歌唱者間相互作用を再現するDNN重唱歌声合成の検討

兵藤弘明[1,a)]　高道慎之介[1,b)]　中村友彦[2,c)]　小口純矢[3,d)]　猿渡洋[1,e)]

概要：本研究では，歌声合成手法に歌唱者間相互作用を計算機で再現する機構を導入することで，より一体感のある重唱音声を合成する手法を検討する．既存の歌声合成手法の多くは独唱音声の合成を前提としており，人間の歌唱者間相互作用，すなわち他者の歌声に合わせて自身の歌声を調節することを考慮していない．歌唱者間相互作用を無視して合成された独唱音声から重唱音声を制作することは，重唱音声の一体感の損失につがなると考えられる．そこで本研究では，合成対象の声部の楽譜に加え他声部の楽譜も利用するアーキテクチャ，および歌声特徴量レベルの相互作用に相当する目的関数を導入することで，歌唱者間相互作用を計算機で再現する歌声合成手法を検討する．実験結果から，他声部の楽譜特徴量を利用する機構が歌声の F_0 予測精度の向上に有用であることや，声部間で音量のバランスを合わせる目的関数が重唱音声の一体感向上に寄与することを示す．

1. はじめに

声を用いて音楽を作り出す歌唱は，コミュニケーションや芸術表現の手段として多くの人に親しまれており，人間にとって不可欠なものである．歌唱は歌唱人数や声部の数によりいくつかの形態に分類される．楽曲が 1 つの声部により構成されるとき，歌唱者が 1 人である場合を独唱，2 人以上の歌唱者が存在する場合を斉唱という．また楽曲が複数の声部により構成されるとき，各声部の歌唱者が 1 人である場合は重唱，2 人以上の歌唱者が存在する場合を合唱という．斉唱・重唱・合唱において複数歌唱者が同時に歌唱する際，歌唱者は互いの歌声を聴きながら，歌声全体に一体感が生まれるように自身の歌声の発音や音高を調節する [1–3]．図 1 上に示すようなこの歌唱者間相互作用は，一体感のある歌唱表現を実現する上で非常に重要な要素である．

また，人間の歌唱音声を計算機により再現する歌声合成技術は，芸術作品や仮歌への使用など，様々な用途で使用されており [4,5]，一般に広く普及している．近年はディープニューラルネットワーク（deep neural network: DNN）に基づく歌声合成手法 [6–9] が登場し，非常に高い品質を持つ歌声の合成が可能となった．

しかしながら，既存の歌声合成手法の多くは独唱音声の合成を前提としており，人間の歌唱者間相互作用，すなわち他者の歌声に合わせて自身の歌声を調節することを考慮していない．従って，合成音声を用いて斉唱・重唱・合唱音声を制作する際は，歌唱者間相互作用を無視して合成された独唱音声を混合することとなり，歌声全体の一体感が損なわれる（図 1 左下）．

そこで本研究では重唱を例に挙げ，歌唱者間相互作用を計算機で再現する重唱歌声合成の検討，および歌唱者間相互作用が合成された歌声の聴感に与える影響の検証を行う（図 1 右下）．重唱音声を合成する際に，人間の歌唱者間相互作用に相当する効果を生じさせることは，歌声に一体感をもたらすだけでなく，人間が一体感を知覚する原理の解明にも繋がる．提案手法では，楽譜レベルと歌声レベルの歌唱者間相互作用の実現を試みる．前者について，従来は各声部の歌声を合成する際に当該声部の楽譜のみを用いるが，提案手法では当該声部の楽譜だけでなく，他声部の楽譜も用いる．後者については，各声部の合成歌声特徴量の間で定義される目的関数を DNN の学習時に導入する．以上の手法により，他声部と混合した際により強い一体感を生み出す独唱音声を合成する．提案手法により合成された独唱音声・重唱音声に対し評価実験を行い，歌唱音声の品質の評価，および歌唱者間相互作用が聴感に与える影響の検証を行う．

[1] 東京大学
Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan
[3] 明治大学
Meiji University
[a)] hyodo-hiroaki783@g.ecc.u-tokyo.ac.jp
[b)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
[c)] tomohiko.nakamura.jp@ieee.org
[d)] korguchi@gmail.com
[e)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

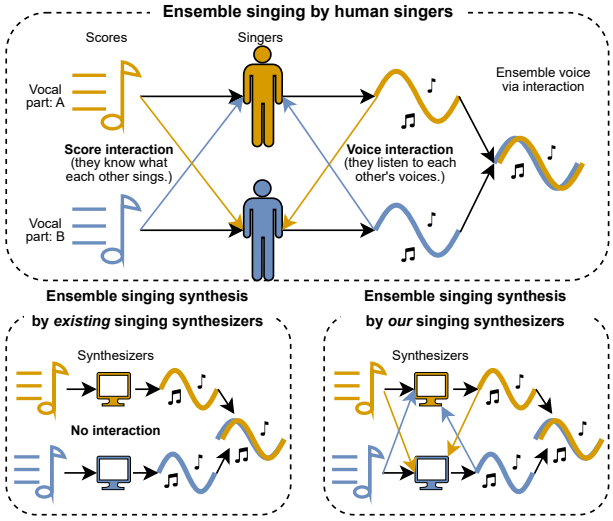

図 1 歌唱者間相互作用が生じる場合における歌唱者と歌声合成の比較．複数歌唱者が歌唱する場合，歌唱者間で相互作用が生じ一体感のある歌声が得られる．一方で既存の歌声合成技術ではそのような相互作用は生じない．本研究では，歌声合成間に相互作用機能をもたらし，一体感のある歌声の合成を目指す．

