Vol.2021-MUS-131 No.11
Vol.2021-SLP-137 No.11
2021/6/18

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

勝瑞　雄介[1,a)]　齋藤　大輔[1,b)]　峯松　信明[1,c)]

概要：複数人の歌唱における歌唱者間の音程が知覚に与える影響は，合唱に関しての調査はあるが，斉唱に関しては少ない．そこで本研究では音程の定常成分，変動成分の 2 点から合成斉唱音声の F0 系列のモデルを提案した．モデルに基づいて合成した音声の評価を行った結果，定常的な歌唱者間音程が全くない場合には音声の自然性が低くなることがわかった．

キーワード：斉唱，音声合成，基本周波数

1. はじめに

音楽は人間の文化活動の中でも古くから行われてきたものの一つで，芸術的価値のある作品としての音楽だけでなく，自分の感情を表現したり，聴くことで感情をコントロールしたりするための音楽などがあり，多くの用途で使われてきた表現様式である．そのなかでも歌唱は楽器などの道具を必要とせず，広く親しまれている．一口に歌唱と言ってもその形態はいくつかあり，歌唱者の数やパートの分かれ方によってそれぞれ呼称が異なる．本研究で対象とする斉唱は複数人による歌唱のうち，全員が同じパートを歌うという形態で，パート分けがある歌唱の場合には合唱もしくは重唱と呼ばれる．合唱も斉唱も古くから親しまれてきたが，近年の日本のポピュラー音楽の場面ではグループアーティストによる斉唱が多く見られ，現代でも多くの人間が斉唱に触れて暮らしている．

音声合成技術の進歩により，VOCALOID[1] を始めとする歌声合成システムが一般に使用されるようになり，多くの作品が公開されている．その作品の多くは人間の歌唱のような自然性は低く，歌唱の「上手さ」を押し出しているものではない．近年はカラオケで採点機能を使用した歌唱なども行われるようになり，歌唱の「上手さ」に人々の関心が向いていると言える．歌唱における「上手さ」に対する研究は以前から行われており，歌唱者が一人である独唱については片岡らによって音高，音長，音量に対して調査が行われている [2]．複数人の歌唱の分析は歌唱者間の相互作用により複雑になることが考えられるが，同様に調査されるべき課題である．音響特徴量の中でも，音高に対応する基本周波数は歌唱の中でメロディーに密接に関わる部分であり，歌唱において重要な要素である．人間は自分の声の基本周波数を完全に制御できるわけではない．その一方で，基本周波数が本来の音高から大きくずれている歌唱は一般的に，「上手な」歌唱とはみなされない．このことからも基本周波数は歌唱の「上手さ」に影響を与えていると考えられる．特に，複数人で歌唱する場合には歌唱者ごとに異なる基本周波数のずれを持って歌うことになり，一人で歌唱する独唱のときよりも聴取者が基本周波数のずれを認識しやすくなると考えられる．斉唱で歌唱者間の基本周波数にずれが生じるのは必然であり，知覚できないほどずれが小さい歌唱では斉唱らしさが失われると考えられる．そこで，「上手な」斉唱になるように斉唱のモデル化をすることで，より「上手な」斉唱音声の合成が可能になると期待される．

本研究では基本周波数を制御した斉唱音声を合成し，聴取実験を行うことで歌唱者間の音程が音声の自然性に与える影響を定量的に調査した．歌唱者間の音程も，歌唱全体を通してみられる定常的な音程と，引き込み現象を考慮した動的な音程の 2 つの観点から考察した．

2. 関連研究

本研究で対象とする，複数の歌唱者が全員同じ音高で歌う形態である斉唱に関して上手さや自然性についての調査は少ないが，異なる音高のパートを 1 名ずつで歌う重唱や，異なる音高のパートをそれぞれ複数人で歌う合唱を対象と

[1] 東京大学
[a)] shozui@gavo.t.u-tokyo.ac.jp
[b)] dsk_saito@gavo.t.u-tokyo.ac.jp
[c)] mine@gavo.t.u-tokyo.ac.jp

