人間–AI斉唱における合成歌声の呼吸パラメータの歌唱者間リアルタイム同期

深尾 貫太[1,a)]　三井 啓史[1,b)]　小野 晶子[1,c)]　上原 崇寛[1,d)]　高道 慎之介[1,2,e)]

概要：本研究では，人間–AI 斉唱において，人間歌声に作用されて呼吸パラメータをリアルタイムに制御する歌声合成を提案し，この手法が人間歌声及び人間歌手の主観的体験に与える影響を実験的評価により明らかにする．人間斉唱及び人間合唱では歌唱者間でリアルタイムに相互作用が生じることで歌声が調和することが知られている．AI 斉唱及び AI 合唱においてもその再現が試みられている一方で，人間-AI 斉唱において，人間歌声に合わせてリアルタイムに変化する合成歌声の実現については，依然として限定的である．本稿では，歌声パラメータから呼吸パラメータへの回帰を利用して，人間歌声に対して合成歌声を動的に切り替える方法を提案する．

1. はじめに

他者と群を成して歌唱するアンサンブル（本稿では**人間斉唱**, **人間合唱**と称する）においては, 歌唱者同士がリアルタイムに相互作用する．自分の歌声が他者の歌声に影響を受けたり，逆に自分の歌声が他者の歌声に影響を与えることで，全体として歌声の調和が起こる [1]（図 1 左）．相互作用の具体例としては, F_0, ビブラートそして singer's formant といった音響特徴量が挙げられること [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] に加えて，身体運動の一つである呼吸も，相互作用することが知られている [11], [12].

他方，歌声合成の技術は著しい発展を続けており，人間歌声と遜色ない程度にまで自然な歌声が目指されている [13]．既存の歌声合成技術は基本的に独唱を想定しており [14]，いかなる他者との相互作用も発生しない．これに対し，複数の歌声合成による斉唱や合唱（本稿では *AI* **斉唱**，*AI* **合唱**）において，合成歌声特徴量間の関連に着目した，相互作用の再現が試みられている [15].

さらなる歌唱形態として，人間と歌声合成による

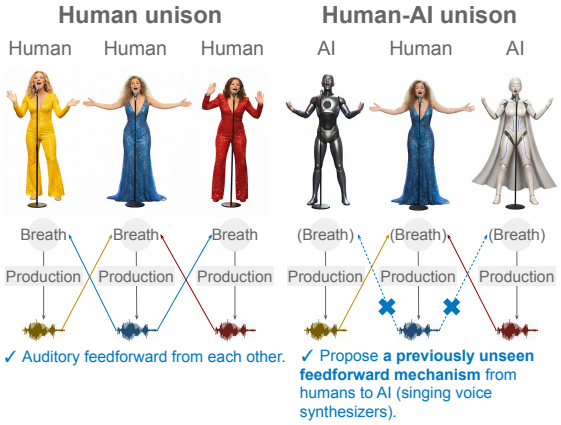

図 1: 本研究の概要．人間斉唱（左）では自身の歌声が他者にリアルタイムに作用するが，従来の人間–AI 斉唱（右）では歌声合成は他の人間歌声に作用されない．本研究では，人間歌声の呼吸パラメータにリアルタイムに作用される歌声合成を提案する．

アンサンブル（本稿では**人間**–*AI* **斉唱**，**人間**–*AI* **合唱**）が考えられる（図 1 右）．この実現にあたっては，人間歌声にリアルタイムに影響される歌声合成，ならびに，それを受けた人間の挙動の調査が必要である．

本研究では，人間–AI 斉唱において，人間歌声に作用されて呼吸パラメータをリアルタイムに制御する歌声合成を提案し，この手法が人間歌声及び人間歌手の主観的体験に与える影響を実験的評価により明らかにする．この提案モデルでは，人間歌声から

[1] 慶應義塾大学
Keio University
[2] 東京大学
University of Tokyo
[a)] kanta.fukao@keio.jp
[b)] kcmitt21@keio.jp
[c)] himiko0406@keio.jp
[d)] doubl.tool.dc215.book@gmail.com
[e)] shinnosuke_takamichi@keio.jp

その呼吸パラメータをリアルタイムに逆推定し，その呼吸パラメータを持つよう合成歌声を動的に切り替える（図 1 右）．この試みは，人間歌手の斉唱相手を，人間歌声に影響を受けて歌声を変えるような AI に拡張するものである．さらに，そのような AI の変化を受けて人間歌手が歌唱を調整するという間接的な相互作用の可能性を検証する点において，アンサンブルの新しいあり方と，音楽への新たな向き合い方を示すことを目指す．

2. 関連研究

2.1 LF モデル

本研究では歌声合成をリアルタイムに制御するが，その対象となる呼吸パラメータと，それらを含めた声帯の挙動を表すようなパラメータ（声帯パラメータ）について述べる．LF（Liljencrants-Fant）モデルとは，ソース・フィルタ理論を根拠として，基本周期 T_0 において声帯を通る息量の変化量をモデル化したものである [16]．このモデルでは，図 2 のように基本周期が開放期，閉鎖期，静止期の 3 つのフェーズから構成され，それぞれにおいて振幅 $u'(t)$ は以下のように定式化される．

$$u'(t)=\begin{cases} E_1 e^{at}\sin(\omega_g t) & 0 \le t \le T_e \\ -E_2\left(e^{-b(t-T_e)}-e^{-b(T_0-T_e)}\right) & T_e \le t \le T_c \\ 0 & T_c \le t \le T_0 \end{cases} \quad (1)$$

以上の式に現れるパラメータのうち，a, b, ω_g, E_1, E_2 はグラフの形状を決定づけるパラメータであるが，特に時間に関するものが T_e, T_c, T_0 である．T_e は声門が開放している時間を，T_c は声門が開放し始めてから閉鎖が終わるまでの時間をそれぞれ表す．さらに，これらの時間に関するパラメータを用いて，$T_p, T_a, \alpha_m, O_q, R_q$ を定義できる．T_p は息量が最大値を取る時間を，T_a は指数減衰の鋭さをそれぞれ表す．α_m, O_q, R_q はいずれも割合を示す指標であり，それぞれ尖度，声門開放率，声帯接触速度率と呼ばれ，以下のように定義される．