2. 関連研究

2.1 歌唱者間相互作用とその歌唱者意識・音響分析

歌唱者間相互作用が生じる際の歌唱者意識については，幾つかの指摘が存在する．例えば，発声技術訓練では「(他歌唱者の楽譜内容を知っている状態で，その歌唱者の声と混ざったときに）自分の声が消えるように歌う」「リズムをピッタリ揃える」 [3]，「全員の音量のバランスを取る」 [10]，「声を合わせる感覚の習得」 [11] などの指摘がある．

また，歌唱者間相互作用の効果が歌声特徴量に現れることも確認されている．Dai ら [2] は，他声部の歌声を聴きながら歌唱する場合と，他声部の歌声を聴かずに歌唱する場合の歌声を比較分析し，他声部の歌声の聴取に依って歌声の基本周波数（F_0）に変化が生じることを示している．Cuesta ら [1] は，特定の歌唱者の組み合わせにおいて，斉唱時の歌唱者間の F_0 が高く相関することを示している．また上原 [12] は，同じ音高で複数人が単母音を歌唱する場合の声色の変化について調査を行い，他歌唱者の歌声の有無により第 $1, 2$ フォルマントが変化することを明らかにしている．これらの研究は，重唱合成を実現する上で重要な知見となる．

2.2 斉唱・重唱らしさの人工付与

楽譜からの歌声合成ではないが，歌声加工により自然な斉唱・重唱音声を実現する方法については，いくつかの先行研究が存在する．Ternström [13] は，合成音声を用いて斉唱音声の聴感と F_0・フォルマント周波数の分散の関係を調査し，受聴者が好ましいと感じる歌声の F_0 の分散の傾向と，歌唱する母音や声部が主観評価値に与える影響を示している．山内ら [14] は重唱音声について分析を行い，歌唱者ペアの歌声の調和と歌唱者間のメルケプストラム歪みの関係を示している．また，勝瑞ら [15] と宮沢ら [16] はそれぞれ，斉唱音声について歌唱者間の F_0 差分および変調スペクトルに着目した加工の有効性を示している．これらの研究では，他声部から独立に合成された独唱音声を足し合わせることで制作した斉唱音声・重唱音声が使用されているため，歌唱者間相互作用と歌声特徴量の関係は不明である．しかしながら，これらの研究で言及されている歌声レベルの分析結果は本研究で活用できる可能性がある．

2.3 歌声の一体感に関する主観評価指標

合成された重唱音声を主観評価するには，その評価軸を策定する必要がある．上原 [12] は，重唱の一体感（声が合っているときの聴感）を表す言葉について大規模な調査を実施している．一体感と対応する 6 つの指標（まとまり，ピッチ，呼吸，響き，声質，母音）を策定し，それぞれの評価軸に関する主観評価実験を実施している．本指標は歌声特徴量との相関も明らかであり見通しの良いものであるため，本稿の主観評価もこの指標に従う．

2.4 歌声合成技術

近年の歌声合成手法は，楽譜特徴量を入力，楽譜特徴量に対応する歌声を出力として，入力と出力の関係を DNN により学習するものが主流である [6,7]．本稿では，一般的な time-lag, duration-aware な機構を採用する [6]．本機構は，楽譜解析モデル，楽譜の音素開始時刻からのズレを予測する time-lag モデル，音素継続長を予測する duration モデル，歌声特徴量を予測する音響モデル，そして歌声特徴量から歌声波形を合成するボコーダから成る．なお，より最先端の独唱合成の枠組みとして，end-to-end 機構 [17] や自己教師あり学習の特徴量の使用 [18] が考えられるが，本稿ではこれらを採用しない．これは，end-to-end 機構や自己教師あり学習特徴量はその推論と表現がブラックボックス的であり，これまでに述べた重唱歌声分析の知見と関連付けることが難しいためである．本稿では，明示的な推論構造とパラメトリックな特徴量表現を用いることで，知見を活用したモデル設計と評価を行う．

3. 提案手法

本節では，歌唱者間相互作用に基づいた重唱歌声合成手法を提案する．提案手法ではまず，ある声部の歌声特徴量を音響モデルで予測する際に，当該声部の楽譜に加え他声部の楽譜も利用するアーキテクチャを提案する．これは，歌唱者同士が互いの楽譜内容を把握することに対応する．歌声特徴量としてメルケプストラム係数，連続対数 F_0，帯域非周期性指標，有声／無声フラグを用いる．次に，複数声部の歌声特徴量間で，特徴量レベルの相互作用に相当す

