人間–AI斉唱において 合成歌声特徴量の変調が斉唱らしさにもたらす効果

三井 啓史[1,a)]　松下 嶺佑[1,b)]　深尾 貫太[1]　高道 慎之介[1,2,c)]

概要：本研究では，人間と歌声合成モデルによる斉唱（人間–AI 斉唱）において，合成歌声の特徴量操作が斉唱の一体感に与える影響を検討する．人間同士の斉唱では，歌唱者間のピッチやフォルマントに完全な一致ではない微小なばらつきが存在し，このばらつきが知覚的な一体感の形成に寄与することが知られている．一方，従来の歌声合成は独唱を前提として設計されており，人間と斉唱した場合には一体感が得られにくいという課題がある．本研究では，人間と合成歌声がそれぞれ一人ずつ参加する人間—AI 斉唱を対象とし，合成歌声のピッチを人間歌声に同調させる手法を提案する．人間斉唱における知覚実験で報告されている許容範囲に基づき，ばらつきを制御した斉唱音声を作成し，主観評価により一体感の変化を分析した．

1. はじめに

人間の歌唱形態には，独唱，斉唱，重唱，合唱の 4 形態が存在する [1]．独唱は単一パートを一人で歌唱する形態であり，斉唱は単一パートを複数人で歌唱する形態である [1], [2]．一方，重唱は複数パートを各一人が担当する歌唱形態であり，合唱は複数パートを各パート複数人で歌唱する形態である [1], [2]．本研究では，斉唱のみを扱うこととする．

斉唱時には，複数の歌手が完全に同一のピッチ・タイミングになることはなく，歌唱者間でピッチやタイミング，音色の微小差が生じ，この微小差の重なりが斉唱における一体感を生み出している [3], [4]．ここでいう一体感とは，歌唱者間の同調によって複数人の歌声が重なり合い，単一の声のように聞こえる人間の知覚に基づくものをいうこととする．斉唱形成は音響的指標と知覚的指標の両面から評価され，合唱のまとまりが聴覚的に判断されることを示している [5]．

本研究は，人間斉唱に見られる歌唱表現の同調を参考に，歌声合成モデルで合成された合成歌声と人間で構成される人間–AI 斉唱の実現を目的とする．ここでいう人間斉唱とは，複数人の人間で構成される斉唱のことをいい，人間–AI 斉唱とは，人間と合成歌声で構成される斉唱のこととする．従来の合成歌声は独唱を前提に作成されていて，人間や他の合成歌声と歌唱することは想定されていない [3]．そのため，合成歌声と人間で斉唱を行うと，個々の独立した歌唱となり，斉唱における一体感を生み出すことはできないという問題があった．人間–AI 斉唱において，本研究では人間の歌声に対して合成歌声が同調されることによって一体感を生み出すことを目指す．

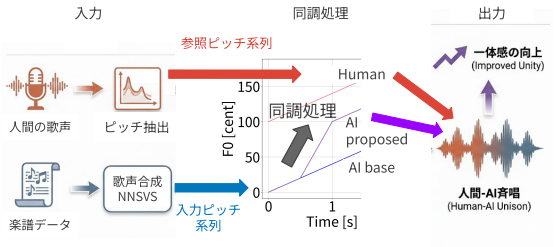

図 1　本研究の概要

本研究の意義は，人間同士の斉唱において成立している歌唱表現の同調を，人間–AI 斉唱という枠組みの中でモデル化・検証する点にある．これにより，従来は人間同士の相互作用としてのみ議論されてき

[1] 慶應義塾大学
Keio University
[2] 東京大学
University of Tokyo
[a)] kcmitt21@keio.jp
[b)] ryosuke.jp66@keio.jp
[c)] shinnosuke_takamichi@keio.jp

た斉唱の一体感を，AI が人間に合わせる能力として定式化することが可能となる．

本研究では，人間と歌声合成モデルがともに歌唱する人間–AI 斉唱において，合成歌声が人間歌唱に同調することで斉唱の一体感を向上させる枠組みを提案する．具体的には，斉唱における一体感に寄与する主要な音響的要素としてピッチに着目し，合成歌声のピッチを斉唱相手である人間の歌声に同調させる処理を導入し，人間–AI 間の音響的ばらつきを制御する．人間斉唱においては，歌手間のピッチのばらつきが人間の知覚と密接に関係し，それぞれに許容範囲および好ましい範囲が存在することが報告されている [6]．本研究では，この知覚的評価が人間–AI 斉唱にも適用可能であるという仮説を立てる．実験的評価では，同調処理を行わない人間–AI 斉唱と，同調処理を行った人間–AI 斉唱を比較し，人間の知覚において斉唱の一体感が向上するかを評価する．

