ユーザ歌唱のための generative moment matching network に基づく neural double-tracking *

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

1 はじめに

音楽制作のボーカルレコーディング時に，同一歌唱者が同一フレーズを 2 回歌唱して多重録音することは double-tracking (DT) と呼ばれる [1, 2, 3, 4]．人間の歌唱は，同一歌唱者が同一フレーズを歌っても，歌うたびに変動 (発話間変動) が生じるため，DT を行うと単純に音量が上がるのではなく，歌声に厚み・豊かな質感を与えることができる．DT はビートルズが活用した [3] ことで知られる．しかし，DT においては同一フレーズを複数回歌唱する際，節回しや発音タイミング，音符を伸ばす時間長が概ね揃っている必要がある．それは歌唱者にとって難しく，手間がかかる．

DT を人工的に行う手法として，artificial (または automatic) double-tracking (ADT) が提案された．ADT は，原音のピッチ系列を信号処理的に変調した音を，原音にミックスする技術である [2]．具体的な手法にはバリエーションがあるが，代表的な手法ではピッチ系列を正弦波変調する，すなわち原音のピッチ系列に正弦波を加算することで変調を行う．しかし，ADT では類似した 2 音をミックスすることによる不自然音が生じる [2] 上，自然な発話間変動を考慮していないため，自然な多重録音感を実現できない．

そこで本稿では，発話間変動をモデル化し，歌声をあたかも再度歌ったかのように自然に変動させるポストフィルタを設計し，変調した歌声を原音とミックスすることにより，歌声に自然な多重録音感を付与する neural double-tracking (NDT) を提案する．著者らは以前，深層生成モデルである generative moment matching network (GMMN) [5, 6] に基づく NDT を提案した [7, 8] が，その際は deep neural network (DNN) を用いて合成された歌声を入力としていた．本稿では，新たに繰り返し歌唱データベースを用いることによってそれを拡張し，より実用性が高いであろう，自然歌声に対する NDT を提案する．

主観評価実験により，歌唱者依存モデルを用いた 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，さらに不特定歌唱者モデルを用いた NDT でも，歌唱者依存モデルと同等の性能が出せることを示す．

2 従来手法 (ADT)

Fig. 1 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手法である [1, 2, 3, 4]．人間は，同一歌唱者・同一曲であっても複数回歌唱すると変動 (発話間変動) が生じるため，複数テイクをミックスすると，単純に音量が大きくなるのではなく，音に厚みや豊かさが生じる．また，複数の歌声を左右に振り分けることで，ステレオ感のある歌声を用いた音楽表現を行うこともできる．しかし，DT は，2 回，歌唱タイミング・継続長や節回しをなるべく揃えて歌う必要があり，難しく，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 [4]．近年では，コーラスエフェクト [2] 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 (LFO) 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が，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 [2] ほか，発話間変動を考慮していないため，自然な多重録音感が再現できない．

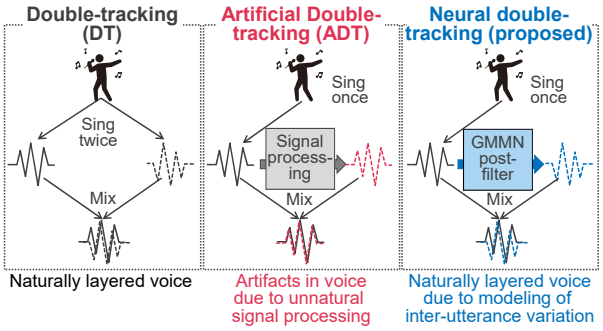

Fig. 1 Double-tracking (DT), artificial double-tracking (ADT) と提案する neural double-tracking (NDT).

3 GMMN に基づくポストフィルタを用いた NDT

本節で提案する NDT は，ADT のように自動で行えて，かつ DT のような自然な多重録音感を実現する技術である．まず，同一歌唱者が同一曲を複数回歌唱したデータベースを事前準備する．ピッチの超文節的構造を捕捉するため，連続対数 F_0 の変調スペクトル (modulation spectrum: MS) [9] を各歌声について抽出する．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 に基づくポストフィルタの学習を行う．最後に，入力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，それを元の歌声とミックスすることにより NDT を行う．

---

*Neural double-tracking based on generative moment matching network for users' singing, by Hiroki TAMARU, Yuki SAITO, Shinnosuke TAKAMICHI, Tomoki KORIYAMA, and Hiroshi SARUWATARI (The University of Tokyo)