る目的関数を導入する．これは，歌唱者同士が互いの音声を聴きながら自らの歌声を調節することに対応する．

以降では，声部数を 2（声部 A，声部 B）と仮定して提案手法の内容を述べる．ただし 3.5 節で述べるように，提案手法は 3 声部以上に容易に拡張可能である．

3.1 前処理：データ分割

DNN 歌声合成の学習において，学習データの 1 曲全体を一度にミニバッチに載せることは GPU メモリ量の観点から現実的ではない．そこで歌声合成では，無音フレームを検出しその時刻で楽譜特徴量と歌声特徴量を分割する．通常の DNN 歌声合成のように各声部を独立に学習する場合には，このデータ分割方法を採ればよい．しかしながら，声部間に相互作用をもたらすには，声部間で時刻同期した分割を行わなければならない．そこで本稿では，2 声部の両方が無音フレームである時刻において分割することで，声部間で時刻同期した楽譜特徴量・歌声特徴量をそれぞれ得る．以降では，声部 A，B の楽譜特徴量系列をそれぞれ $\boldsymbol{x}^{(\mathrm{A})} := \{\boldsymbol{x}^{(\mathrm{A})}[n]\}_{n=0}^{N^{(\mathrm{A})}-1}, \boldsymbol{x}^{(\mathrm{B})} := \{\boldsymbol{x}^{(\mathrm{B})}[n]\}_{n=0}^{N^{(\mathrm{B})}-1}$ とする．ここで，$N^{(i)}$ は声部 $i \in \{\mathrm{A}, \mathrm{B}\}$ の楽譜特徴量系列長である．また，声部 i の F_0 系列とメルケプストラム系列をそれぞれ $\boldsymbol{y}_{\mathrm{F0}}^{(i)} := \{y_{\mathrm{F0}}^{(i)}[t]\}_{t=0}^{T-1}, \boldsymbol{y}_{\mathrm{mgc}}^{(i)} := \{\boldsymbol{y}_{\mathrm{mgc}}^{(i)}[t]\}_{t=0}^{T-1}$ と定義する．ここで，T は F_0 またはメルケプストラム系列長である．一般に $N^{(\mathrm{A})}, N^{(\mathrm{B})}$ は異なるため，本稿では長い方の系列長と一致するように短い方の系列を zero padding する[*1]．この操作で得られた声部 i の楽譜特徴量を $\tilde{\boldsymbol{x}}^{(i)}$ とする．

3.2 アーキテクチャ

提案手法のアーキテクチャを図 2 に示す．アーキテクチャは 2 並列構造であり，各構造は各声部に対応する．本稿では time-lag モデル，duration モデル，音響モデルに限り相互作用機能を持たせ，それ以外のモデル（楽譜解析モデル，ボコーダ）は従来手法と同様に各声部で独立に学習・推論する．相互作用機能を持たせるモデルは多歌唱者モデルとし，歌唱者ベクトルを別途入力する．

Time-lag モデル，duration モデルについては，$\tilde{\boldsymbol{x}}^{(\mathrm{A})}$ と $\tilde{\boldsymbol{x}}^{(\mathrm{B})}$ を特徴量方向に結合した系列を入力とし，それぞれ time-lag フレーム数と音素継続長を予測する．予測した time-lag フレーム数と音素継続長に従って時間方向に楽譜特徴量を展開し，音響モデルに入力する．

音響モデルについては，はじめにフレームレベルに展開された楽譜特徴量系列を対数 F_0 モデルに入力として与え，対数 F_0 系列を予測する．次に，フレームレベルに展開された楽譜特徴量，対数 F_0 系列をメルケプストラム係数モデル，帯域非周期性指標モデル，有声／無声フラグモデルに入力として与え，この 3 つの歌声特徴量を予測する．本稿では，音響モデルとして LSTM（long short term memory）ベース音響モデル，拡散モデルベース音響モデルの 2 種類を使用する．

[*1] 本来は，$\boldsymbol{x}^{(A)}$ と $\boldsymbol{x}^{(B)}$ の間でアライメントをとり，系列間で対応する要素（直感的には，楽譜上で同じ時刻に対応する音符）が同じ系列インデックスに配されるよう padding するべきである．しかしながら本稿では実装の簡便さを優先し，zero padding を選択した．Padding の方法については今後検討する．