- 尖度：$\alpha_m = T_p/T_e$
- 声門開放率：$O_q = T_e/T_0$
- 声帯接触速度率：$R_q = T_a/((1-O_q)T_0)$

このうち，声門開放率は息漏れを示すパラメータであり，値が大きいほど音声の息っぽさが増す．一方で，声帯接触速度率は声門閉鎖の鋭さを示すパラメータであり，値が大きいほど柔らかい声質になる [17]．尖度はグラフの形状における非対称性を示すパラメータであり，例えばファルセットでは声帯振動が部分的になるため値が小さくなる [18]．以上において言及された，グラフの形状を決定づけるパラメータのうち，時間あるいはその割合に関する 8 つ（$T_0, T_e, T_c, T_p, T_a, \alpha_m, O_q, R_q$）を**声帯パラメータ**とする．なお，LF モデルは声帯レベルの現象（ソース）をモデル化したものであり，声道共鳴（フィルタ）についてはその範囲外である．本研究では，まず声帯パラメータに限定して検討を行い，その効果の範囲を明らかにすることを目指す．

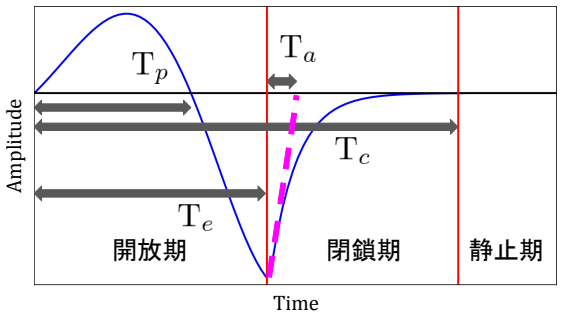

図 2: LF モデルの波形の例.

2.2 人間斉唱・人間合唱における傾向

音響特徴量への影響．本研究では人間–AI 斉唱を人間斉唱・人間合唱を拡張したものとして位置づけることができるが，人間斉唱・人間合唱におけるリアルタイム相互作用に関して，様々な先行研究が存在する．まずは音響特徴量，ここでは特に F_0，ビブラート，singer's formant における傾向について述べる．F_0 に関して，斉唱では 16〜30 cents の分散が歌声間に起こることが知られているが [3]，F_0 誤差を歌唱の形態ごとに比較した先行研究 [1] によると，斉唱における F_0 誤差は独唱よりも大きくなることが確認されている．一方でビブラートについては，歌唱者が熟達しているほど，合唱におけるビブラートの周期と幅が独唱よりも制御されることが報告されている [4]．合唱効果と関連する客観指標についての先行研究 [5] でも，同様に合唱ではビブラートが抑制されることが明らかになっており，加えてビブラートをするほとんどの歌唱者の間でその同期が確認されている．また主観評価実験により，このような独唱と合唱の間でのビブラートの変化は自発的であり，さらに被験者が調整した要素のうち最も共通するものがビブラートであることが示されている [6]．また singer's formant について，合唱ではその周波数帯でのエネルギーが独唱よりも弱まる傾向にあり，同時に基本周波数帯でのエネルギーは強まる傾向にあることが報告されている [7], [8]．実際に合唱の聴取実験においては，singer's formant の弱い声の方が有意

に好まれることが確認されている [19]．さらに，アマチュア合唱歌手は独唱においても singer's formant をほとんど使わないのに対して [9]，修練度が上がるほど歌唱形態に応じた声質の切り替えが顕著になる傾向が報告されている [10]．

呼吸への影響．音響特徴量以外にも，呼吸が斉唱と関連することが報告されている．聴取評価語指標と音響特徴量の重回帰分析を実施した研究 [20] において，「息の流れがあっている，無理のない発声」を意味する呼吸という聴取評価語指標は，声門開放率，声帯接触速度率，中期から後期における音高変動率，第 2 声道共鳴周波数と因果関係を持つことが示されている．なお本研究では，**呼吸パラメータ**を声門開放率，声帯接触速度率，音高変動率の総称とし，音高変動率は F_0 を中央値を基準とした cent 偏差の平均絶対値と定義する．第 2 声道共鳴周波数については，LF モデルはソース・フィルタ理論におけるフィルタを対象としていないため，呼吸パラメータの定義には含めていない．また合唱における呼吸パターンについて，歌唱者のそれは指揮者の後に続いており，このような因果が，目を閉じていてお互いが見えない場合でも存在することが示されている [11]．

2.3 AI 斉唱・AI 合唱，人間–AI 斉唱における傾向

AI 合唱において声部間の相互作用をモデリングした歌声合成の技術も存在し，複数声部の間で，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とで，時刻ずれと持続長の予測精度が向上することが報告されている [15]．また，AI 斉唱の自然さに寄与する因子として，タイミングと F_0 があることが明らかになっている [14]．その上近年では，人間歌声と合成歌声との調和に関しても検討されており，VocaListener2 ではピッチやダイナミクスだけでなく，音色も人間歌声に寄せることができる [21]．

一方で，人間が合成歌声の録音に合わせて歌唱する条件では，人間歌声との歌唱よりも同調が小さくなることが報告されており，この理由として，呼吸のような人体固有の動きの有無による可能性が示唆されている [22]．また人間歌声に合わせてリアルタイムに変化する合成歌声の実現については，依然として限定的である．

3. 提案手法

提案する歌声合成は，人間–AI 斉唱において人間歌声に作用されて呼吸パラメータをリアルタイムに制御する．具体的には以下の手順からなる．

(1) **人間歌声から呼吸パラメータを推定する．**その前準備として，歌声パラメータから呼吸パラメータをリアルタイムに推定できる機械学習モデルを事前学習する．

(2) **推定した呼吸パラメータに対応する歌声を合成する．**この実現には呼吸パラメータを入力とする歌声合成モデルの学習及び推論が理想だが，リアルタイム制御の観点から現実的でない．そこで，各種呼吸パラメータに対応する合成歌声を事前に作成しておき，それを動的に切り替える．