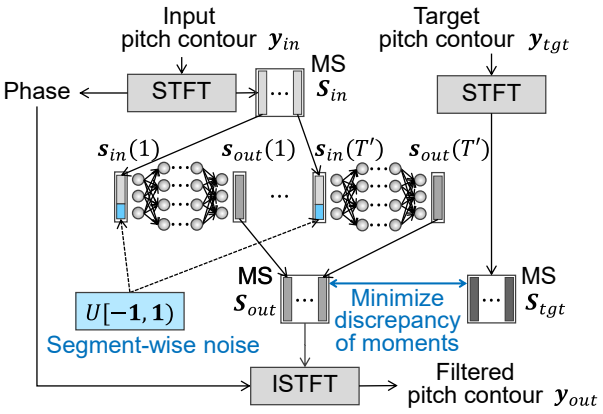

Fig. 2　提案するポストフィルタ．

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 $\boldsymbol{S}_{\boldsymbol{y}}$ 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．

$$\boldsymbol{S}_{\boldsymbol{y}} = [\boldsymbol{s}_{\boldsymbol{y}}(1), \cdots, \boldsymbol{s}_{\boldsymbol{y}}(\tau), \cdots, \boldsymbol{s}_{\boldsymbol{y}}(T')] \tag{1}$$

$$\boldsymbol{s}_{\boldsymbol{y}}(\tau) = [s_{\boldsymbol{y}}(\tau,0), \cdots, s_{\boldsymbol{y}}(\tau,m), \cdots, s_{\boldsymbol{y}}(\tau,M)]^\top \tag{2}$$

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．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 [9] 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．変調周波数の高い成分に変動を与えると，高速で不自然なピッチの振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

入力された連続 F_0 系列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. 2 に，提案するポストフィルタを示す．$\boldsymbol{y}_{in}$ は入力の連続 F_0 系列，$\boldsymbol{y}_{tgt}$ はターゲットの連続 F_0 系列であり，$\boldsymbol{S}_{in}$，$\boldsymbol{S}_{tgt}$ はそれぞれの MS である．GMMN [5]，条件付き GMMN [6] は深層生成モデルであり，generative adversarial network (GAN) [10] よりも安定した学習が可能である．本稿では，条件付き GMMN を用いる．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，ある歌声の連続 F_0 の MS とセグメントごとのノイズの結合ベクトルを入力とし，同一曲の別テイクの連続 F_0 の MS をターゲットとする．

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．$\boldsymbol{n}(\tau) \sim U[-\boldsymbol{1},\boldsymbol{1})$ 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，結合ベクトル $[\boldsymbol{s}_{in}(\tau)^\top, \boldsymbol{n}(\tau)^\top]^\top$ であり，出力は変動が加わった MS $\boldsymbol{s}_{out}(\tau)$ である．すなわち，$\boldsymbol{s}_{out}(\tau) = G([\boldsymbol{s}_{in}(\tau)^\top, \boldsymbol{n}(\tau)^\top]^\top)$ である．$\boldsymbol{S}_{out} = [\boldsymbol{s}_{out}(1), \cdots, \boldsymbol{s}_{out}(\tau), \cdots, \boldsymbol{s}_{out}(T')]$ とおくと，以下に示す条件付き MMD (conditional MMD: CMMD) [6] を最小化するように学習が行われる．

$$\begin{aligned} L(\boldsymbol{S}_{in}, \boldsymbol{S}_{tgt}, \boldsymbol{S}_{out}) = \frac{1}{T'^2}\{&\mathrm{tr}(\boldsymbol{L}_{\boldsymbol{S}_{in}} \cdot \boldsymbol{K}_{\boldsymbol{S}_{tgt},\boldsymbol{S}_{tgt}}) \\ &+ \mathrm{tr}(\boldsymbol{L}_{\boldsymbol{S}_{in}} \cdot \boldsymbol{K}_{\boldsymbol{S}_{out},\boldsymbol{S}_{out}}) \\ &- 2\mathrm{tr}(\boldsymbol{L}_{\boldsymbol{S}_{in}} \cdot \boldsymbol{K}_{\boldsymbol{S}_{tgt},\boldsymbol{S}_{out}})\} \end{aligned} \tag{3}$$

$$\boldsymbol{L}_{\boldsymbol{S}_{in}} = \tilde{\boldsymbol{H}}_{\boldsymbol{S}_{in}}^{-1} \boldsymbol{H}_{\boldsymbol{S}_{in}} \tilde{\boldsymbol{H}}_{\boldsymbol{S}_{in}}^{-1} \tag{4}$$

$$\tilde{\boldsymbol{H}}_{\boldsymbol{S}_{in}} = \boldsymbol{H}_{\boldsymbol{S}_{in}} + \lambda \boldsymbol{I}_{T'} \tag{5}$$

