Generative moment matching net に基づく歌声のランダム変調ポストフィルタと double-tracking への応用 *

☆田丸 浩気，齋藤 佑樹，高道 慎之介 (東大院・情報理工)，
郡山 知樹 (東工大)，猿渡 洋 (東大院・情報理工)

1 はじめに

近年，歌声合成を用いた音楽制作が盛んに行われている．歌声合成の目的の一つは，ユーザの性別や歌唱技量に関係なく，表現豊かな歌声を制作することである．数ある手法の中でも，deep neural network (DNN) に基づく手法 [1, 2] は，高品質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，発話間変動を欠く (Fig. 1)．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り，コンサートの臨場感が高まるほか，音楽制作において複数回録音した同一フレーズの取捨選択が可能になるなど，豊かな音楽体験につながる．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，前述の豊かさがないほか，double-tracking (DT) [3, 4] (Fig. 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・録音して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を持たせる手法である．また，1 回の録音しか存在しない場合でも，artificial double-tracking (ADT) と呼ばれる，1 回の録音を信号処理的に変調して原音にミックスする代替法が存在する．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，合成歌声にも利用できる．しかし，ADT は不自然な音質の変化を生じさせることが知られている [5]．

そこで，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，複雑な分布をモデル化できる，深層生成モデルを用いる．我々の以前の研究 [6] と同様に，学習が容易な generative moment matching network (GMMN) [7, 8] を用いる．提案手法では，ピッチの超分節的構造を，変調スペクトル (modulation spectrum: MS) [9] を用いて捕捉し，GMMN を用いて自然歌声の MS の発話間変動をモデル化する．DNN で合成されたピッチ系列の MS と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．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，自然な厚みを持った歌声を生成する neural double-tracking (NDT) を提案する．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を示す．

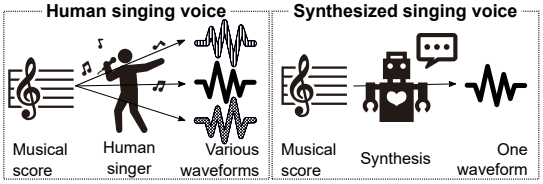

Fig. 1 人間の歌声と合成歌声の比較．

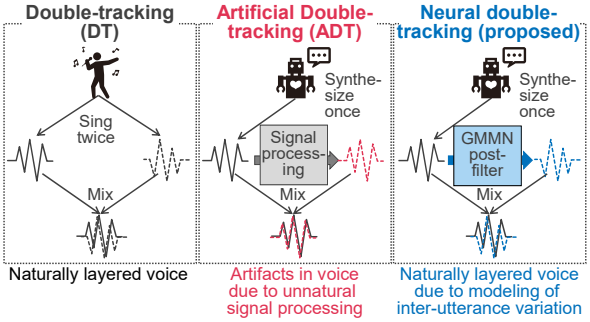

Fig. 2 Double-tracking (DT), artificial double-tracking (ADT) と提案する neural double-tracking (NDT)．この図では，ADT を合成歌声に対して適用しているが，同技術は人間の歌声に対しても適用可能である．

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 [1] では，楽譜と対応する歌声の音声パラメータとの関係を DNN でモデル化する．まず，楽譜は言語・音楽的コンテキストの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 squared error: MSE) を最小化 [1] するように学習する．

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

ただし $\boldsymbol{y}$ は自然歌声の，$\hat{\boldsymbol{y}}$ は合成歌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 $\boldsymbol{y}$ を 1 次元の連続対数基本周波数 (F_0) [10] 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

Fig. 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・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 [3, 4]．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

* Generative moment matching network-based random modulation post-filter for singing voices and its application to double-tracking, by Hiroki TAMARU, Yuki SAITO, Shinnosuke TAKAMICHI (The University of Tokyo), Tomoki KORIYAMA (Tokyo Institute of Technology), and Hiroshi SARUWATARI (The University of Tokyo)

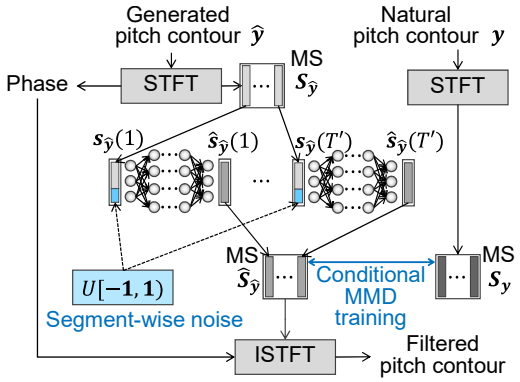

Fig. 3　提案するポストフィルタ.

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 [3]．近年では，コーラスエフェクト [5] 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，コピーした原音のピッチを 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，適度な多重録音感 (人間が同一フレーズを 2 回歌唱および録音した感じ) を与えるため，変調歌声に微少な時間遅延の付与と音量の減少処理を行ったのち，変調歌声を原音にミックスする．ADT は 1 個の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 [5].