した研究はいくつかされている．野田らは複数人で同時に歌唱した際に他の歌唱者の音声に合わせてタイミングや基本周波数がつられる，「引き込み」と呼ばれる現象について研究している [3]．歌唱の訓練を行った歌唱者はそうでない歌唱者に比べて歌唱者間の音程が小さくなることが示されている．桑原らの研究では 2 名での重唱において，下のパートの基本周波数を協和音程から 10 cent 単位で 40 cent までずらした合成音声について聴取実験が行われており，10 cent と 20 cent，30 cent と 40 cent の間で「上手さ」についての知覚のギャップがあることが示されている [4]．

また，基本周波数ではない他の音響特徴量を用いて複数人での歌唱の「上手さ」を分析した研究もされている [5]．他の歌唱者が存在する場合の基本周波数系列のモデル化については，ばね質量系を用いた研究がある [6]．

既存研究から，2 名の合唱，重唱音声において音響特徴量の差が小さいほど歌唱が「上手い」と評価される傾向にあることが言える．その一方で，音響特徴量が過度に近しい場合には評価が下がることがあると報告されている．斉唱についても合唱，重唱と同様の傾向が見られることが期待される．このことから，斉唱の基本周波数系列のモデル化についても歌唱者間音程を適度に小さくするのが良いことが予想される．

3. 実験

この章では本研究で行った実験について示す．まず定常的な音程に着目した実験について示し，次に動的な音程に着目した実験について示す．

3.1 実験 1: 定常的な音程を付加した合成斉唱音声の自然性の調査

3.1.1 データセット

本実験ではデータセットとして，100 名の歌唱音声が含まれる JVS-MuSiC [7] を用いた．JVS-MuSiC には全員が歌っている共通曲として日本童謡「かたつむり」がある．本実験ではこの音声データのうち女性歌唱者 51 名のデータの中で，すべての歌唱者についてテンポが同一になるように調整されている音声から ID002，004，007，030，035，065，085，092，093，095 の 10 名を選んで用いた．ただし，テンポが同一であっても音高の変化のタイミングは歌唱者間で完全に同期しているわけではない．すべての音声のサンプリング周波数は 24 kHz であり，音声長は 31 秒である．

3.1.2 斉唱音声の合成

音声分析合成システム WORLD[8](D4C edition[9]) を用いて歌唱者間音程のある 10 名斉唱の合成音声を作成した．歌唱者が異なる 10 個の音声データをもとに 10 個の独唱音声を合成し，重ね合わせることで擬似的な斉唱音声とした．合成に使用する特徴量は，基本周波数系列以外は WORLD を使用して抽出した各歌唱者のスペクトルと非周期性指標をそのまま用いた．窓長は 1024 サンプル（音声長 42.7 ms）でシフト長は 5 ms である．基本周波数系列は本実験で制御する対象であり，音声データの時系列音素ラベルと楽譜で示される音高をもとに作成した．時系列音素ラベルは，Julius[10] で強制アライメントを取った後に手動で修正がなされている先行研究のデータを用いた [5]．

各合成独唱音声の合成に用いる基本周波数系列はそれぞれで音程が生じるように，楽譜通りの基本周波数系列を一律に数 cent だけシフトすることで作成した．シフトする量は平均が 0 cent，標準偏差が 0，10，20，30，40 cent の正規分布から 10 回ランダムサンプリングすることで決定した．この研究では歌唱者間音程に着目しているので，サンプリングしたシフト量の標準偏差が目標の標準偏差の上下 10% 以内にし，サンプリングしたシフト量の平均値が母集団の 0 cent からずれることは考慮をしなかった．

こうして得られる基本周波数系列はステップ状であり，そのまま音声合成すると不自然な音声となる．これは，人間の歌唱音声の基本周波数に，メロディーによる変動以外に動的成分が含まれていることに依拠する [11]．