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．$\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{B}}$ は行列 $\boldsymbol{A}$ と行列 $\boldsymbol{B}$ との間の T'-by-T' のグラム行列を表し，例えば $\boldsymbol{K}_{\boldsymbol{S}_{tgt},\boldsymbol{S}_{out}}$ は $\boldsymbol{S}_{tgt}$ と $\boldsymbol{S}_{out}$ 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 $k(\boldsymbol{s}_{tgt}(i), \boldsymbol{s}_{out}(j))$ (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，$\boldsymbol{H}_{\boldsymbol{S}_{in}}$ は $\boldsymbol{S}_{in}$ に関するグラム行列であり，(i,j) 成分は $h(\boldsymbol{s}_{in}(i), \boldsymbol{s}_{in}(j))$ 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，このモデルは歌声ピッチ系列の MS $\boldsymbol{S}_{in}$ を入力すると，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，入力歌声の連続 F_0 系列 $\boldsymbol{y}_{in}$ から MS $\boldsymbol{S}_{in}$ 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 $\boldsymbol{S}_{in}$ をフィルタリングして変動させた MS $\boldsymbol{S}_{out}$ 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 系列 $\boldsymbol{y}_{out}$ を生成する．これにより，あたかも再度歌唱したかのように，ランダムに異なるピッチ系列を得ることができる．

3.3　NDT

歌声に自然な厚みを持たせる技術である NDT を提案する．歌声に対し，ボコーダ分析によってピッチ系列 $\boldsymbol{y}_{in}$ とスペクトル包絡，有声/無声ラベルを抽出する．$\boldsymbol{y}_{in}$ に上記のポストフィルタをかけてランダム変調を行い，変動させたピッチ系列 $\boldsymbol{y}_{out}$ を得る．$\boldsymbol{y}_{out}$ と，元のスペクトル包絡・有声/無声ラベルを用いてボコーダ合成を行うことで，あたかも再度歌ったかのような歌声を得ることができる．その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，元の歌声にミックスすることを，NDT と呼ぶ．

3.4　考察

著者らの以前の研究 [7, 8] では，DNN を用いて決定論的に合成された歌声に対して人間のようなランダムな発話間変動を付与し，NDT を行うためのポストフィルタを提案した．その際，入力は合成されたピッチ系列，ターゲットは対応する自然歌声 (1 テイクのみ) を用いた．

一方，本研究は，実際に同一曲を複数回歌唱したデータベースを用いて，自然歌声の発話間変動を深層生成モデルでモデル化し，自然歌声に対しランダムに変動を与えた，および多重録音感を付与した最

初の研究である．n 回歌唱したデータがあれば，入力 n 通りに対してターゲットはそれぞれ $n-1$ 通り考えられるため，$n(n-1)$ 通りの組み合わせのデータを用いることができ，効率的に学習データ量を多くすることができる．

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．

ポストフィルタのモデルには，生成時に入力したい歌声の歌唱者の歌声を用いて学習する場合 (歌唱者依存モデル) と，入力歌唱者とは異なる複数の歌唱者の歌声を用いて学習する場合 (不特定歌唱者モデル) の 2 通りが考えられる．後者でも前者と同程度の多重録音感を得ることができれば，入力歌唱者の歌声を事前に収録する必要がなく，汎用的な活用が可能となる．これら 2 種類のモデルについては実験で評価を行う．

学習に用いるデータの歌唱者と，生成時に入力する歌声の歌唱者をあえて変えることで，ある歌唱者の発話間変動を別の歌唱者に対して転写するという拡張が考えられる．また，提案手法は，同一内容を複数回収録したデータがあれば学習が行えるため，歌声のみならず，発話音声にも応用できることが期待される．

4 実験的評価

4.1 繰り返し歌唱データベース

HTS [11] の歌声合成デモに含まれる童謡のうち 17 曲 (13 分 30 秒) を見本とし，男性 4 名が 5 回ずつ歌唱した in-house データベースを用いた．歌声収録は無響室で行った．歌唱者は，楽譜を見つつ，見本 (女声) およびメトロノームの音をヘッドホンで聴きながら，見本の 1 オクターブ下で歌唱した．音楽制作ソフトを用いたため，各テイクに関してタイミングの同期はとれているものとする．収録歌声のうち 14 曲 (一人当たり 12 分 6 秒) を学習に，3 曲 (一人当たり 1 分 24 秒) を評価に用いた．