2. 関連研究

2.1 歌声合成モデル

歌声合成（singing voice synthesis; SVS）モデルは，歌詞と音高，音価，テンポが記された楽譜を入力として、歌唱波形を生成する深層学習モデルを指す [7]．SVS では楽譜に明示された音高・リズムの厳密な追従と、楽譜に書かれにくいビブラートなどの歌唱表現を両立することを目的としている．代表的な歌声合成モデルには，Sinsy [8] や DiffSinger [9] などがある．ここでは，後述する実験で使用する歌声合成を構成する複数のモデルと処理フローを定義したフレームワークである NNSVS [10] についてその構造を詳述する．

NNSVS は各音素の開始時間を予測する timelag model, 各音素の持続時間を予測する duration model, これらの出力情報を含めたフレームごとの特徴量から音響特徴量を予測する音響モデル（acoustic model），音響特徴量から音声波形を出力する vocoder で構成される [10]．音響モデルでは，ピッチ，スペクトル特徴，有声/無声判定，ブレスなどを含む非周期成分をそれぞれのストリームで予測し，これら 4 つの特徴量が歌唱波形合成時に密接に関わり合うことで，人間らしい歌唱表現を含んだ歌声を合成することを達成している [10]．本研究においては，この NNSVS によって合成された歌声波形を人間と斉唱するベースライン手法の合成歌声として使用する．

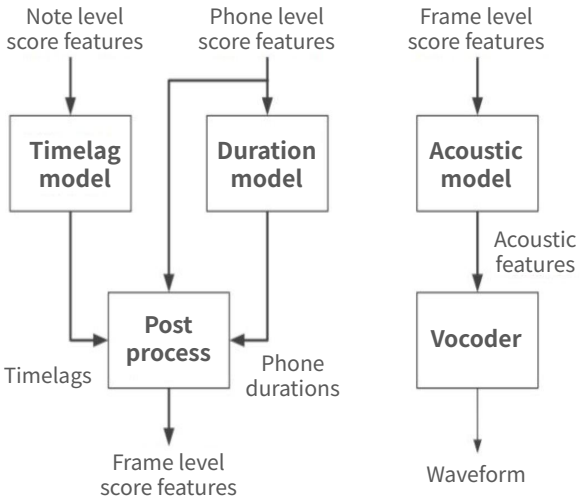

図 2 NNSVS モデル図 [10].

2.2 斉唱の一体感に関する分析

2.2.1 斉唱の同調要素の分析

人間斉唱では，歌唱者間で同調する歌唱表現としてピッチ，タイミング，フォルマントなどが挙げられ，これらの要素が歌唱者間でどのように同調されているのかが以前から研究されてきた [11], [12], [13], [14]．先行研究では，人間斉唱において，複数の歌唱者は互いにピッチ，タイミングの調整を行い，その結果として歌唱の相互作用が発生することが確認されている [11], [12]．特にピッチについては，他者に合わせて素早くピッチを修正する能力があることを実験的に検証し，急なピッチ変化に対して追従することが示されている [13]．また，斉唱において歌唱者が互いの歌声を聴取できるか否かという条件を操作した結果，ピッチ誤差が変化することが報告されており，歌唱の相互作用がピッチ精度に影響を及ぼすことが示されている [12]．これらより，合唱品質はピッチのずれによって決められるといえる．

これらの研究から，人間斉唱では，ピッチの同調が主要で基本的な戦略とされる．本研究においても，斉唱時に合成歌声が人間に同調される要素として，基本的な同調対象とされるピッチを考えることとする．

2.2.2 斉唱のばらつきと一体感の知覚の分析

人間斉唱において，人間が斉唱における一体感を感じることができるピッチとフォルマントのばらつきが存在するとする先行研究がある [6], [11]．斉唱の一体感の成立には許容できる範囲と好ましく感じる範囲が区別され，歌唱者のピッチのばらつきに関しては許容範囲は標準偏差で約 14 cent までであり，好ましく感じる範囲は標準偏差で 5 cent 以下であるということが示されている [6]．さらには，歌唱者間でのピッチのずれが 30 cent を上回ると斉唱として