以降に手法の詳細を述べる．

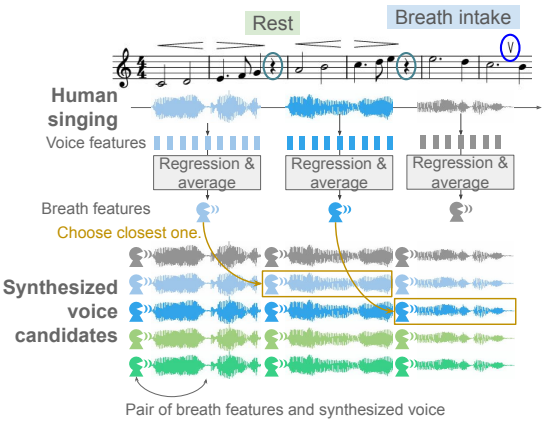

図 3: 合成歌声の動的切り替え

3.1 歌声から呼吸パラメータの推定

複数の人間歌手をそれぞれ独唱させて，その歌声信号と声帯振動信号を時刻同期させて事前収録する．歌声波形から歌声パラメータを，声帯振動から呼吸パラメータをそれぞれ抽出し，歌声パラメータから呼吸パラメータへの回帰を機械学習モデルで学習する．

3.2 呼吸パラメータに対応する歌声の合成

各呼吸パラメータに対応する歌声を合成する．3.1 節にて用意した，呼吸パラメータのデータ集合と，歌声パラメータから呼吸パラメータへの回帰モデルを利用する．呼吸パラメータのそれぞれに対し，データ集合の 12.5 %, 25 %, 50 %, 75 %, 87.5 %の 5 種類の代表値を合成歌声が持つように，学習済みの歌声合成モデルに入力する**合成パラメータ**（後述）を調整する．この調整は，合成パラメータの変更，歌声合成，回帰モデルを用いた呼吸パラメータ推定を手動反復することで行う．3 種類の呼吸パラメータと 5 種類の代表値の全組み合わせにより，計 $5^3 = 125$ 種類の歌声を合成する．

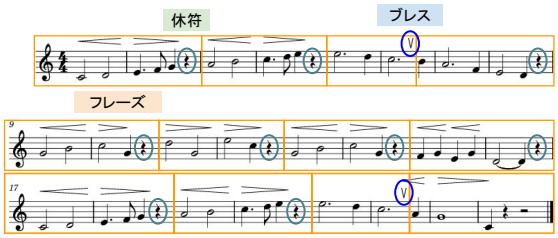

図 4: コンコーネ 50 番の第 1 番の楽譜

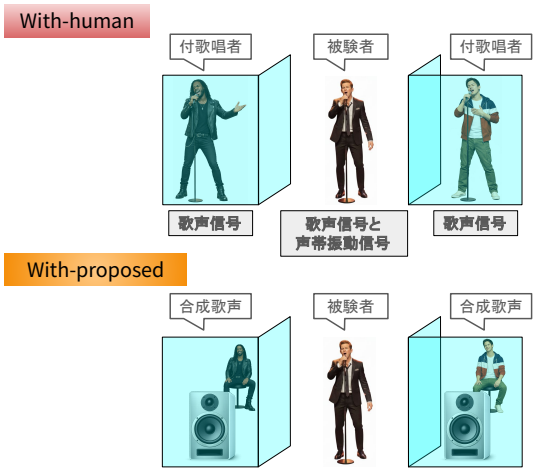

図 5: 実験における歌唱の様子

3.3 人間歌声に対する合成歌声の動的切り替え

人間歌声に応じて合成歌声をリアルタイムに切り替える．ただし，フレーム単位で切り替えると聴感上不自然な音の切り替わりが想定されるため，本研究では事前に決定したフレーズ毎に切り替える．図 3 はその概要である．

まず，あるフレーズに対する人間歌声を観測する．その各フレームの歌声パラメータを抽出し，フレームごとに回帰モデルで呼吸パラメータを推定する．その呼吸パラメータをそのフレーズの全歌唱時間区間で平均したものを，当該フレーズの呼吸パラメータとする．その後，次のフレーズの合成歌声を決定する．3.2 節で作成した合成歌声のうち，推定した呼吸パラメータと，誤差とペナルティの総和が最小となる合成歌声を，次のフレーズの合成歌声とする．これを各フレーズについて繰り返す．

以上の手続きにより，人間歌声の呼吸パラメータに対応する合成歌声がフレーズ毎に選択され，人間斉唱のように，人間歌声に影響を受けて歌唱を調整する現象を模擬できる．

4. 実験的評価

4.1 実験条件

人間–AI 斉唱において提案手法が人間歌手に与える影響を実験的に調査した．

使用する楽曲． コンコーネ 50 番第 1 番（図 4）を用いた．この曲ではブレス位置が限定され被験者に指定する必要がなく，音域が広いという特徴がある．男声の歌声データベース及び男性被験者（歌唱者）の音域は C3〜E4，女声の歌声データベース及び女性被験者（歌唱者）の音域は C4〜E5 である．BPM = 80 であり，歌唱前に 2 小節間メトロノームのオーディオファイルを再生することで，テンポの提示及びオンセットの時刻同期を図る．人間歌手の歌唱開始と同時にピアノ伴奏のオーディオファイルを同時に再生し，ピッチ参照の 1 つとしている．図 4 に示すように，休符あるいはブレス位置を区切りとして 12 フレーズを設けており，各フレーズの時間は約 4.5〜5.25 s である．このように各フレーズの終わりが非歌唱区間となるように割り当てることで，全歌唱区間を呼吸パラメータへの回帰に用いることができる．また実験を通して，本曲を /a/ の音素で歌唱した．

計測機器と環境． 声帯振動信号の測定には VoceVista 社の EGG（electroglottograph）[*1]を使用した．喉頭部分の左右に小さな電極パッドを装着し，微弱な高周波電流を流した際の電気インピーダンスの変化を電圧波形として記録することで，声帯間の接触の程度を計測した．歌声を収録するマイクには，audio technica AT2035[*2]を用いた．実験は，静音な環境で実施した．

歌声パラメータから呼吸パラメータの回帰． 歌声パラメータを入力として，スプライン回帰における誤差と GMM（混合ガウスモデル）におけるペナルティの和が最小となるような呼吸パラメータの代表値を推定し，その組み合わせに対応する合成歌声を選択した．回帰誤差 $D_{\mathrm{reg}}(c)$ とペナルティ $D_{\mathrm{prior}}(c)$ は，それぞれ以下のように定義した．$\hat{O}_q$ と $\hat{R}_q$ は回帰により得られる声門開放率と声帯接触速度率の推定値を，$c = (O_c, R_c)$ は声門開放率と声帯接触速度率の代表値を，σ_O と σ_R は声門開放率と声帯接触速度率の事前分布における標準偏差をそれぞれ表す．