オーバーシュート　音高が変化した直後に目標とする音高を超えるよう変動する動的成分

プリパレーション　音高が変化する直前に音高変化と逆に変動する動的成分

ビブラート　5〜8Hz で変動する準周期的な動的成分

微細変動　発声区間全体に渡り不規則に微小に変動する動的成分

この中で，オーバーシュート，プリパレーション，ビブラートを考慮した基本周波数系列はステップ状の基本周波数系列にフィルタをかけることで得られる．既存研究 [11] で提案されている手法では 2 次系の伝達関数

$$H(s) = \frac{K}{s^2 + 2\zeta\omega s + \omega^2} \tag{1}$$

のインパルス応答

$$h(t) = \begin{cases} \frac{K}{2\sqrt{\zeta^2-1}} \left(\exp(\lambda_1 \omega t) - \exp(\lambda_2 \omega t)\right) & |\zeta| > 1 \\ \frac{K}{\sqrt{1-\zeta^2}} \exp(-\zeta\omega t) \sin(\sqrt{1-\zeta^2}\omega t) & 0 < |\zeta| < 1 \\ \frac{K}{\omega} \sin(\omega t) & |\zeta| = 0 \end{cases}$$

をフィルタとして用いている．式中の λ_1，λ_2 はそれぞれ $\lambda_1 = -\zeta + \sqrt{\zeta^2 - 1}$，$\lambda_2 = -\zeta - \sqrt{\zeta^2 - 1}$ である．オーバーシュートとプリパレーションは減衰振動のモデルである $0 < |\zeta| < 1$ のインパルス応答をかけることで表現でき，ビブラートは定常振動モデルである $|\zeta| = 0$ のインパルス応答をかけることで表現できる．パラメータである K，ζ，ω の値は先行研究で示されている値と同じものを使用した．各動的成分を付与する際のパラメータの値を表 1 に示す．

微細振動は以上 3 つの動的成分を付与した基本周波数系列に，10Hz でカットオフし，最大振幅が 5Hz になるよう

表 1　フィルタのパラメータの値
Table 1　Value of the filter parameter

F0 動的成分	ω(rad/ms)	ζ	K
オーバーシュート	0.0348	0.5422	0.0348
プリパレーション	0.0292	0.6681	0.0292
ビブラート	0.0345	-	0.0018

にした白色雑音を発声区間の全体に足し合わせることで付与した．こうして得られた基本周波数系列を用いて独唱音声を合成した．

歌唱者が 10 名いる場合，全歌唱者の中でどの歌唱者が高い音高で歌うのか，低い音高で歌うのかによって斉唱の自然性が変化することが考えられる．そこで，事前に歌唱者間音程の標準偏差 10，20，30，40 cent ごとに各歌唱者の音高シフト量が異なる 3 種類の斉唱合成音声に対して，Web 上のクラウドソーシングサービスで音声の自然性について一対比較を行い，自然性がより高いと評価された回数が最も多かった歌唱者ー音高シフト量の組み合わせを採用し，以降の音声合成の際に用いた．

3.1.3　聴取実験条件

聴取実験は，音高シフト量の標準偏差が異なる 2 つの音声を順に聴いてもらい，どちらがより自然な斉唱であったかを答えてもらう一対比較を行った．被験者には事前に，聴かせる音声が 10 名による斉唱を意図して作成された合成音声であることを伝えたうえで実験に参加してもらった．実験はクラウドソーシングで行い，各組み合わせについて，それぞれ 25 人の被験者が順序効果を考慮して各人 2 回評価した．

さらに，声楽経験のある 2 名の被験者に対して合成した 5 種類の音声を聴いてもらい，1〜7 の 7 段階で音声の自然性を評価してもらった．1 が最も自然性が低く，7 が最も自然性が高くなるように評価を行った．

3.2　実験 2:動的な音程を付加した合成斉唱音声の自然性の調査

3.2.1　基本周波数系列の動的制御

先行研究では引き込み現象を考慮した基本周波数系列を，ばね質量系を用いてモデル化をしているが，他者の歌唱の影響を受け始めるまでは時間差が生じるはずであることと，この手法は他歌唱者の音高が既知である場合の手法であり，すべての歌唱者が互いに影響し合うモデルとして不都合であることから新たな手法を提案する．

提案する基本周波数系列の制御法では，楽譜から得られたステップ状の基本周波数系列を 100 ms ごとに直前の 100 ms の全歌唱者の基本周波数の平均値と対象の歌唱者の基本周波数の平均値の比を元に基本周波数の変動速度を決定する．更新時間の幅が 100 ms であるのは，先行研究 [12] において，聴覚刺激の反応時間が最速で 169 ms であったことから，音高が変化した直後の 100 ms は引き込み現象が起こらないと考えたからである．