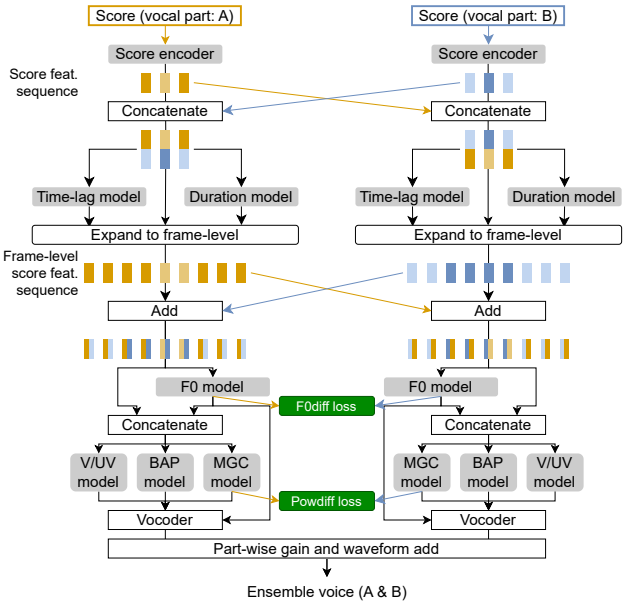

図 2 提案手法のアーキテクチャと目的関数．実際には各モデルは多歌唱者モデルであり，歌唱者ベクトルも入力とするが本図では省略している．また，従来手法で用いられる目的関数（各声部における ground-truth との距離関数）も，提案手法の学習に使用するが本図では省略している．

3.2.1 LSTM ベース音響モデル

LSTM ベース音響モデルでは，対数 F_0 モデル，メルケプストラム係数モデル，帯域非周期性指標モデル，および有声／無声フラグモデルとして，双方向 LSTM [6] を使用するモデルである．

3.2.2 拡散モデルベース音響モデル

拡散モデルベース音響モデルでは，メルケプストラム係数モデルと帯域非周期性指標モデルとして拡散モデル [7] を，対数 F_0 モデルと有声／無声フラグモデルとして LSTM ベース音響モデルと同様の LSTM [6] を使用するモデルである．

3.3 学習の目的関数

Time-lag モデル，duration モデルについては既存手法 [6] と同様に，各声部について ground-truth の値と予測値の間の L2 ロスを計算する．

LSTM ベース音響モデルでは，ground-truth の値と予測値の間の L_1 ロスを計算する．拡散モデルベース音響モデルでは，ground-truth に足し合わせるノイズ値と，音響モデルに ground-truth とノイズを足し合わせたものを入力

して予測したノイズ値の間の L_1 ロスを計算する [7]．声部 i に関する予測 F_0 系列および予測メルケプストラム系列を $\hat{\boldsymbol{y}}^{(i)}_{\mathrm{F0}} := \{\hat{y}^{(i)}_{\mathrm{F0}}[t]\}_{t=0}^{T-1}, \hat{\boldsymbol{y}}^{(i)}_{\mathrm{mgc}} := \{\hat{\boldsymbol{y}}^{(i)}_{\mathrm{mgc}}[t]\}_{t=0}^{T-1}$ とすると，対数 F_0 モデルとメルケプストラム係数モデルに関するロス関数 $\mathcal{L}_{\mathrm{F0}}$ と $\mathcal{L}_{\mathrm{mgc}}$ は以下のように書ける．

$$\mathcal{L}_{\mathrm{F0}} = \sum_{i\in\{\mathrm{A,B}\}}\sum_{t=0}^{T-1}\left|y^{(i)}_{\mathrm{F0}}[t] - \hat{y}^{(i)}_{\mathrm{F0}}[t]\right| \tag{1}$$

$$\mathcal{L}_{\mathrm{mgc}} = \sum_{i\in\{\mathrm{A,B}\}}\sum_{t=0}^{T-1}\left\|\boldsymbol{y}^{(i)}_{\mathrm{mgc}}[t] - \hat{\boldsymbol{y}}^{(i)}_{\mathrm{mgc}}[t]\right\|_1 \tag{2}$$

ここで，$\|\cdot\|_1$ は L_1 ノルムを表す．本稿では省略するが，非周期性指標モデルと有声／無声ラベルモデルについても同様にロス関数を定義する．

以降では，2 節の内容を元に，音響モデルの声部 A，B の間で定義される，新たな目的関数を導入する．

F_0 差分ロス： 2 節で説明したように，歌唱者間の F_0 差分には相互作用の影響が生じる．そこで，2 つの F_0 系列の差について，ground-truth 値と予測値の間で L_1 ロスを計算する．声部 i において有声であるフレームインデックスの集合を $\mathcal{V}^{(i)}$ と書くと，この目的関数は以下のように書ける．

$$\mathcal{L}_{\mathrm{F0diff}} = \sum_{t\in\mathcal{V}^{(\mathrm{A})}\cap\mathcal{V}^{(\mathrm{B})}}\left|\Delta_{\mathrm{F0}}[t] - \hat{\Delta}_{\mathrm{F0}}[t]\right| \tag{3}$$

$$\Delta_{\mathrm{F0}}[t] := y^{(\mathrm{A})}_{\mathrm{F0}}[t] - y^{(\mathrm{B})}_{\mathrm{F0}}[t] \tag{4}$$

$$\hat{\Delta}_{\mathrm{F0}}[t] := \hat{y}^{(\mathrm{A})}_{\mathrm{F0}}[t] - \hat{y}^{(\mathrm{B})}_{\mathrm{F0}}[t] \tag{5}$$

