系列の例を示す．提案するポストフィルタは，ランダムかつ連続性を保ったピッチ系列を生成できていることがわかる．本研究は，1) GMMN を用いて MS をモデル化することにより，自然な発話間変動を生成した，また 2) そのようなシステムを歌声のポストフィルタとして導入した最初の研究である．1 節で述べたように，発話間変動は，複数の歌声から好みのものを選択することを可能にする．提案法では，入力のノイズを固定することにより，ピッチ系列を保存することができるため，数回ランダム生成を行い，好きな歌い方をフレーズごとに選んで，それらを結合して 1 曲にする「良いとこ取り」も可能である．

従来の ADT は，ピッチの自然な変動を考慮せずに原音を決定論的に変調して生成した波形を用いる．一方，NDT はピッチの自然な変動を考慮して変調を行う点が新しく，4.4 節に示すように，従来の ADT よりも多重録音感が高くなる．

データ拡張は DNN のモデリング精度を向上させるための強力な手法である．本研究では STFT の際にデータ拡張を用いた．MS の算出には STFT を用いるため，セグメント分析の開始フレーム位置は，MS の値に大きな影響を与える．そのような摂動をカバーするため，とり得るすべてのオフセットを分析開始位置に対して加え，それぞれについて MS を算出した．この手法により，ポストフィルタの学習データを増やし，学習の精度を向上することができる．

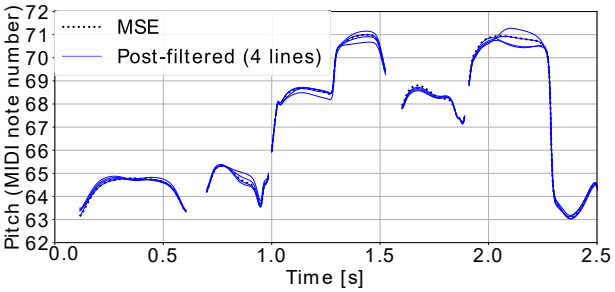

図 4 最小二乗誤差規範 (MSE) に基づく従来の DNN 歌声合成によるピッチ系列と，ポストフィルタをかけた 4 種類のピッチ系列．縦軸は 1 が半音に相当．

Fig. 4 Example of generated pitch contours. We used proposed method to sample four contours. Value of unity on vertical axis is equal to semitone.

提案手法ではピッチ系列をニューラルネットワークの入出力としているため，NDT を人間の歌声に対して行う，すなわち人間が 1 回歌った録音に対して自然な厚みを持たせる手法にも拡張できることが期待される．

4. 実験的評価

4.1 実験条件

歌声のデータベースとして，HTS [17] の歌声合成デモから 31 曲，JSUT-song コーパス [18] から 26 曲，我々のデータベース (歌唱者は JSUT-song と同じ) から 9 曲を用いた．これらのコーパスから，58 曲を歌声合成の DNN の学習に，HTS のデモから 28 曲をポストフィルタの学習に，HTS のデモのうち学習に用いなかった 3 曲を評価に用いた．DNN 歌声合成の学習時には，半音上げ，もしくは半音下げのトランスポーズによるデータ拡張 [5] を行った．サンプリング周波数は 16 kHz とした．音声特徴量の抽出と音声波形合成には WORLD [19] を用いた．フレームシフトは 5 ms とした．MSE に基づく音声パラメータの予測に用いた DNN は Feed-Forward 型で，705 次元の入力層，3×256 ユニットの gated linear unit (GLU) [20] からなる隠れ層，127 ユニットの線形出力層から構成された．ラーニングレートは 0.005，バッチサイズは 500，勾配は AdaGrad [21] とし，50 エポックの学習を行った．705 次元の入力特徴量は 688 次元の言語・音楽特徴量，one-hot の曲コードと one-hot の歌手コード [22] から構成された．出力は，127 次元のベクトルで，40 次元のメルケプストラム係数，連続対数 F_0，非周期性指標 [23], [24]，それら 42 次元の動的特徴量 (1 次と 2 次) [25]，2 値の有声・無声ラベルからなった．条件付き GMMN に用いた DNN は，Feed-Forward 型で，11 次元の入力層，3×128 ユニットの GLU からなる隠れ層，input-to-output residual net [26] から構成された．ラーニングレートは 0.005，バッチサイズは 13000，勾配は

表 1　発話間変動を知覚したと回答した率
Table 1　Answer rate of perceived inter-utterance difference

Proposed	MSE	p-value
0.276	0.176	7.45×10^{-3}

表 2　時間長条件 middle と long に関する，歌声の自然性の評価スコアと p 値
Table 2　Preference scores of singing voice naturalness and their p-values for the middle and long conditions

Length condition	Proposed	MSE	p-value
Middle	0.504	0.496	8.58×10^{-1}
Long	0.480	0.520	3.72×10^{-1}