楽譜をもとに作成したステップ状の基本周波数系列 $f_0^i(t)$ を以下の式で $f_0^{i\prime}(t)$ へと更新する．i は歌唱者 ID を，t は時刻を表している．

$$f_0^{i\prime}(t) = f_0^i(t) + v^i * (t - t_s) + b \quad (t_s \leq t \leq t_s + 100\text{ms}) \tag{2}$$

ただし，(2) 式内の v^i，b はそれぞれ音高変化の速度，基本周波数系列更新の際のバイアス項であり，以下の式で表される．

$$v^i = -k * \log_2 \left(\frac{\overline{f_0^i}}{\overline{f_0}} \right) \tag{3}$$

$$\overline{f_0^i} = \frac{1}{T} \sum_{t=t_s-T}^{t_s} F_0^i(t) \tag{4}$$

$$\overline{f_0} = \frac{1}{10T} \sum_{i=0}^{9} \sum_{t=t_s-T}^{t_s} F_0^i(t) \tag{5}$$

$$b = f_0'(t_s - 1) - f_0(t_s - 1) \tag{6}$$

この更新を $t_s = 0$ から t_s を 100 ms ずつ進めることにより引き込み現象を考慮した基本周波数系列の作成を行う．(3)，(4)，(5) 式は，ステップ状の基本周波数系列 $f_0^i(t)$ にオーバーシュートや微細振動などの動的成分を付与した $F_0^i(t)$ について 100 ms 間の IDi の平均音高と全 10ID の平均音高の比を IDi の音高変化の速度として用いることを示している．(3) 式中の k はハイパーパラメータであり，この値が引き込み現象の強さを決定づける．k の値は 0.005 ごとに変化させていき，全員の音高が収束するまでの時間が先行研究 [3] で示されている 0.6 s〜1.0 s に最も近かった値である 0.015 とした．(4)，(5) 式中の T は基本周波数系列の 100 ms 分のフレーム数であり，今回の研究ではその値は 2400 である．(6) 式のバイアス項は更新の際に $f_0^{i\prime}(t)$ が $f_0^{i\prime}(t-1)$ と連続してつながるように補正するための項である．

3.2.2　斉唱音声の合成

実験 1 と同じデータセットを用いて WORLD により合成を行った．窓長は 1024 サンプル（音声長 42.7 ms）でシフト長は 5 ms である．使用した特徴量は基本周波数系列以外は実験 1 と同様である．基本周波数系列は実験 1 で使用した，音高のシフト量の標準偏差が 0，10，20，30，40 cent の動的制御のない 5 種類，これに提案手法を用いて得た動的制御のある 5 種類，提案手法における音高変化の速度を平均 0，標準偏差 0.005 の正規分布からランダムに決定して得た動的制御のある 5 種類の計 15 種類を使用した．いずれも合成する際にオーバーシュートや微細振動などの動的成分を加えている．

3.2.3 聴取実験条件

提案手法の有効性を確認するため，聴取実験を行った．聴取実験は Web 上のクラウドソーシングサービスで一対比較を行った．音高シフト量の各標準偏差で，提案手法ーステップ状，提案手法ーランダム，ステップ状ーランダムの比較についてそれぞれ 25 人の被験者が順序効果を考慮して各人 2 回評価した．ただし，この組み合わせの中で，0 cent の提案手法ーステップ状の組み合わせは完全に同じ音声の組み合わせである．

3.3 実験 3:収束値を考慮した動的な音程を付加した合成斉唱音声の自然性の調査

実験 2 で提案した手法では基本周波数を動的に制御した合成斉唱音声の音高が収束した際，音程が 0 となっている．しかし，実際の歌唱では音高が収束した際に音程が生じていることが報告されている．そこで，実験 2 の提案手法に音高収束時の音程の値をパラメータとして加えたモデルを作成して斉唱音声を合成し，自然性の評価を行った．