の質が低下する [11]．また，斉唱における歌唱者間の第 3 フォルマント周波数から第 5 フォルマント周波数の全体の平均に対するばらつきは，許容範囲は標準偏差で 12 %，好ましく感じる範囲は標準偏差で 7 %とされている [6]．本研究におけるピッチとフォルマントの補正についても，先行研究の許容範囲と好ましく感じる範囲についての値を利用し，人間–AI 斉唱でもこれらの値が適用できるのか検証することとする．

2.3 ピッチ補正の従来手法

ピッチ補正とは，歌唱や演奏において生じる音高のずれを検出し，目標とする音高に近づくよう基本周波数を調整する処理を指す．これは，音程の正確性を向上させることを目的として，音楽制作や音声処理の分野で広く用いられている技術である．このうち，DPW（dynamic pitch warping）[15] という技術は，歌声のピッチを楽譜の音符のピッチに合わせるためのピッチ補正技術として，歌声補正の研究で参照されている [16], [17]．DPW は目標とするピッチに寄せつつビブラートなどの歌唱表現を潰さない設計思想に基づくものである [16]．

DPW は，入力 F_0 軌跡を目標とする F_0 軌跡へ“動的にワープ（warping）”することで補正する．補正は常時最大量でかけるのではなく，F_0 の時間変化（例：ピッチ速度）に基づき，ビブラート等の微細変動を保持しつつ主たるピッチを目標へ近づけるよう設計されている [16]．そして，ピッチ補正は、ピッチが周波数軸の一定範囲内に臨界時間以上とどまっている場合にのみ発動するため、ピッチが安定しており、歌唱者が目標とするピッチを意図的に狙っている場合に限って補正が行われる [16]．また，ピッチ補正は補正対象の歌声が有声である区間のみで行われ，無声の区間では行われない [16]．

本研究では DPW を単純化して動的なピッチ補正の代わりに静的なピッチ補正を使用することとした．

3. 提案手法

本研究では，歌声合成モデルから得られた合成歌声に対してピッチの補正処理を加えることによって，合成歌声を斉唱相手の人間の歌声に同調させる．以下に，本研究で用いる手法の詳細を述べる．

3.1 ピッチの補正

人間歌声の F_0 系列 $p_{\mathrm{human}}[t]$ と 合成歌声の F_0 系列 $p_{\mathrm{synth}}[t]$ を用いて，補正後の合成歌声の F_0 系列 $p'_{\mathrm{synth}}[t]$ を求める．t はフレームインデックスである．具体的には

- 補正強度の最大値：$\alpha_{\max}$
- 補正開始時間：T_{start}
- 補正漸近時間：T_{shift}

を定め，以下のように補正する．

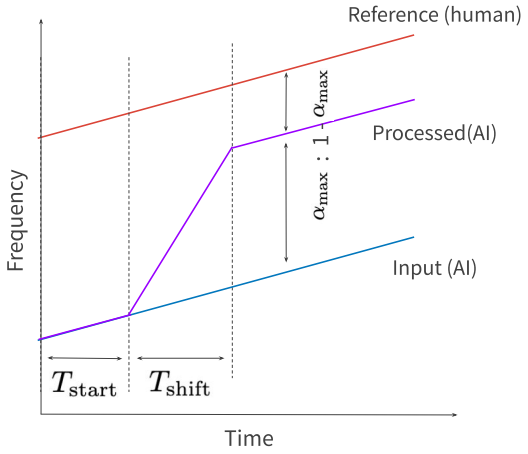

図 3 各パラメータ変化とその影響

$$p'_{\mathrm{synth}}[t] = (1-\alpha(t))\,p_{\mathrm{synth}}[t] + \alpha(t)\,p_{\mathrm{human}}[t], \tag{1}$$

$$\alpha(t) = \begin{cases} 0, & 0 \le t < T_{\mathrm{start}} \\ \alpha_{\max} \cdot \dfrac{t - T_{\mathrm{start}}}{T_{\mathrm{shift}}}, & T_{\mathrm{start}} \le t < T_{\mathrm{start}} + T_{\mathrm{shift}} \\ \alpha_{\max}, & T_{\mathrm{start}} + T_{\mathrm{shift}} \le t \end{cases} \tag{2}$$