4.2 歌声処理条件

歌声は収録時の 48 kHz から 16 kHz にダウンサンプリングして処理した．スペクトル包絡の抽出と音声波形合成には WORLD [12] を用い，F_0 の抽出には STRAIGHT [13] を用いた．フレームシフトは 5 ms とした．F_0 の無声区間は線形補間して連続 F_0 とした．

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 gated linear unit (GLU) [14] からなる隠れ層，input-to-output residual net [15] 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 (4) における $\boldsymbol{L}_{\boldsymbol{S}_{in}}$ 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 [16, 17] を用いて近似を行った．11 次元の入力ベクトルは，入力ピッチ系列の 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．入力として 5 テイク，ターゲットとして残りの 4 テイクを用いることで，20 パターンの組み合わせを用いた．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 ($\exp\{-||\boldsymbol{s}_{tgt}(i)-\boldsymbol{s}_{out}(j)||^2/\sigma^2\}$)．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・出力どちらも 0.1 とした．これらの値は実験的に定めた．入力とターゲットの 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．これらの値は，不自然な変動が生じないよう，実験的に定めた．また，GMMN の学習時にのみ，分析開始位置に 0 フレーム以上 47 フレーム以下のオフセットを加えるデータ拡張 [7, 8] を行った．

歌唱者依存モデルは，評価時の入力歌唱者の歌声を用いて学習し，不特定歌唱者モデルは，評価時の入力歌唱者以外の 3 人の歌唱者の歌声を用いて学習した．

4.3 評価方法

ADT，NDT (歌唱者依存モデル)，NDT (不特定歌唱者モデル)，DT の 4 手法に関して，多重録音感を主観評価した．それぞれの歌声は，以下のように作成した．

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入力歌声に重ねた．正弦波の周波数は 0.775 Hz，振幅は半音の 10%とした．これらのパラメータは [2] を参考に定めた．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．

NDT (歌唱者依存モデル: SD): 連続対数 F_0 系列を歌唱者依存モデル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入力歌声に重ねた．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．

NDT (不特定歌唱者モデル: SI): 連続対数 F_0 系列を不特定歌唱者モデル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入力歌声に重ねた．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．

DT: 同一歌唱者の別テイクを入力歌声に重ねた．別テイクは 3 dB 音量を下げて原音にミックスした．別テイクは，入力歌声と顕著に違いがあって重ねると不自然になるものは手作業で避けた．

4 手法とも，入力の歌声は同じテイクを用いた．評価を容易にするため，曲を short，long の 2 種類の時間長に手動で切り分けた．平均した時間長は，それぞれ 4.9 s，10.5 s であった．

評価はクラウドソーシングサービスのランサーズ [18] 上で行った．short，long のそれぞれに関して 100 人が，48 問 (歌唱者 4 人，4 手法の 16 組それぞれについて 3 サンプルずつをランダムな順序で混ぜたもの) を聴取し，mean opinion score (MOS) 方式で多重録音感 (2 回歌って重ねたような感じであるか) を 1 (非常にそう思わない) から 5 (非常にそう思う) の 5 段階スコアで評価した．

4.4 評価結果

Fig. 3, 4 に，結果を示す．short・long いずれの条件でも，多重録音感スコアは ADT で低く，2 種類の NDT は DT に近い．また，Table 1，2 に，全歌唱者統合データについて，ADT 以外の 3 手法に関して，データに対応のある対の平均の差の両側検定を