3.3.1 斉唱音声の合成

実験 1，2 と同様に WORLD を用いて合成を行った．窓長は 1024 サンプル（音声長 42.7 ms）でシフト長は 5 ms である．使用した特徴量は基本周波数系列以外は実験 1，2 と同様である．基本周波数系列は実験 2 で使用した，音高のシフト量の標準偏差が 0，10，20，30，40 cent で動的制御のある 5 種類に加え，元の標準偏差 20 cent の場合に収束時の標準偏差が 10 cent となるもの，元の標準偏差 30 cent の場合に収束時の標準偏差が 10，20 cent となるもの，元の標準偏差 40 cent の場合に収束時の標準偏差が 10，20，30 cent となるものの 6 種類を作成して使用した．実験 2 の提案手法における v^i に対して、元の標準偏差 σ_s，目標の標準偏差 σ_g，現在の標準偏差 σ_t を用いて $v^{i\prime} = v^i \frac{\sigma_t - \sigma_g}{\sigma_s - \sigma_g}$ とするモデル化を行っている．実験 2 と同様にいずれも合成する際にオーバーシュートや微細振動などの動的成分を加えている．

3.3.2 聴取実験条件

聴取実験は，元の音高シフト量の標準偏差が同一で，収束時の標準偏差が異なる音声，動的制御を行わない音声の内，2 つの音声を順に聴いてもらい，どちらがより自然な斉唱であったかを答えてもらう一対比較を行った．被験者には事前に，聴かせる音声が 10 名による斉唱を意図して作成された合成音声であることを伝えたうえで実験に参加してもらった．実験はクラウドソーシングで行い，各組み合わせについて，それぞれ 25 人の被験者が順序効果を考慮して各人 2 回評価した．

さらに，声楽経験のある 2 名の被験者に対して聴取実験を行った．元の音高シフト量の標準偏差が同一で収束時の標準偏差が異なる音声，動的制御を行わない音声に対して自然性が高いと感じられた順に順位付けを行ってもらった．

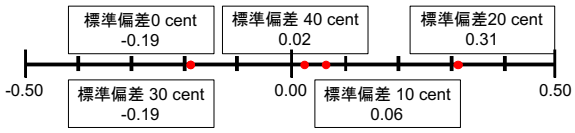

図 1 定常的音程を付与した 10 名斉唱合成音声の自然性の間隔尺度
Fig. 1 Interval measures of naturalness of synthesized 10-person unison with dynamic pitch interval

表 2 10 名斉唱合成音声の自然性の評価値
Table 2 Evaluation value of naturalness of synthesized 10-person unison

[cent]	0	10	20	30	40
評価者 1	4	6	6	5	7
評価者 2	6	5	4	2	3

4. 実験結果

4.1 実験 1

サーストンの一対比較法を用いて音高シフト量の標準偏差が異なる合成音声の自然性についての間隔尺度を算出した結果を図 1 に示す．数値が大きい方が自然性が高いと評価された音声である．

また，声楽経験がある被験者に対する聴取実験の結果を示す．これは数値が大きいほど自然な音声だと評価されたことを示している．コメントは付録にて示す．2 名での評価の平均値は音高シフト量 10 cent の時に 5.5 で最も高くなっている．また，音高シフト量 0 cent の音声に対しては 10 名斉唱音声に聴こえないとフィードバックを得た．

4.2 実験 2

各比較の評価に対してサーストンの一対比較法を用いて間隔尺度を計算した．図 2 に各音声の自然性についての間隔尺度を示す．数値が大きい方が自然性が高いと評価された音声である．音高シフト量の標準偏差の違いによって間隔尺度の値に差があるため，間隔尺度の値は全標準偏差での平均値と 95%の信頼区間を示している．ランダムな動的制御を行って得た基本周波数系列を用いた合成音声は他の手法を使用して得られた基本周波数系列を用いて合成した音声に比べ有意に自然性が低いと評価されたことがわかる．

4.3 実験 3

各比較の評価に対してサーストンの一対比較法を用いて間隔尺度を計算した．図 3 に元の標準偏差 40，30，20，10 cent ごとの各音声の自然性についての間隔尺度を示す．動的な制御を行わなかった合成音声は，収束値がそれぞれ 40，30，20，10 cent の音声とみなせるのでそのように表記する．数値が大きい方が自然性が高いと評価された音声である．声楽経験のある被験者に対する聴取実験の結果を表 3 に示す．これは音声の自然性について順位付けをしたも