- 回帰誤差：$D_{\mathrm{reg}}(c) = \left((O_c - \hat{O}_q)/\sigma_O\right)^2 + \left((R_c - \hat{R}_q)/\sigma_R\right)^2$
- ペナルティ：$D_{\mathrm{prior}}(c) = -\log p_{\mathrm{GMM}}(O_c, R_c)$

1 つの歌声パラメータから 1 つの声帯パラメータを推定するように学習されているが，その組み合わせにおける回帰曲線の決定係数で重みづけをすることにより，結果を統合するときに性能の悪い回帰器の

[*1] https://www.vocevista.com/en/products/electroglottograph-egg/
[*2] https://www.audio-technica.co.jp/product/AT2035

影響を小さくしている．加えて GMM を適用し，ある領域における密度の高低に基づいてペナルティの大きさを変えることで，外れ値の影響を抑えている．なお本研究では，**歌声パラメータ**は F_0，メル周波数ケプストラム係数（5 次元），スペクトル重心，スペクトル帯域幅，ゼロ交差率，音高変動率，RMS レベルの 11 種類のパラメータを指す．メル周波数ケプストラム係数はメル尺度で変換された周波数成分をケプストラム分析して得られる音色に関する特徴量を，スペクトル重心は時間軸におけるスペクトルの中心を，スペクトル帯域幅はその時間軸におけるスペクトルの分布の広がりを[*3]，ゼロ交差率は信号が正から負または負から正に切り替わる頻度を，RMS レベルは RMS（root mean square）を対数尺度として dB 表現したものを，それぞれ表す．回帰モデルの学習には，男性 6 名と女性 2 名による独唱をマイクと EGG で時刻同期して観測したデータを用いた．

歌声合成モデルと合成歌声． 歌声合成ソフトウェアとして Synthesizer V Studio Pro[*4]を用いた．男性声の歌声合成として，Kevin[*5]，Ryo[*5]，Jin[*5]そして Mo Chen[*5]，女性声の歌声合成として，花隈千冬[*6]，小春六花[*7]，京町セイカ[*8]そして Saki AI[*5]を用いた．このソフトウェア上の合成パラメータ（breath, tension, vibrato）が，呼吸パラメータを構成する声門開放率，声帯接触速度率，音高変動率にそれぞれ対応していると仮定して，合成パラメータの調整と合成歌声の事前作成を行った．

被験者． 後述するように，本実験では人間–AI 斉唱と比較するために，合成歌声を人間に交替した人間斉唱も実施する．以降では，両方の斉唱に参加する人間を単に被験者，合成歌声から交替して参加する人間を付歌唱者（つけかしょうしゃ）と称する．付歌唱者は性別ごとにペアで参加し，男性 2 ペアと女性 2 ペアの計 8 名が参加した．付歌唱者はいずれも合唱経験者で，ペア同士で互いに面識があり普段から共に合唱活動をしている．被験者は男性 4 名，女性 3 名の合計 7 名であり，合唱経験があるのは男女 1 名ずつに限られるが，全員演奏経験がある．被験者と付歌唱者の間で人間の重複はない．

4.2 比較手法

以下の 4 歌唱形態を比較する．

- **Solo (独唱)**：他歌唱者を設けず，被験者単独で歌唱する．
- **With-human (斉唱)**：被験者は同性 2 名の付歌唱者と共に歌唱する．
- **With-baseline (斉唱)**：被験者は同性声 2 つの歌声合成と共に歌唱する．合成歌声は固定の合成パラメータから合成されており，被験者からの影響を受けない．
- **With-proposed (斉唱，提案法)**：With-baseline と同形態だが，合成歌声は被験者の呼吸パラメータに影響される．

With-proposed 形態における被験者の挙動や評価が，With-baseline 形態と比較して With-human 形態における挙動や評価に近づくことを期待する．

4.3 実験の流れ

図 5 に実験の様子を示す．実験室中央に被験者の立つ位置を用意し，その眼前にマイクを配した．With-proposed 形態においては，このマイクに収録された歌声に基づいて呼吸パラメータが推定される．被験者の位置を両側から挟むように，付歌唱者あるいは合成歌声再生用スピーカを配した．この際，付歌唱者あるいは合成歌声再生用スピーカが被験者から見えないように，その間に不透明パーティションを設けた．

被験者が実験室に来る前に，付歌唱者を実験室に召喚し所定の位置に立たせた．被験者が来たのちに被験者を所定の位置に立たせた．被験者はその場で十分な歌唱練習を行った後，まず Solo 形態の歌唱を行う．その後，With-human，With-baseline，With-proposed 形態を被験者毎にランダムな順番で実施した．練習を除く全ての歌唱形態において，EGG とマイクによる同期収録を実施した．Solo 形態において観測された呼吸パラメータの平均値から，当該話者が斉唱中に生じさせると思われる呼吸パラメータ候補を事前に定めた．具体的には，$5^3 = 125$ 種類の候補のうち，観測したパラメータ平均値の近傍 $3^3 = 27$ 種類を候補とした．

付歌唱者は With-human 形態のときのみ歌唱した．このとき With-proposed 形態と条件を揃えるために，被験者に合わせて歌唱するように指示をした．それ以外の形態においても付歌唱者はその場に立っており，被験者は全ての形態において，付歌唱者の発する物音などを聴取できる環境であった．

4.4 歌唱後アンケート

全歌唱形態の収録が終了したのち，被験者に歌唱

[*3] https://librosa.org/doc/main/generated/librosa.feature.spectral_bandwidth.html
[*4] https://dreamtonics.com/ja/synthesizerv/
[*5] https://www.ah-soft.com/synth-v/dreamtonics/
[*6] https://www.ah-soft.com/synth-v/chifuyu/
[*7] https://www.ah-soft.com/synth-v/rikka/
[*8] https://www.ah-soft.com/synth-v/seika/