AdaGrad とし，10 イテレーションの学習を行った．式 (5) における $\boldsymbol{L}_{\boldsymbol{S}_{\hat{\boldsymbol{y}}}}$ の計算では，逆行列の演算が現実的な時間では不可能なため，1024 次元の random Fourier feature [27] を用いて近似を行った．11 次元の入力ベクトルは，MSE に基づいて合成されたピッチの 1 次 ($m=1$) の MS と，一様分布 $U[-\mathbf{1},\mathbf{1})$ から生成された 10 次元のノイズベクトルからなった．学習の安定化のため，ノイズベクトルは各セグメントに対して生成し，学習の間は固定した．正則化係数 λ は 0.01 とした．カーネル関数は，ガウシアンとした ($\exp\{-||\boldsymbol{s}_{\boldsymbol{y}}(i)-\hat{\boldsymbol{s}}_{\hat{\boldsymbol{y}}}(j)||^2/\sigma^2\}$)．入力特徴量に関してもガウシアンカーネルを用いた．σ は，入力に関しては 100.0，出力に関しては 1.0 とした．これらの値は実験的に定めた．自然音声の MS は区間 $[0.01, 0.99]$ に収まるように正規化した．STFT には，96 フレーム (480 ms) のハニング窓，48 フレーム (240 ms) のセグメントシフトを用いた．ピッチ変調の効果を明瞭にするため，ボコーディングの際，スペクトルパラメータ，非周期性指標，有声・無声ラベルは，自然音声のものを用いた．

本稿では，(1) 提案ポストフィルタは知覚できる水準の発話間変動を生じさせるか，(2) 提案ポストフィルタは合成歌声の自然性が低下させないか，(3) NDT は ADT よりも自然な多重録音感を再現できるかの 3 点を評価した．評価はクラウドソーシングサービスであるランサーズ [28] 上で行った．参加者の判断を容易にするため，曲を short，middle，long の 3 種類の時間長に手動で切り分けた．平均した長さは，それぞれ 3.01 s，4.88 s，10.24 s であった．

4.2　発話間変動の知覚

発話間変動が人間に知覚できるかを調べるため，25 人の参加者に，二つの歌声の対を聴いてそれらの間に違いがあると感じたか否かの回答を求めた．時間長の短い歌声を提示した方が細かい違いを記憶しやすいと考えられるため，時間長条件 short の歌声を使用した．各参加者は，ランダムにポストフィルタリングされた 2 種類の歌声 10 対 (proposed) と，MSE に基づく完全に同一の音声 2 回からなる 10 対 (MSE)，合計 20 対をランダムな順序で聴取した．ウェルチの t 検定を行って p 値を算出した．

表 3　時間長条件 middle と long に関する，多重録音感の評価スコアと p 値
Table 3　Preference scores of double-trackedness and their p-values for the middle and long conditions

Length condition	NDT	ADT	p-value
Middle	**0.724**	0.276	$<10^{-10}$
Long	**0.736**	0.264	$<10^{-10}$

表 1 に結果を示す．MSE 条件では，全く同じ音声が提示されたにもかかわらず，17.6%において違いがあると誤認された．一方，提案法における差異の知覚率はそれよりも有意に高い．したがって，提案法は知覚できる水準の発話間変動を生じさせることができていると考えられる．

4.3　ポストフィルタをかけた歌声の自然性

ポストフィルタをかけると合成歌声のピッチの自然性が低下しないか調べるため，25 人の参加者に，ポストフィルタリングされた歌声と，MSE に基づく歌声の対を 10 対聴き，より自然な方を選ぶよう求めた．全体的な自然性の評価を容易にするため，時間長条件は，middle および long とした．

表 2 に結果を示す．統計的に有意な差は，中・長いずれの条件にも見られない．これにより，ポストフィルタをかけてもピッチの自然性を損なわないことが示唆される．

4.4　NDT の評価

多重録音感を，提案する NDT と従来の ADT について評価・比較した．それぞれの条件を以下に示す．

NDT:　連続対数 F_0 系列を提案法のポストフィルタで変調し，それを用いてボコーダで合成した歌声を MSE に基づく歌声に重ねた．

ADT:　連続対数 F_0 系列を low-frequency oscillator で変調し，それを用いてボコーダで合成した歌声を MSE に基づく歌声に重ねた．オシレーションの波形は正弦波，周期は 0.775 Hz，深さは半音の 10%とした．これらのパラメータは [8] を参考に定めた．変調を波形ドメインでなくボコーダパラメータドメインで行ったのは，その方がアーティファクトが発生しにくいと考えられるためである．

通常の ADT の設定 [8] を再現するため，2 種類の手法とも，変調した波形は 20 ms 遅らせ，3 dB 音量を下げて原音にミックスした．25 人の参加者に，これら 2 種類の条件で厚みを持たせた音声の対を 10 対聴き，より実際に多重録音したように聴こえる方を選ぶよう求めた．4.3 節と同様，時間長条件は，middle および long とした．