これは，声部間の F_0 差分を ground-truth の値に近づける効果を持つ．

パワー差分ロス： 2 節で説明したように，パワーについても相互作用の影響が生じると考えられる．そこで，パワー差分についても歌唱者間で整合させる．その目的関数は，

$$\mathcal{L}_{\mathrm{Powdiff}} = \sum_{t\in\mathcal{V}^{(\mathrm{A})}\cap\mathcal{V}^{(\mathrm{B})}}\left|\Delta_{\mathrm{mgc}}[t,0] - \hat{\Delta}_{\mathrm{mgc}}[t,0]\right| \tag{6}$$

$$\Delta_{\mathrm{mgc}}[t,0] := y^{(\mathrm{A})}_{\mathrm{mgc}}[t,0] - y^{(\mathrm{B})}_{\mathrm{mgc}}[t,0] \tag{7}$$

$$\hat{\Delta}_{\mathrm{mgc}}[t,0] := \hat{y}^{(\mathrm{A})}_{\mathrm{mgc}}[t,0] - \hat{y}^{(\mathrm{B})}_{\mathrm{mgc}}[t,0] \tag{8}$$

で定義される．ここで，$y^{(i)}_{\mathrm{mgc}}[t,0]$ と $y^{(i)}_{\mathrm{mgc}}[t,0]$ は，それぞれ声部 i の時刻 t，0 次元目の ground-truth メルケプストラム成分，予測メルケプストラム成分を表す．上式は 0 次元目のメルケプストラムについて，予測値差分と ground-truth 値差分を近づける効果を持つ．

提案手法の最終的なロス関数 $\mathcal{L}$ は以下のように書ける．

$$\mathcal{L} = \mathcal{L}_{\mathrm{F0}} + \mathcal{L}_{\mathrm{mgc}} + w_{\mathrm{F0diff}}\mathcal{L}_{\mathrm{F0diff}} + w_{\mathrm{Powdiff}}\mathcal{L}_{\mathrm{Powdiff}} \tag{9}$$

ここで $w_{\mathrm{F0diff}}, w_{\mathrm{Powdiff}}$ はそれぞれ $\mathcal{L}_{\mathrm{F0diff}}, \mathcal{L}_{\mathrm{Powdiff}}$ に対する重みを表す．

3.4 推論と重唱音声の生成

各声部の歌声波形は各モデルを順に推論して得られる．はじめに，各声部の楽譜データから得た 2 声部の楽譜特徴量を結合したものを time-lag モデル，duration モデルに入力したのち，楽譜特徴量をフレームレベルに展開する．次に，各声部のフレームレベル楽譜特徴量を加算したものを音響モデルに入力し，各声部の歌声特徴量を予測する．ボコーダを用いてこれらを各声部の独唱音声波形に変換する．最後に，各声部の独唱音声波形を重み付き加算することで．最終的な重唱音声波形を得る．

3.5 3 声部以上への拡張可能性について

重唱は 3 声部以上から構成されることもしばしばあり，アカペラでは 6 声部から構成されることもある [19]．そのため，2 声部に限って議論した提案手法の内容を 3 声部以上に拡張する方法を述べる．以降では，声部数を $N > 2$ としてその拡張可能性について述べる．なお，以降の実験的評価では 2 声部のみを扱う．

最も単純な方法は，並列数を声部数 N だけ増やすことである．すなわち，各声部の音響モデルは当該声部以外の $N-1$ 声部の楽譜特徴量の加算結果を入力とし，目的関数には $N(N-1)/2$ 個の項が加わる．この方法には，N 声部のモデルを全て GPU メモリ上に載せなければならないことと，目的関数の組み合わせ数が増加する懸念がある．一方で，実装は容易であり，学習と推論で一貫した枠組みとなる．

もうひとつの方法は，声部の乱択である．すなわち，N 声部のうち 2 声部をミニバッチ毎にランダムに選び，2 声部に対応したアーキテクチャと目的関数で学習する．この方法の懸念は，乱択による学習の不安定性（例えば，明らかに相互作用のない 2 声部を選んでも提案手法の効果は見込めないこと）と，学習と推論で一貫しないこと（学習は 2 声部毎に行うが推論は N 声部で行うこと）である．しかしながら，N が増加しようとも固定の GPU メモリ量と目的関数で学習可能であるというメリットがある．これに類似する方法として，主たる声部（例えばメインボーカル）が明らかである場合，主たる声部を常に選択しもう 1 つの声部を乱択する方法も考えられる．

4. 実験的評価

4.1 実験条件

本実験では，データセットとして「波音リツ」歌声データベース Ver. 2（計 261 分）[20] および日本語のアカペラ重唱曲からなる jaCappella コーパス [21] の一部を用いた．従来手法（各声部で独立に学習を行う手法）および提案手法の学習には音素アライメントが必要であるが，jaCappella コーパスはこれを含まない．そこで，一部のデータに対し手動でアノテーションを行い，音素アライメントを作成し

た．本実験では手動アノテーションを行った 2 名分の歌唱者 (歌唱者 ID: S1, Vo1) のデータ（計 23 分）を用いた．S1 はソプラノ声部, Vo1 はメインボーカル声部の歌唱者である．