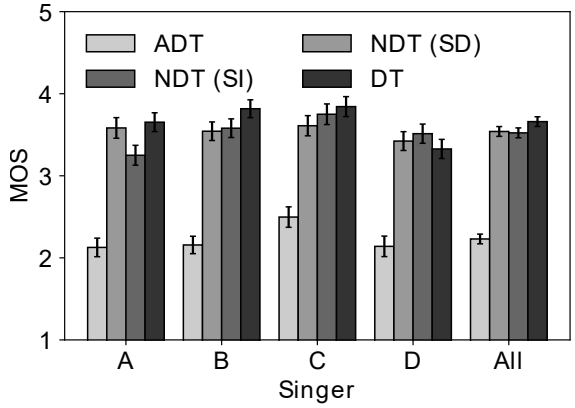

Fig. 3　時間長条件 short における手法・歌唱者ごとの多重録音感の MOS 値．エラーバーは 95%信頼区間．

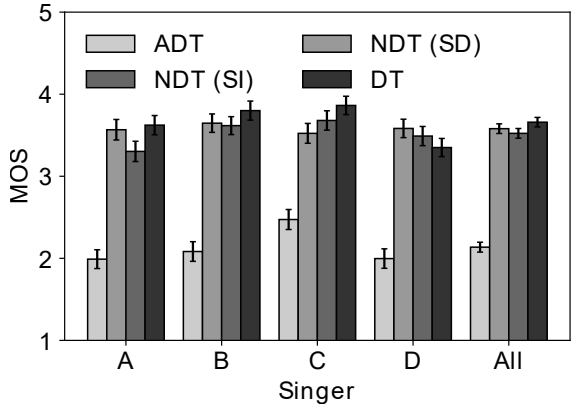

Fig. 4　時間長条件 long における手法・歌唱者ごとの多重録音感の MOS 値．エラーバーは 95%信頼区間．

行った結果を示す．時間長 short・long のいずれにおいても，2 種類の NDT 間に有意差は観測されなかった．また，DT と比べると 2 種類の NDT はスコアが有意水準 5%で有意に低かった．

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方が，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．また，不特定歌唱者モデルでも，歌唱者依存モデルと同等の多重録音感が得られることから，歌唱者に依存しない汎用的なモデルの構築が可能と言える．

5　まとめ

本稿では，自然歌声に対して多重録音感を付与する技術である NDT を提案した．

提案ポストフィルタは，GMMN を用いて，繰り返し歌唱データベースから歌声の MS の分布を学習し，発話間変動をモデル化する．次に，ポストフィルタを用いて，元の歌声にあたかも再度歌ったかのような自然なピッチの変動を付与し，その歌声を原音とミックスすることで，多重録音感を再現する．実験評価により，提案するポストフィルタを用いた NDT は，歌唱者依存モデル・不特定歌唱者モデルともに，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．また，不特定歌唱者モデルでも歌唱者依存モデルと有意差が観測されない程度の多重録音感が出せるため，ユーザごとにモデル学習を行う必要はなく，汎用性の高い手法であることが示された．

今後は，発話間変動や多重録音感の客観評価指標の検討，MS 以外の手法を用いたピッチ系列のモデリング，歌唱タイミングやスペクトルパラメータに関する発話間変動のモデリングを検討する．

Table 1　手法間の，対応のある平均の差の両側検定 (short 条件)

Compared methods	p 値
NDT (SD) & NDT (SI)	5.66×10^{-1}
NDT (SD) & DT	8.21×10^{-5}
NDT (SI) & DT	2.19×10^{-5}

Table 2　手法間の，対応のある平均の差の両側検定 (long 条件)

Compared methods	p 値
NDT (SD) & NDT (SI)	5.12×10^{-2}
NDT (SD) & DT	1.10×10^{-2}
NDT (SI) & DT	1.73×10^{-5}

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] J. Tolbert, *Take Control of Recording with GarageBand '09*, ser. Take Control. TidBITS Publishing, Incorporated, Sep. 2009.
[2] R. Izhaki, *Mixing Audio: Concepts, Practices, and Tools*. Taylor & Francis, Oct. 2017.
[3] K. Womack, *The Beatles Encyclopedia: Everything Fab Four*. ABC-CLIO, Jun. 2014.
[4] R. Brice, *Music Engineering*. Elsevier Science, Oct. 2001.
[5] Y. Li, K. Swersky, and R. Zemel, "Generative moment matching networks," in *Proc. ICML*, Lille, France, Jul. 2015, pp. 1718–1727.
[6] Y. Ren, J. Li, Y. Luo, and J. Zhu, "Conditional generative moment-matching networks," in *Proc. NIPS*, Barcelona, Spain, Dec. 2016, pp. 2928–2936.
[7] H. Tamaru, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, "Generative moment matching network-based random modulation post-filter for DNN-based singing voice synthesis and neural double-tracking," in *Proc. ICASSP*, Brighton, U.K., May 2019, pp. 7070–7074.
[8] 田丸浩気, 齋藤佑樹, 高道慎之介, 郡山知樹, and 猿渡洋, "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," in **音講論（春）**, no. 2-10-5, 東京, Mar. 2019.
[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] 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.
[11] "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/.
[12] 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.
[13] 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.
[14] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *arXiv*, vol. abs/1612.08083, 2016.
[15] 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.
[16] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. NIPS*, Vancouver, Canada, Dec. 2008, pp. 1177–1184.
[17] 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.
[18] "Lancers," https://www.lancers.jp/.