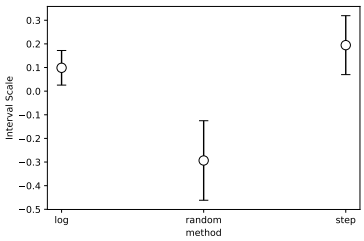

図 2 音程を制御した 10 名斉唱合成音声の自然性の間隔尺度
Fig. 2 Interval measures of naturalness of pitch-controlled 10-person unison with dynamic pitch interval

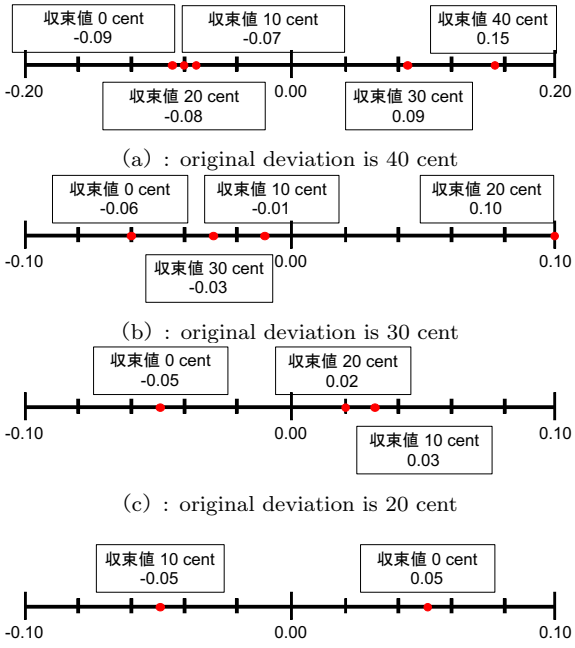

図 3 動的音程を付与した 10 名斉唱合成音声の自然性の間隔尺度
Fig. 3 Interval measures of naturalness of synthesized 10-person unison with dynamic pitch interval

ので，数値が小さいほど自然な音声だと評価されている．コメントは付録にて示す．

5. 考察

実験 1 の図 1 の間隔尺度から，音高シフト量の標準偏差 20 cent の音声が最も自然性が高いと評価されたと考えられる．声楽経験者を対象とした聴取実験の評価値の平均値を見ると，音高シフト量の標準偏差 10 cent の音声が最も自然性が高いと評価されているが，それと同程度に標準偏差 0，20，40 cent の音声も高く評価されている．

音高シフト量の標準偏差 30 cent の音声と 40 cent の音声では，最も音高の高い歌唱者と最も音高の低い歌唱者とでは 100 cent 以上，つまり半音以上離れている．半音離れているとき，同じ音高で歌っているとみなすことができないにも関わらず，歌唱者間音程の無い音声と同じかそれ以上の自然性があると評価されたことから，多人数での斉唱では歌唱者間音程があることが歌唱の自然性を生じさせるのに重要であると考えられる．

実験 2 ではランダムに制御した音声の自然性が低く評価された．実際の歌唱で引き込み現象が起こることを考えると，楽譜上で同一音高を歌っている複数の歌唱者の音程が徐々に大きくなることは起こりえない．ランダムな動的制御を行った基本周波数系列では徐々に歌唱者間音程が大きくなる箇所がある．主にこの箇所の音声により自然性が低下していると考えられる．

ランダムに基本周波数系列の動的制御を行った合成音声は自然性が低く評価されたことから，歌唱者間の基本周波数の比が動的に変化することが知覚に影響を与えると考えられる．しかし，実験 2 の提案手法で基本周波数系列の動的制御を行った音声の自然性は，動的制御を行わないで合成した音声と有意な差はみられなかった．これは，ランダムな動的制御は常に音高が変化するのに対して，提案手法では途中で音高が収束して音高が変化しない箇所があることが影響していると考えられる．実際，実験 2 の条件では定常的な音程は平均して 2 割減少していた．