以上のデータを，学習／検証／評価用データに分割して使用した．評価用データは，手動アノテーションを行った jaCappella コーパスからランダムに選択し，残りのデータをランダムに学習／検証用データに分割した．分割は楽曲単位で行い，学習／検証／評価用データの楽曲数の比が 8 : 1 : 1 となるように行った．評価実験では，評価用データに含まれる楽曲のうち，本実験で用いた jaCappella コーパス [21] の 2 名の歌唱者（歌唱者 ID: S1, Vo1）のどちらも歌唱している楽曲（七つの子，雪）を使用した．

全ての音声のサンプリング周波数は 48000 Hz, 音声分析窓のシフト長は 5 ms である．また，歌声特徴量はメルケプストラム係数（60 次元），連続対数 F_0（1 次元），帯域非周期性指標（5 次元），有声／無声フラグ（1 次元）から成り，WORLD [22] (D4C edition [23]) を用いて学習データから抽出したものを使用し，歌唱者ベクトルの次元数は 256 とした．

楽譜解析モデルでは，NEUTRINO [9] を用いて MusicXML [24] 形式の楽譜データをフルコンテキストラベルに変換し，これを score encoder により楽譜特徴量に変換した．ルールベースで変換を行うためモデルの学習は不要である．提案手法は NNSVS [25] に準拠して実装を行った．また，ボコーダには WORLD [22] を用いた．学習時には Adam optimizer [26] を用いて学習パラメータの最適化を行った ($\alpha = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$)．

4.2 音響モデルの比較

重唱音声を評価する前に，自然性の高い独唱音声を合成できる音響モデルを決定するため，独唱音声の自然性に関する主観評価実験を実施した．比較した手法は，3.2.1 節と 3.2.2 項で述べた 2 種類であり，ラベルを以下のように定義する．ただし，提案手法のアーキテクチャ及び目的関数を使用せず，従来手法の構造に歌唱者ベクトルを加えたアーキテクチャ，および従来手法と同様の目的関数を用いて学習を行ったことに注意する．

- **LSTM**: NNSVS [25] の MultistreamSeparateF0ParametricModel に基づいた音響モデル．ハイパーパラメータは NNSVS レシピの既定値[*2]に従った．
- **Diffusion**: NNSVS [25] の NPSSMDNMultistreamParametricModel に基づいた音響モデル．ハイパーパラメータは NNSVS レシピの既定値[*3]に従った．

[*2] https://github.com/nnsvs/nnsvs/blob/master/recipes/amaboshi_cipher_utagoe_db/dev-48k-world/config.yaml
[*3] https://github.com/nnsvs/nnsvs/blob/master/recipes/namine_ritsu_utagoe_db/dev-48k-world/config.yaml

表 1 独唱音声の自然性に関する MOS 値（95%信頼区間）

Method	MOS(↑)
LSTM	2.82 ± 0.15
Diffusion	2.56 ± 0.15

合成した独唱音声に対し，独唱音声の自然性に関する 5 段階の mean opinion score (MOS) 評価実験を行った．被験者は Lancers[*4]で募集したクラウドワーカ 30 名であり，各被験者は 12 サンプルの音声を評価した．各サンプルは，評価用データを 3.1 節の方法で分割したものであり，平均時間長は 5.11 秒である．

表 1 に評価結果を示す．2 手法の間における MOS 値の有意性を調査するため，t 検定を行った．はじめに 2 手法の MOS 値の等分散性を確認するため，有意水準を 0.05、帰無仮説を「2 手法の各サンプルの評価値の分散に差がないこと」として F 検定を行った．得られた p 値は 0.58 であり有意水準を上回ったことから，不等分散でないことが示された．

次に等分散性を仮定し，有意水準を 0.05、帰無仮説を「2 手法の MOS 値に差がないこと」として Student の t 検定を行った．得られた p 値は 0.017 であり．有意水準を下回ったことから MOS 値に有意差があることが示された．

以上より，音響モデルとして LSTM を用いる方が拡散モデルを用いる場合より自然性の高い独唱音声を合成できることが確認できた．従って以降の実験では音響モデルとして LSTM を使用した．

4.3 独唱音声の品質に関する客観評価

次に，提案手法を用いて合成した各声部の独唱音声について客観評価実験を実施した．比較手法は以下の通りである．なお，モデルと学習のハイパーパラメータは全手法で共通であり，学習の反復回数は 100 epoch とした．

- **Baseline**: 4.2 節の LSTM と同一のモデル．
- **MT**: 3.2 節に記載のモデル（MultiTrack）．F_0 差分ロスおよびパワー差分ロスは使用しない．
- **MT+F0diff**: 3.2 節に記載のモデルに，F_0 差分ロスを導入したもの．
- **MT+Powdiff**: 3.2 節に記載のモデルに，パワー差分ロスを導入したもの．
- **MT+F0diff+Powdiff**: 3.2 節に記載のモデルに，F_0 差分ロス，パワー差分ロスを導入したもの．

提案手法の目的関数の重みは全て 1.0 とした．客観評価指標にはメルケプストラム歪み（mel-cepstral distortion: MCD）[dB]，対数 $F0$ 二乗平均平方根誤差（RMSEs of log $F0$: F0-RMSE）を用いた．

表 2 に評価結果を示す．

[*4] https://www.lancers.jp/

表 2 独唱音声の自然性に関する客観評価値.