表 3 に結果を示す．中・長いずれの条件でも，NDT の評価スコアは従来の ADT と比べ，有意に大きい．これにより，従来の信号処理的なアプローチよりも，提案する深

層生成モデルを用いた手法のほうが，より自然な DT のような聴覚的印象を出せることが示された．

5. まとめ

本稿では，歌声の発話間変動を生成する，GMMN に基づくポストフィルタを提案し，また ADT への応用についても述べた．人間は同じ歌を完全に同じように 2 回歌うことはあり得ないが，従来の歌声合成システムでは，最も自然とみなされる歌声 1 個が合成されるのみであった．我々のポストフィルタは，GMMN を用いて，合成音声と自然音声の MS の統計量を揃えるように学習し，発話間変動をモデル化する．実験評価により，提案するポストフィルタは，歌声の自然性を損なうことなく，人間に知覚可能な発話間のピッチ変動を生成できることが示された．さらに，ADT にポストフィルタを応用した際の有効性についても議論した．実験評価により，提案する NDT は従来の ADT よりも，知覚的に自然な DT に近い音になることが示された．

今後は，音符の継続長やスペクトルパラメータに関して発話間変動のモデリングを行い，ピッチのそれと組み合わせてより自然な発話間変動を生成することや，人間の歌声の MS を変調できるポストフィルタの設計を検討する．

謝辞： 本研究の一部は，セコム科学技術支援財団の助成を受け実施した．

参考文献

[1] H. Kenmochi and H. Ohshita, “Vocaloid - commercial singing synthesizer based on sample concatenation,” in *Proc. INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 4011–4012.

[2] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, “An HMM-based singing voice synthesis system,” in *Proc. ICSLP*, Pittsburgh, U.S.A., Sep. 2006, pp. 2274–2277.

[3] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, “Recent development of the HMM-based singing voice synthesis system - Sinsy,” in *Proc. SSW7*, Kyoto, Japan, Sep. 2010, pp. 211–216.

[4] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Singing voice synthesis based on deep neural networks,” in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 2478–2482.

[5] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer modeling timbre and expression from natural songs,” *Applied Sciences*, vol. 7, no. 12, Dec. 2017.

[6] R. Brice, *Music Engineering*. Elsevier Science, Oct. 2001.

[7] K. Womack, *The Beatles Encyclopedia: Everything Fab Four*. ABC-CLIO, Jun. 2014.

[8] R. Izhaki, *Mixing Audio: Concepts, Practices, and Tools*. Taylor & Francis, Oct. 2017.

[9] S. Takamichi, K. Tomoki, and H. Saruwatari, “Sampling-based speech parameter generation using moment-matching network,” in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 3961–3965.

[10] Y. Li, K. Swersky, and R. Zemel, “Generative moment matching networks,” in *Proc. ICML*, Lille, France, Jul. 2015, pp. 1718–1727.

[11] Y. Ren, J. Li, Y. Luo, and J. Zhu, “Conditional generative moment-matching networks,” in *Proc. NIPS*, Barcelona, Spain, Dec. 2016, pp. 2928–2936.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. NIPS*, Montreal, Canada, Dec. 2014, pp. 2672–2680.

[13] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv*, vol. abs/1312.6114, 2013.

[14] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, “Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors,” in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5274–5278.

[15] 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, Apr. 2016.

[16] K. Yu and S. Young, “Continuous F0 modeling for HMM based statistical parametric speech synthesis,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1071–1079, Jul. 2011.

[17] “HMM-based speech synthesis system (HTS),” http://hts.sp.nitech.ac.jp/.

[18] “JSUT-song,” https://sites.google.com/site/shinnosuketakamichi/publication/jsut-song.

[19] 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. E99-D, no. 7, pp. 1877–1884, Jul. 2016.

[20] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” *arXiv*, vol. abs/1612.08083, 2016.

[21] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *EURASIP Journal on Applied Signal Processing*, vol. 12, pp. 2121–2159, Jul. 2011.

[22] N. Hojo, Y. Ijima, and H. Mizuno, “An investigation of DNN-based speech synthesis using speaker codes,” in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 2278–2282.

[23] H. Kawahara, J. 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*, Firentze, Italy, Sep. 2001, pp. 1–6.

[24] 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.

[25] 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, Jun. 2000, pp. 1315–1318.

[26] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, “Generative adversarial network-based postfilter for STFT spectrograms,” in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 3389–3393.

[27] A. Rahimi and B. Recht, “Random features for large-scale kernel machines,” in *Proc. NIPS*, Vancouver, Canada, Dec. 2008, pp. 1177–1184.

[28] “Lancers,” https://www.lancers.jp/.