定常的な音程の減少を考慮した音声について聴取実験を行った実験 3 の一般被験者の評価では，元の標準偏差が 40 cent だった場合を除いて，収束時の標準偏差が元の標準偏差より 10 cent 低い合成音声が最も自然性が高いと評価されている．元の標準偏差が 40 cent だった場合でも，音程の動的制御を行わない音声についで自然性が高いと評価されている．このことから，同一の音高を歌っているはずの歌唱内で極端に音高が変化すると歌唱の自然性が低くなると考えられる．

声楽経験者の評価は表 3 の，元の標準偏差 40 cent の音声では評価がほとんど逆になっている．しかし，コメントを見ると評価者 1 は収束値が 10 cent の音声，20 cent の音声で迷っていたことが明記されており，評価者 2 は 10 cent の音声，20 cent の音声に対して全く同じコメントをしており，30 cent の音声に対してもほとんど同一のコメントをしている．このことから音声の違いは声楽経験者からしてもシビアであることがわかる．

本研究の聴取実験では斉唱の「自然性」を評価してもらった．しかし，合成音声としてではなく，斉唱としての「自然性」という評価基準が被験者に正確に認識されているかは明らかでない．例えば，10 名斉唱を意図した合成音声であると明記したにもかかわらず，「2 声がよくあっている」とコメントを頂いたことから，十分に評価基準の意図が伝わっていない可能性が存在する．

6. 結論

本研究では斉唱における歌唱者間音程のモデル化について新たな手法を提案した．複数の歌唱者間音程の合成音声を用いた聴取実験の結果，斉唱では歌唱者の音高が完全に

表 3　斉唱合成音声の自然性の評価順位

Table 3　Evaluation rank of naturalness of synthesized unison

（a）：original deviation is 40 cent

収束値 [cent]	0	10	20	30	40
評価者 1	3	2	1	5	4
評価者 2	5	2	4	3	1

（b）：original deviation is 30 cent

収束値 [cent]	0	10	20	30
評価者 1	1	4	3	2
評価者 2	1	3	4	2

（c）：original deviation is 20 cent

収束値 [cent]	0	10	20
評価者 1	1	3	2
評価者 2	2	3	1

（d）：original deviation is 10 cent

収束値 [cent]	0	10
評価者 1	1	2
評価者 2	2	1

同一になるようにするのではなく，歌唱者間の基本周波数のシフト量の標準偏差が 10 cent となるような歌唱者の基本周波数系列のモデル化をすることによって，合成音声がより自然になるモデルとなることが示された．100 ms ごとに各歌唱者の音高と全歌唱者の音高の平均値との比から各歌唱者の音高の変動速度を決定する提案手法は，実際の歌唱に存在する引き込み現象を考慮した複数の歌唱者の基本周波数系列の同時作成を可能にした．

本研究では 10 名の斉唱までを対象としているが，100 名などのより多い人数での斉唱では歌唱者間の基本周波数のシフト量の標準偏差が 10 cent であっても最も高い音高になるシフトと最も低い音高になるシフトとでは音高比が 100 cent を超えることが予想される．音高比が 100 cent を超えているとき，その 2 名は同一音高で歌唱しているとはいえない．この条件下においても音高シフト量の標準偏差を 20 cent とするモデルが適合するのかについての議論が必要である．

提案手法に存在するハイパーパラメータの値を，今回は全歌唱者の基本周波数が収束するまでの時間から決定したが，実際に収録した斉唱音声を用いて推定することでより正確なモデル化ができることが考えられる．また，歌唱者ごとに引き込み現象の強さが異なることも考えられるので，各歌唱者ごとにハイパーパラメータの値を変化させるモデル化も考えられる．