Method	MCD [dB] (↓)	LF0-RMSE (↓)
Baseline	10.58	0.12
MT	10.82	0.10
MT+F0diff	11.02	0.13
MT+Powdiff	10.94	0.10
MT+F0diff+Powdiff	10.95	0.11

メルケプストラム歪みについて，MT は Baseline と比較して悪化する．これは，提案手法のアーキテクチャの導入がメルケプストラム歪みの悪化につながることを示す．この原因として，音響モデルに入力する特徴量の次元数の増加により，メルケプストラムの推論が困難になった可能性が考えられる．

また，声部間の目的関数を導入した 3 つの手法（MT+F0diff, MT+Powdiff, MT+F0diff+Powdiff）は, MT と比較して悪化する．これは，目的関数の導入がメルケプストラム歪みの悪化につながることを示す．声部間の目的関数は重唱音声の一体感向上を意図した目的関数であり，独唱音声の品質向上につながる要素を持たないことから，独唱音声の歪みが悪化することは自然であると考えられる．

LF0-RMSE について，MT は Baseline と比較して改善が見られた．このことから，重唱の各声部の独唱音声を合成する際に，他の声部の情報が有用である可能性が考えられる．また，MT+F0diff は MT と，MT+F0diff+Powdiff は MT+Powdiff と比較して悪化する．これは，F_0 差分ロスの導入が LF0-RMSE の悪化につながることを示す．この原因として，他声部と音程を合わせようとした結果，各声部において正しい音高から外れた可能性が考えられる．

4.4 重唱音声の一体感に関する主観評価

最後に，提案手法を用いて合成した重唱音声の一体感に関する主観評価実験を実施した．使用する手法とモデルは 4.3 節と同様である．

合成した独唱音声を足し合わせることで得た重唱音声に対し，歌声のまとまり（ハーモニーが心地よく聴こえるか，全体がまとまって聴こえるか）に関する 5 段階の MOS 評価実験を行った．なお，重唱音声の一体感については 2 に述べたようにいくつかの評価基準 [12] があるが，本稿ではこれらの評価基準の中で，被験者にとって最も評価が容易であると考えられるまとまりに注目して評価を行った．独唱音声を足し合わせる際は，メインボーカル声部／ソプラノ声部の独唱音声波形の絶対値の最大値の比が $1:\frac{2}{3}$ となるように重み付けを行った．この設定において，どちらの声部の歌声も十分に聴取できることを確認した．

被験者の募集方法，被験者あたりの評価サンプル数，評価用データの分割方法は 4.3 節と同様である．各サンプルの平均時間長は 5.34 秒である．

表 3 に評価結果を示す．重唱音声のまとまりに関する MOS 値について，F_0 差分ロスを導入した 2 つの手法（MT+F0diff，MT+F0diff+Powdiff）は，他の手法と比べてまとまりに関する MOS 評価値が悪化していた．特に MT+F0diff は，F0-RMSE についても Baseline からの悪化が見られた．原因として，歌唱者間相互作用により他声部と音程を合わせようとした結果，各声部において独唱音声として正しい音高から外れてしまい，結果的に重唱音声の一体感を損ねることにつながった可能性が考えられる．また，その他のモデルについては有意差は見られなかった．この原因として，手法間の聴感の変化が僅かであることから, 細かな聴感の違いの評価が困難であった可能性が考えられる．なお有意差は見られなかったものの，MT+Powdiff は MT と，MT+F0diff+Powdiff は MT+F0diff と比較して MOS 値の改善が見られた．これは，パワー差分ロスの導入が重唱音声のまとまりの改善につながることを示す．これは，声部間で音量のバランスを合わせることが，重唱音声の一体感の向上に寄与することを示唆する．

表 3 重唱音声のまとまりに関する MOS 値（95%信頼区間）

Method	MOS (↑)
Baseline	3.13 ± 0.19
MT	3.08 ± 0.20
MT+F0diff	2.60 ± 0.21
MT+Powdiff	3.16 ± 0.19
MT+F0diff+Powdiff	2.95 ± 0.18

5. おわりに

本研究では，深層学習に基づく歌声合成手法に歌唱者間相互作用を意図する機構を組み込むことで，より一体感のある重唱音声を合成する手法を検討した．その結果，ある声部の歌声を合成する際，他声部の情報が歌声の音高制御性の向上に有用であることや，声部間で音量のバランスを合わせることが重唱音声の一体感向上に寄与することが示唆された．一方で，他声部と音程を合わせる際に各声部の正しい音高から外れてしまう可能性が示唆されたことなどから，歌唱者間相互作用を意図する機構にはさらなる改善の余地があると考えられる．今後の展望としては，歌唱者間相互作用を意図する機構の改善や，細かな聴感の違いを評価するための実験設定の検討が考えられる．

謝辞：アノテーションの方法について，西山陽子様から多くの助言を受けた．本研究は JST 創発的研究支援事業 JPMJFR226V，JSPS 科研費 23H03418，23K18474 の助成を受けた．

参考文献

[1] H. Cuesta, E. Gómez Gutiérrez, A. Martorell Domínguez, and F. Loáiciga, “Analysis of intonation in unison choir singing,” in *Proceedings of the International Conference on Music Perception and Cognition(ICMPC)*, 2018.