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 を自然歌声と合成歌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，合成歌声のピッチ系列の MS にポストフィルタをかけてボコーダにより歌声を合成し，NDT はそれと元の歌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 [9]，系列の時間構造を表す．$\boldsymbol{y}$ 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 (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 $\boldsymbol{S_{\hat{y}}}$ も同様に算出される．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．変調周波数の高い成分を変調すると，F_0 軌跡に高速で不自然な変動が生じるため，緩やかな時間変化に対応する変調周波数の低い成分 のみをフィルタリングの対象とする．

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．Fig. 3 に，提案するポストフィルタを示す．GMMN [7] は深層生成モデルの一つである．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，自然歌声の連続 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 $[\boldsymbol{s_{\hat{y}}}(\tau)^\top, \boldsymbol{n}(\tau)^\top]^\top$ 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，$\hat{\boldsymbol{s}}_{\hat{\boldsymbol{y}}}(\tau) = G([\boldsymbol{s_{\hat{y}}}(\tau)^\top, \boldsymbol{n}(\tau)^\top]^\top)$ 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) [8] を最小化するように学習が行われる．

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

$$\boldsymbol{L_{S_{\hat{y}}}} = \tilde{\boldsymbol{H}}^{-1}_{\boldsymbol{S_{\hat{y}}}} \boldsymbol{H_{S_{\hat{y}}}} \tilde{\boldsymbol{H}}^{-1}_{\boldsymbol{S_{\hat{y}}}} \tag{5}$$

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

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．$\boldsymbol{K_{\hat{S}_{\hat{y}},S_y}}$ は $\hat{\boldsymbol{S}}_{\hat{\boldsymbol{y}}}$ と $\boldsymbol{S_y}$ 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 $k(\hat{\boldsymbol{s}}_{\hat{\boldsymbol{y}}}(i), \boldsymbol{s_y}(j))$ (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，$\boldsymbol{H_{S_{\hat{y}}}}$ は $\boldsymbol{S_{\hat{y}}}$ に関するグラム行列であり，(i,j) 成分は $h(\boldsymbol{s_{\hat{y}}}(i), \boldsymbol{s_{\hat{y}}}(j))$ 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，自然歌声の 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 歌声合成の音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変動させた MS を生成し，元の位相と組み合わせて逆 STFT を行うことによって最終的な F_0 軌跡を生成する．

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 歌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時間遅延の付与と音量を下げる処理をした後，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．

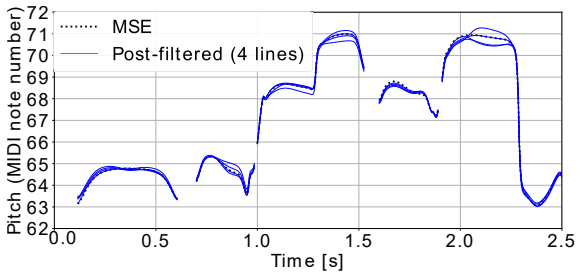

Fig. 4　最小二乗誤差規範 (MSE) に基づく従来の DNN 歌声合成によるピッチ系列と，ポストフィルタをかけた 4 種類のピッチ系列．縦軸は 1 が半音に相当．

3.4　考察

我々の以前の研究 [6] 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が，この手法には二つの問題があった．第一に，フレームごとの変動モデリングでは，F_0 軌跡の連続性が損なわれること，第二に，スパースである言語特徴量で条件づけられた GMMN は出力が縮退することである．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て変調する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，低次元な MS で条件付けることで解決する．

提案手法は自然歌声の MS の分布を考慮するので，変調後の歌声の変動も自然な範囲内にあると考えられる．Fig. 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．提案法では，入力のノイズを固定することにより，ピッチ系列をセグメントごとに保存することが可能であるため，人間の歌声のように，複数の歌声から好みのものを選択することが可能である．

従来の ADT は，原音を単純な波形で信号処理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．

提案手法ではピッチ系列を DNN 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

歌声データベースは，HTS [11] の歌声合成デモから 31 曲，JSUT-song コーパス [12] から 26 曲，in-house 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 から 28 曲をポストフィルタの学習に，HTS 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，半音上げ下げするトランスポーズによるデータ拡張 [2] 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 [13] を用い，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) [14] 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 [15] 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 [16] 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 [17, 18]，それら 42 次元の動的特徴量 (1 次と 2 次) [19]，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 [20] 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 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 [21, 22] 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．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．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，非周期性指標，有声・無声ラベルは，自然歌声の値を用いた．

(1) 提案ポストフィルタは知覚可能な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を低下させないか，(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のランサーズ [23] 上で行った．評価を容易にするため，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．違いの記憶を容易にするため，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歌声 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．

Table 1 に結果を示す．提案法における差異の知覚率は MSE 条件よりも有意に高い．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．

Table 1　発話間変動を知覚したと回答した率

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

Table 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

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

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．

Table 2 に結果を示す．統計的に有意な差は，middle・long いずれの条件にも見られない．ポストフィルタをかけてもピッチの自然性が損なわれることは確認されなかった．

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 系列を正弦波変調し，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．正弦波の周期は 0.775 Hz，振幅は半音の 10%とした．これらのパラメータは [5] を参考に定めた．

通常の ADT の設定 [5] 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歌声の対を 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．

Table 3 に結果を示す．middle・long 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，合成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．提案ポストフィルタは，GMMN を用いて，合成歌声と自然歌声の 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．

Table 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

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

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うことや，人間の歌声の MS を入力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] 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.

[2] 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.

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

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

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

[6] 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.

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

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

[9] 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.

[10] 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.

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

[12] "JSUT-song," https://sites.google.com/site/shinnosuketakamichi/publication/jsut-song.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 郡山知樹, 高道慎之介, and 小林隆夫, "グラム行列のスパース近似を用いた生成的モーメントマッチングネットに基づく音声合成の検討," in **音講論（春）**, no. 3-10-6, 東京, Mar. 2019.

[23] "Lancers," https://www.lancers.jp/.