表 1: 被験者が 7 段階の MOS で回答する項目

番号	内容
独唱のみ	
Q01	この曲の独唱を難しいと感じましたか？
独唱・斉唱共通	
Q02	合唱全体としてエネルギッシュな演奏でしたか？
Q03	合唱全体として力強いと感じる演奏でしたか？
Q04	満足のいく演奏ができましたか？
Q05	演奏していて心地良いと感じましたか？
Q06	演奏していて楽しいと感じましたか？
Q07	演奏していて緊張しましたか？
Q08	総じて自分の演奏に自信は持てましたか？
Q09	自分のピッチに自信は持てましたか？
斉唱のみ	
Q10	相手が自分をリードしていると感じましたか？
Q11	相手の演奏に対して，自分の演奏はバランスが取れていると感じましたか？
Q12	自分が相手をリードしていると感じましたか？
Q13	自分の演奏が相手の演奏とうまく混ざっている（溶け込んでいる）と感じましたか？
Q14	自分の演奏が相手の演奏と調和していると感じましたか？
Q15	自分の演奏に対して，相手の演奏はバランスが取れていると感じましたか？

後アンケートを実施した．具体的な項目を表 1 に示す．Q02〜Q09 では歌唱に対する肯定感を問う質問を，Q10〜Q15 では自分と相手との相互作用及びその結果としての一体感を問う質問をそれぞれ設けている．

4.5 主観評価

4.5.1 アンケート結果の分析

Q03 と Q05 のスコア，Q12 のスコアを，それぞれ図 6，図 7 に示す．全質問のスコアは付録 A.1 を参照されたい．

Q03（力強さ）． With-baseline 形態から With-proposed 形態にかけて 7 名中 6 名のスコアが上昇し，With-human 形態に接近することを確認できる．このことから，With-baseline 形態よりも With-proposed 形態の方が，人間歌手が加わった場合と近い変化が起こることが示唆される．ただし，いずれも With-human 形態と有意な差が生じていた．

Q05（心地よさ）． まず，Solo 形態から With-human 形態にかけては中央値が上昇する一方で，With-baseline 形態にかけては減少することがわかる．これは，独唱から斉唱に歌唱形態が変化すると，その相手が人間であればより心地良く感じるようになるが，AI であるとかえって心地良さが損なわれることを示唆している．対して With-proposed 形態はその減少度合いを緩和できていることがわかる．また，With-baseline 形態と With-proposed 形態の被

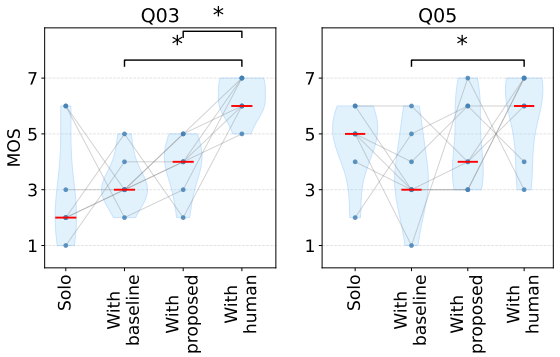

図 6: Q03 と Q05 の回答結果．点は被験者，赤線は中央値を表す．∗ は，With-{baseline, proposed} と With-human の間に $p < 0.05$ の有意差があることを示す．

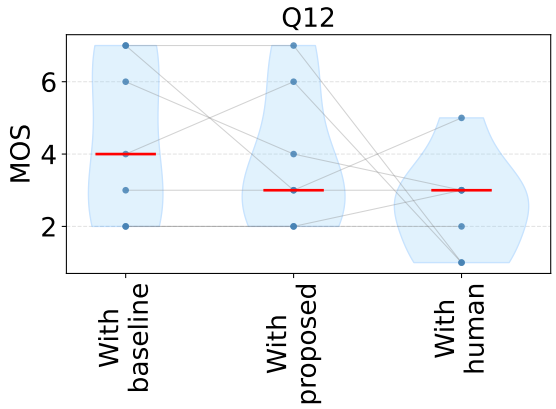

図 7: Q12 の回答結果．点は被験者，赤線は中央値を表す．

験者毎のスコアを比較すると，7 名中 6 名において With-proposed 形態によるスコアの減少は見られない．以上より，人間–AI 斉唱特有の被験者の感じる心地悪さは軽減されている可能性が示唆される．

Q12（相手をリード）． With-baseline 形態と異なり，With-proposed 形態は With-human 形態と同じ中央値をとっている．この結果は，提案手法では合成歌声を動的に切り替えたことで，その影響を受けた被験者がさらに相手に影響を与えるというような相互作用が生じており，人間斉唱の場合と同様に自分から相手を過剰にリードしようとする必要がなかった可能性を示唆している．

以上のように，力強さや心地よさといった肯定感を問う質問においては，付録 A.1 にあるように 8 個中 6 個において改善が見られた．この結果は，これらの項目が呼吸パラメータの動的制御によって影響を受けやすい体感であり，呼吸パラメータと合成パラメータの対応の仮定にある程度の妥当性が認められることを示唆している．一方で，一体感を問

う質問においては，With-proposed 形態の中央値が With-baseline 形態と比較して With-human 形態に近づいたのは，付録 A.1 にあるように 6 個中 2 個に留まった．この結果は，これらの項目は，声道共鳴や singer's formant，合成歌声を切り替える時間粒度に挙げられるような呼吸パラメータ以外の要因が支配的であることを示唆している．

4.6 客観評価

4.6.1 評価方法

被験者の歌声波形におけるフレーズ単位での音響特徴量の中央値を，図 9 左側に示す．また，各斉唱形態における 3 つの歌声波形（被験者と合成歌声あるいは付歌唱者）を入力として推定した音響特徴量に対して，被験者と斉唱相手とのフレーム単位での絶対誤差の和を，図 8，図 9 右側に示す．なお SPR（singing power ratio）とは，基本周波数の領域（0〜2000 Hz）のエネルギーに対する singer's formant を含む領域（2000〜4000 Hz）におけるエネルギーの比を dB 表記したものである．