[2] J. Dai and S. Dixon, “Analysis of interactive intonation in unaccompanied satb ensembles.” in *Proc. ISMIR*, 2017, pp. 599–605.
[3] あっしー, **アカペラ・パーフェクト・ブック 練習方法・楽譜アレンジ・ボイパこれ１冊！** ドレミ楽譜出版社, 2012.
[4] 剣持秀紀, “音楽情報処理技術の最前線: 3. 歌声合成とその応用,” **情報処理**, vol. 50, no. 8, pp. 723–728, 2009.
[5] 藤本健, “AI 歌声合成が歌手を超えた’22 年。「Synthesizer V」の進化に驚愕した by 藤本健,” accessed on January 27, 2024. [Online]. Available: https://av.watch.impress.co.jp/docs/topic/pb2022/1466374.html
[6] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Sinsy: A deep neural network-based singing voice synthesis system,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2803–2815, 2021.
[7] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, “DiffSinger: Singing voice synthesis via shallow diffusion mechanism,” in *Proc. AAAI*, vol. 36, no. 10, 2022, pp. 11 020–11 028.
[8] “CeVIO,” accessed on January 27, 2024. [Online]. Available: https://cevio.jp/
[9] “NEUTRINO,” accessed on January 27, 2024. [Online]. Available: https://studio-neutrino.com/
[10] あっしー, **アカペラ・パーフェクト・ブック〜アドバンス〜 アカペラ！アレンジ！ボイパ！さらに上達する１００のコツ**. ドレミ楽譜出版社, 2015.
[11] 河秀哉, **うたごえの戦後史**. 人文書院, 2016.
[12] 上原崇寛, “複数人歌唱・同一音高での 「声を合わせること」に関する音響特徴量の振る舞いと聴感の関係,” Ph.D. dissertation, 東京藝術大学, 2022.
[13] S. Ternström, “Perceptual evaluations of voice scatter in unison choir sounds,” *Journal of Voice*, vol. 7, no. 2, pp. 129–135, 1993.
[14] 山内孔貴, 須田仁志, 齋藤大輔, and 峯松信明, “ソースフィルタ分解に基づく複数歌唱者の調和制御に関する検討,” **研究報告音声言語情報処理** *(SLP)*, vol. 2020, no. 35, pp. 1–6, 2020.
[15] 勝瑞雄介, 齋藤大輔, and 峯松信明, “自然な斉唱音声合成のための複数歌唱者の基本周波数パターン制御に関する検討,” **研究報告音楽情報科学** *(MUS)*, vol. 2021, no. 11, pp. 1–7, 2021.
[16] 宮沢宙, 菊地晏南, 齋藤大輔, and 峯松信明, “音響特徴量系列の変調に基づいた斉唱音声合成の検討,” **研究報告音声言語情報処理** *(SLP)*, vol. 2023, no. 53, pp. 1–6, 2023.
[17] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, “Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis,” in *Proc. ICASSP*. IEEE, 2022, pp. 7237–7241.
[18] T. Jayashankar, J. Wu, L. Sari, D. Kant, V. Manohar, and Q. He, “Self-supervised representations for singing voice conversion,” in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.
[19] 中村友彦, 高道慎之介, 丹治尚子, 深山覚, and 猿渡洋, “jaCappella コーパス: 重唱分離・合成に向けた日本語アカペラ歌唱コーパス,” **日本音響学会研究発表会講演論文集** *(CD-ROM)*, vol. 2022, pp. 3–26, 2022.
[20] “音源_カノンの落ちる城,” accessed on February 01, 2024. [Online]. Available: https://www.canon-voice.com/voicebanks/
[21] T. Nakamura, S. Takamichi, N. Tanji, S. Fukayama, and H. Saruwatari, “jaCappella corpus: A Japanese a cappella vocal ensemble corpus,” in *Proc. ICASSP*. IEEE, 2023.
[22] 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. 99, no. 7, pp. 1877–1884, 2016.
[23] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, vol. 84, pp. 57–65, 2016.
[24] “MusicXML for exchanging digital sheet music,” accessed on February 1, 2024. [Online]. Available: https://www.musicxml.com/
[25] R. Yamamoto, R. Yoneyama, and T. Toda, “NNSVS: A neural network-based singing voice synthesis toolkit,” in *Proc. ICASSP*. IEEE, 2023.
[26] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.