参考文献

[1] 剣持秀紀，大下隼人，歌声合成システム VOCALOID，情報処理学会研究報告．[音楽情報科学]，Vol. 72，No. 102，pp. 25 – 28，2007.
[2] 片岡靖景，伊藤一典，池田操，中澤達夫，米沢義道，今関義弘，橋本昌巳，歌唱支援システム構築のための歌声の分析と評価，情報処理学会研究報告．[音楽情報科学]，Vol. 98，No. 74，pp. 23 – 30，1998.
[3] 野田雄也，合唱における基本周波数の同期現象に関する基礎研究．2008.
[4] 桑原彰宏，徳田功，合唱音声の同期解析，電子情報通信学会技術研究報告，Vol. 110，No. 122，pp. 91-95，2010.
[5] 山内孔貴，須田仁志，齋藤大輔，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，情報処理学会研究報告音楽情報科学，Vol. 127，No. 35，pp. 1-6，2020.
[6] 川岸基成，川渕将太，宮島千代美，北岡教英，武田一哉，合唱における歌声の引き込みを利用した歌声 F_0 制御の検討，情報処理学会研究報告．[音楽情報科学]，Vol. 102，No. 13，pp. 1-6，2014.
[7] Hiroki Tamaru, Shinnosuke Takamichi, Naoko Tanji, and Hiroshi Saruwatari, JVS-MuSiC: free Japanese multi-speaker singing-voice corpus, arXiv, 2001. 07044, 2020.
[8] Masanori Morise, Fumiya Yokomori, and Kenji 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, 2016.
[9] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, Vol. 84, pp. 57-65, 2016.
[10] Akinobu Lee, Tatsuya Kawahara and Kiyohiro Shikano, Julius - an Open Source Real-Time Large Vocabulary Recognition Engine, Proceedings of European Conference on Speech Communication and Technology, Vol. 3, pp. 1691-1694, 2001.
[11] Takeshi Saitou, Masashi Unoki and Masato Akagi, Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis, Speech Communication, Vol. 46, pp. 405 – 417, 2005.
[12] 高岡佳弘，長谷川賢一，堀内和之，岡本途也，聴覚の反応時間測定法に関する研究，Audiology Japan，Vol. 27，No. 5，1984.

付　　録

A.1　実験 1，3 における声楽経験者の評価コメント

実験 1,3 の評価コメントについて表 A.1-A.5 に示す

表 A·1 実験 1 に対する声楽経験者のコメント

音程	評価者 1	評価者 2
0 cent	2 声がよく合っている	音色が同じものが重なって聴こえる。全体の豊かな響きの為に声の質が増えた方がよいのでは。
10 cent	音色の違いが気にならずよく合っている	自然に聴こえる。
20 cent	ビブラートがいかにも人工音声っぽい	自然に聴こえる。
30 cent	下方向のビブラートが自然ではない	ピッチの低いものがあるのが気になってしまう。
40 cent	上方向のずれが自然ではない	それぞれの声の質が重なって豊かな響きに聴こえる。

表 A·2 実験 3 に対する声楽経験者のコメント（元の標準偏差 40 cent）

音程	評価者 1	評価者 2
制御なし	2 声がやや分離して聴こえる	10 人の斉唱として自然に聴こえる。
0 cent	人工音声としては十分自然のうちか。	ピッチの理由か響きに明るさが欲しいと感じる。
10 cent	音色の違いはあまり気にならないがどちらかといえば 04 が 1 位か	声の種類のバランスがよいと感じる。
20 cent	音色の違いはあまり気にならい	声の種類のバランスがよいと感じる。
30 cent	1 番音色のずれが気になる	バランスがよいと感じる。

表 A·3 実験 3 に対する声楽経験者のコメント（元の標準偏差 30 cent）

音程	評価者 1	評価者 2
制御なし	長い音の時に特に不自然さが気になる	歌の響きとして、もっと明るい方がよいと感じる。ピッチが理由か。
0 cent	響きがこの中では 1 番自然	自然に聴こえる。
10 cent	音色音程の違いがかなり気になる	若干バラつきを感じる。
20 cent	上方向のずれの方が 03 のりややまし	若干バラつきが気になる。

表 A·4 実験 3 に対する声楽経験者のコメント（元の標準偏差 20 cent）

音程	評価者 1	評価者 2
制御なし	不自然ではないが 02 の方がより自然的	10 人の斉唱として好感がもてる響きと感じる。
0 cent	比較的聞き辛くはない	自然な斉唱と感じる。
10 cent	響きが不自然に聴こえる。	若干ピッチのバラつきを感じる。

表 A·5 実験 3 に対する声楽経験者のコメント（元の標準偏差 20 cent）

音程	評価者 1	評価者 2
制御なし	よく合っているが若干不自然な気がする	自然に聴こえる
0 cent	音色はよく合って気にならない	ピッチがやや低いような声が気になるが自然である。