4.6.2 実験結果

メル周波数ケプストラム係数（1 次元）． With-baseline 形態から With-proposed 形態にかけて全員の値が小さくなっており，With-human 形態に接近することを確認できる．メル周波数ケプストラム係数の低次元は，スペクトル包絡を集約して音色を表現する指標である．したがって，この結果は被験者と斉唱相手との音色の差が小さくなったことを示唆しており，独唱から合唱にかけて他者の歌声と混ざり合うように音色を調整する傾向があることを報告する先行研究 [2], [23] と，整合するものであると言える．

SPR． 図 9 左側において，With-baseline 形態と With-proposed 形態に中央値の明確な差異は確認できない．本研究では，呼吸パラメータへの回帰の学習及び合成歌声の事前作成が声帯レベルの仮定に基づいていることを考慮すると，フィルタと密接に関係する SPR が歌唱形態ごとに大きく変化しないことを示すこの結果は妥当であると言える．また Solo 形態から With-human 形態にかけても大きな変化はないことが確認できる．本研究では被験者のうち合唱経験があるのは 7 名中 2 名に留まっており，アマチュア合唱歌手において singer's formant を強める傾向は限定的であること [9] が，原因として考えられる．また図 9 右側における被験者と斉唱相手との差異に着目すると，With-baseline 形態と With-proposed 形態では顕著な差はない一方で，With-human 形態は

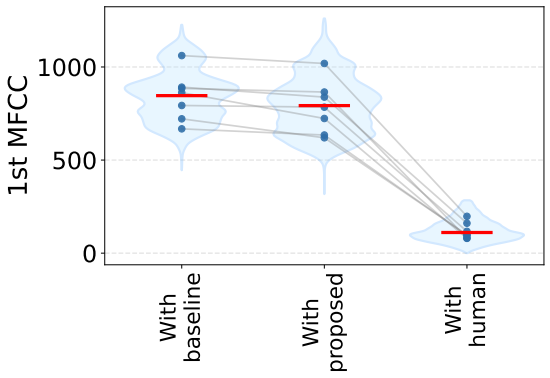

図 8: メル周波数ケプストラム係数（1 次元）の分布．点は被験者，赤線は中央値を表す．

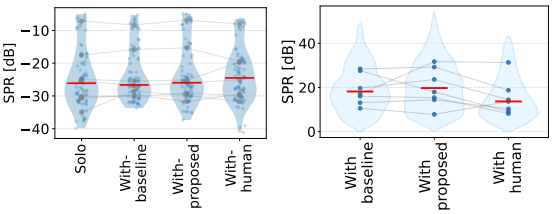

図 9: SPR(singing power ratio) の分布．左図は被験者の推定値，右図は被験者と斉唱相手との推定値における絶対誤差の和である．また点は被験者，赤線は中央値を表す．

それら 2 形態よりも中央値が小さいことがわかる．この結果は，singer's formant を調整する程度は修練度と関連しており [10]，実際に付歌唱者はペア同士で互いに面識があり普段から共に合唱活動をしているため，被験者よりも声を合わせる能力が高い可能性を示唆している．

以上のように，メル周波数ケプストラム係数（1 次元）に見るように，客観的には音色の差が縮まったことを示唆する結果が得られたが，付録 A.1 にあるように主観的な溶け込み（Q13）や調和（Q14）の評価は改善しなかった．このギャップは，聴感上の「溶け込み」「調和」には，メル周波数ケプストラム係数で捉えられる音色特徴以外の要因が影響している可能性があり，呼吸パラメータの制御による効果範囲の限界を示唆する結果であると言える．実際にその一つである SPR では，被験者と斉唱相手との差異において，With-baseline 形態と With-proposed 形態との差は限定的である一方，With-human 形態はそれらと比較して中央値はより小さかった．この結果は，被験者と付歌唱者における合唱経験の差だけでなく，被験者は斉唱相手を知らされていないのに対して付歌唱者は人間斉唱を前提として歌唱していることに起因している可能性も示唆している．

5. まとめ

本研究では，人間–AI 斉唱において，人間歌声に作用されて呼吸パラメータをリアルタイムに制御する歌声合成を提案した．この手法を導入することで，呼吸パラメータの動的な制御が人間歌声及び人間歌手の主観的体験に与える影響について実験的に調査し，主観評価と客観評価の両方を行うことでその影響について明らかにした．具体的には，力強さや心地よさといった肯定感に関する 8 個の項目のうち 6 個，相互作用感を含めた一体感に関する 6 個の項目のうち 2 個，そして客観評価としては音色特徴の一つであるメル周波数ケプストラム係数（1 次元）において，With-baseline 形態よりも With-proposed 形態の方が With-human 形態に近い結果を示した．この結果から，これらの要素は声帯レベルの変化によって影響を受けやすい体感及び音響特徴量であることが示唆される．一方で，楽しさや緊張度，溶け込みや調和といった項目，フィルタと密接に関係する音響特徴量である SPR においては，提案手法の効果は限定的であった．これらの要素には，声道共鳴や singer's formant，合成歌声を切り替える時間粒度といった本研究で制御していない要因が支配的であることが示唆される．

また，本研究には以下の懸念点も残されている．

- 合成歌声を事前作成している点である．リアルタイムな歌声合成が技術的制約から困難であるため，合成歌声の事前作成を行ったが，歌声合成そのものをリアルタイムに行うことで，本手法を異なる楽曲にも適用しやすくなり，また 1 フレーズあたりの時間（4.5〜5.25 s）を短縮し，かつ動的にすることで，人間斉唱の相互作用をより緻密に模擬できる可能性がある．
- 呼吸パラメータと合成パラメータの対応を厳密な検証を行っていない点である．これらの対応について，語義と聴感の両方において妥当性を有し，またその仮定を支持するような結果も確認できるが，ある合成パラメータの値の変化が，それとは対応しない呼吸パラメータの値に影響を与える可能性がある．
- 人間歌手間の相互作用に対応した設計ではない点である．実験において，付歌唱者には被験者に合わせて歌唱するように指示をしたが，人間–AI 斉唱において人間歌手が複数存在する場合には，人間歌手間の相互作用を観測し，それに基づき合成歌声を変化させる必要がある．
- /a/ の音素に限定されている点である．/a/ の音素のみで歌唱する場合は稀であり，多様な母音を含む歌詞への一般化が必要である．
- 被験者数が少なく，声部や声区の統制を行っていない点である．被験者数が 7 名と限定的であり，その属性について分析を行うのは困難である．また 1 回の実験における被験者と付歌唱者の声部に条件はなく，声区の指定も行わなかったが，これらの要素が人間歌声及び人間歌手の主観的体験に与える影響として支配的である可能性がある．
- 声道共鳴が呼吸パラメータの範囲外である点である．呼吸パラメータは LF モデルを定義の根拠としているため，フィルタは範囲外である．しかし付録 A.1 に見るように，SPR といった声道共鳴に基づく音響特徴量が，斉唱の人間らしさにおける重要な因子であることを示唆する結果が確認できるため，歌声合成の制御対象に加える必要がある．

本研究は，人間–AI 斉唱において，人間歌声の影響を受けてリアルタイムに歌声合成に着目した初期的な事例である．今後は，先述した懸念点について検討するとともに，人間斉唱により近い人間–AI 斉唱を実現することで，アンサンブルの新しいあり方と，音楽への新たな向き合い方を示すことが期待される．

謝辞： 本研究は，JST 創発的研究支援事業 JPMJFR226V，JSPS 科研費 23K28108 の支援を受けて実施した．

参考文献

[1] J. Dai and S. Dixon, "Singing together: Pitch accuracy and interaction in unaccompanied unison and duet singing," *Acoustical Society of America*, vol. 145, pp. 663–675, 2019.

[2] E. C. Carterette, "Choir acoustics - an overview of scientific research published to date," *TMH-QPSR*, vol. 43, no. 1, pp. 001–008, 2002.

[3] M. A. L. F. Cuesta H, Gómez E, "Analysis of intonation in unison choir singing," *Music Perception and Cognition*, pp. 23–28, 2018.

[4] E. C. Carterette, "The science of the singing voice," *Music Perception: An Interdisciplinary Journal*, vol. 7, no. 2, pp. 187–195, 1989.

[5] H. J. . S. Ternström, "Intonation analysis of a multi-channel choir recording," *TMH-QPSR*, vol. 47, no. 1, pp. 001–006, 2005.

[6] L. M. Mann, "Effects of solo and choral singing modes on vibrato rate, extent, and duration exhibited by undergraduate female singers," 2015.

[7] T. D. Rossing, J. Sundberg, and S. Ternström, "Acoustic comparison of voice use in solo and choir singing," *Acoustical Society of America*, vol. 79, no. 6, pp. 1975–1981, 1986.

[8] T. D. Rossing, J. Sundberg, and S. Ternström, “Acoustic comparison of soprano solo and choir singing,” *Acoustical Society of America*, vol. 82, no. 3, pp. 830–836, 1987.
[9] S. Ternström and J. Sundberg, “Formant frequencies in choir singers,” *STL-QPSR*, vol. 28, no. 4, pp. 43–55, 1987.
[10] B. B. Carter, *An acoustic comparison of voice use in solo and choral singing in undergraduate and graduate student singers*. The University of Texas at Austin, 2007.
[11] V. Müller and U. Lindenberger, “Cardiac and respiratory patterns synchronize between persons during choir singing,” *PLOS ONE*, vol. 6, pp. 1–15, 09 2011.
[12] A. Hemakom, V. Goverdovsky, L. Aufegger, and D. P. Mandic, “Quantifying cooperation in choir singing: Respiratory and cardiac synchronisation,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 719–723.
[13] Y.-P. Cho, F.-R. Yang, Y.-C. Chang, C.-T. Cheng, X.-H. Wang, and Y.-W. Liu, “A survey on recent deep learning-driven singing voice synthesis systems,” in *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2021, pp. 319–323.
[14] K. Nishizawa, R. Yamamoto, W.-C. Huang, and T. Toda, “Investigating factors related to the naturalness of synthesized unison singing,” in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
[15] H. Hyodo, S. Takamichi, T. Nakamura, J. Koguchi, and H. Saruwatari, “Dnn-based ensemble singing voice synthesis with interactions between singers,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 660–667.
[16] G. Fant, “The source filter concept in voice production,” *STL-QPSR*, vol. 22, no. 1, pp. 21–37, 1981.
[17] J. P. Cabral and A. R. Meireles, “Transformation of voice quality in singing using glottal source features,” in *Workshop on Speech, Music and Mind*, 2019, pp. 31–35.
[18] H. Motoda and M. Akagi, “A singing voices synthesis system to characterize vocal registers using arx-lf model,” in *2013 International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP’13)*. 2013 International Workshop on Nonlinear Circuits, Communications and Signal ⋯, 2013.
[19] J. K. Ford, *The preference for strong or weak singer’s formant resonance in choral tone quality*. The Florida State University, 1999.
[20] 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,” 博士論文, 東京藝術大学, 9 月 2022.
[21] T. Nakano and M. Goto, “Vocalistener2: A singing synthesis system able to mimic a user’s singing in terms of voice timbre changes as well as pitch and dynamics,” 06 2011, pp. 453 – 456.
[22] R. Nishiyama and T. Nonaka, “Can a human sing with an unseen artificial partner? coordination dynamics when singing with an unseen human or artificial partner,” *Frontiers in Robotics and AI*, vol. 11, p. 1463477, 2024.
[23] J. F. Daugherty, “On the voice: Rethinking how voices work in a choral ensemble,” *The Choral Journal*, vol. 42, no. 5, pp. 69–75, 2001.

付　　録

A.1 主観評価の結果

表 1 にある 15 個の質問のうち，4.5.1 節で述べていない 12 個の質問に対する回答結果を図 A·1，図 A·2 に示す．With-proposed 形態の With-human 形態に対する中央値の差が With-baseline 形態の With-human 形態に対する差よりも小さい結果となった質問は 5 個あり，4.5.1 節で言及した質問を含めれば 8 個と，過半数の質問が該当した．また斉唱のみの質問（Q10，Q11，Q13〜Q15）では，With-human における Violin 図の形状はいずれも上方が膨らんだ形であり，どの歌唱形態でも中央値が同じであった Q14 を除いては，人間–AI 斉唱の Violin 図はこのような形状とはならなかった．

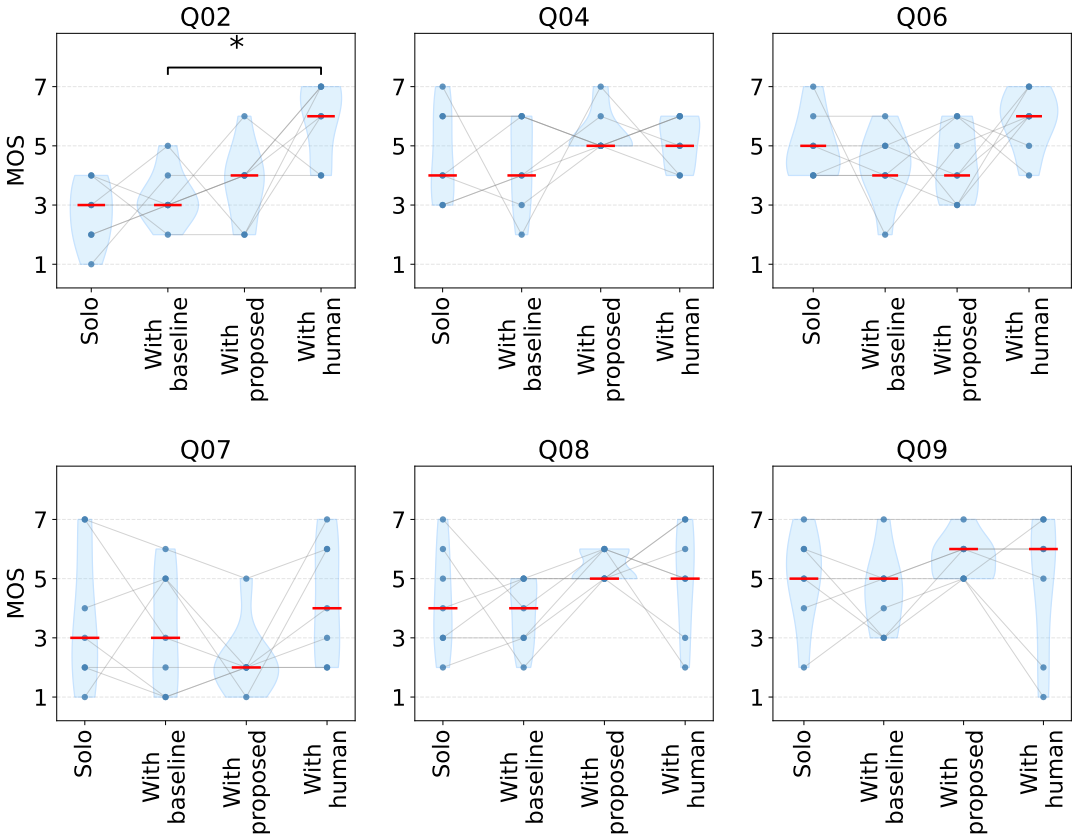

図 A·1: （a）Q02，Q04，Q06，Q07，Q08，Q09 の回答結果.

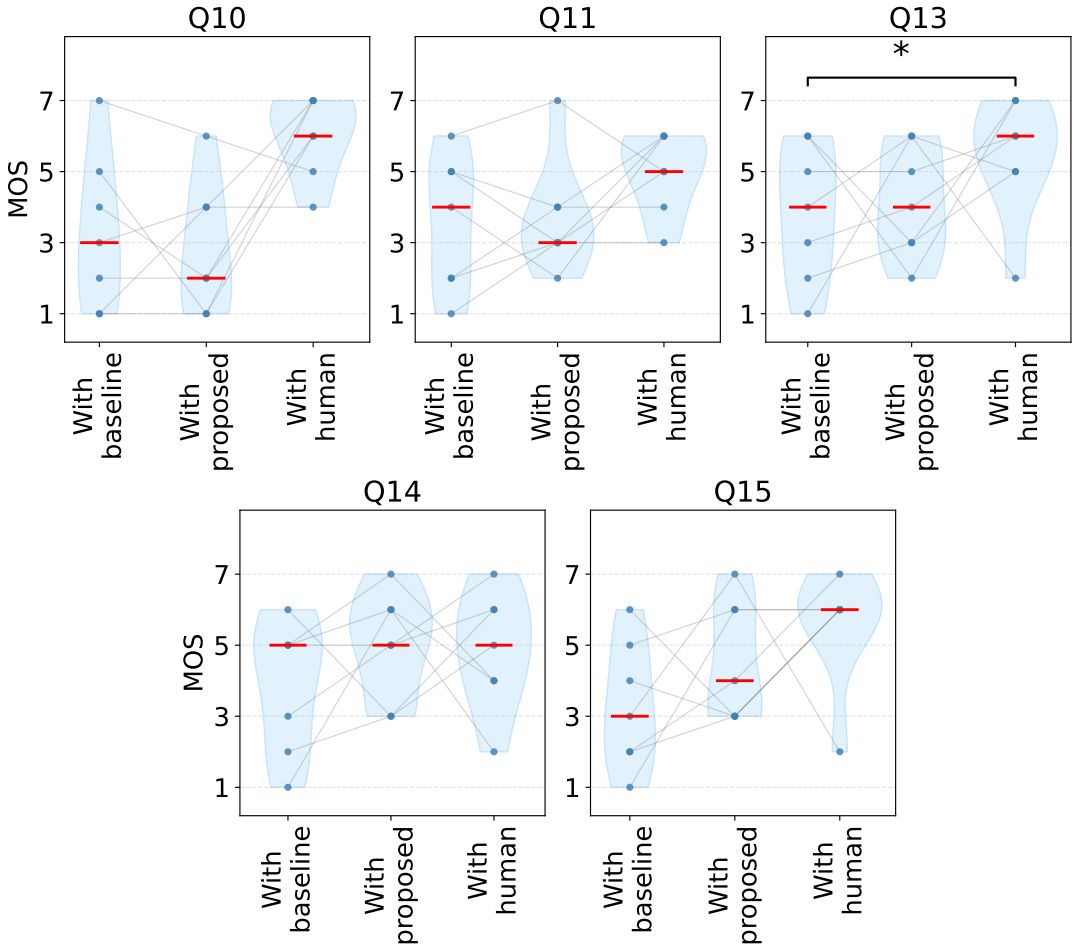

図 A·2: （b）Q10，Q11，Q13，Q14，Q15 の回答結果.