式 (1) のように，人間歌声と合成歌声の音高系列を時間的に重み付け平均することで，合成歌声のピッチを人間歌声へ段階的に接近させる．本節では，補正関数 $\alpha(t)$ に含まれる各パラメータが接近挙動に与える影響の例を図 3 に示す．図の Input の黒線が補正前のピッチ系列，Reference の黒線が参照のピッチ系列を表している．

図 3 に，ピッチ補正パラメータ（$\alpha_{\max}$, T_{start}, T_{shift}）を基準設定から増減させた際の F_0 系列の変化を示す．$\alpha_{\max}$ は参照（人間歌声）への補正割合，T_{start} は補正開始時刻，T_{shift} は補正漸近時間をそれぞれ制御する．各パラメータを大きくすると参照への追従が強く，早く，緩やかに進み，小さくすると入力（合成）F_0 の保持が優先される．

4. 実験的評価

4.1 実験条件

人間–AI 斉唱においてピッチの一体感に関する影

表 1　各手法の主観評価結果

手法	全体平均	赤とんぼ	ふるさと	春が来た	男性	女性
Baseline	3.38	4.08	2.34	3.70	3.58	3.17
Exact-static	3.43	3.53	2.67	4.10	3.06	3.81
Approximate-static	3.38	4.02	2.32	3.81	3.10	3.66
Exact-progressive	3.54	3.88	2.56	4.17	3.42	3.65
Approximate-progressive	3.54	3.88	2.77	3.97	3.23	3.86

響を実験的に調査した．

4.1.1　楽曲

実験に使用する楽曲は 3 曲とし，ピッチが高くなく，隣り合う音符間で音程の跳躍が少ない楽曲，BPM は 100 程度，使用言語は日本語，という 4 条件を満たす楽曲を選曲した．これらの条件を踏まえ，使用する楽曲は “赤とんぼ” [18]，“ふるさと” [19]，“春が来た” [20] の 3 曲を使用した．

4.1.2　被験者

人間歌声を歌唱する被験者属性は，上記で述べた 3 つの使用楽曲を歌唱した経験があり，日本語母語話者である条件を満たした男女それぞれ 3 人とした．

斉唱相手の合成歌声は同性の歌声とし，女性は男性より 1 オクターブ高いピッチで歌唱する．

4.1.3　合成歌声

本研究では，合成歌声を既存歌声合成モデル NNSVS [10] を使用し，3 つの使用楽曲の歌声を，男性声，女性声の両者について合成した．使用した合成歌声のレシピは，男性声を “natsume-singing” [21]，女性声を “kiritan-singing” [22] とした．

4.1.4　歌唱時

被験者には歌声を録音された歌声がはっきりと聞こえるような環境で事前に録音してもらう．

歌唱時には，各性別用のピアノによるガイド音源をイヤホンで聞きながら歌唱してもらった．ガイド音源のピッチと歌唱ピッチが完全に一致する必要はないが，歌唱オクターブとテンポをメロディと合わせるように歌唱してもらった．

録音の後処理として，MyEdit [23] を用いて環境音除去を行い Audacity [24] を用いて，合成歌声と録音歌声（人間歌声）の音量を揃えた．また，歌唱開始タイミングについても両歌声について手作業で揃えた．

4.2　実験方法

人間–AI 斉唱において，合成歌声が人間歌声にピッチを同調させたときの斉唱歌唱の一体感の向上について調査する．ピッチ補正前の合成歌声と人間歌声による人間–AI 斉唱をベースライン手法とした．

提案手法として，以下の 4 手法を提案する．

- **Exact-static**：式 (2) の $T_{\text{start}} = 0$, $T_{\text{shift}} = 0$, $\alpha_{\max} = 1$
- **Approximate-static**：式 (2) の $T_{\text{start}} = 0$, $T_{\text{shift}} = 0$，std ≤ 14 cent を満たす $\alpha_{\max}$
- **Exact-progressive**：式 (2) の $T_{\text{start}} = 200$, $T_{\text{shift}} = 1000$，$\alpha_{\max} = 1$
- **Approximate-progressive**：式 (2) の $T_{\text{start}} = 200$, $T_{\text{shift}} = 1000$, std ≤ 14 cent を満たす $\alpha_{\max}$

$T_{\text{start}} = 200$, $T_{\text{shift}} = 1000$ は，Grell の研究 [13] で述べられている補正開始時間と補正漸近時間の被験者の平均値として紹介してあるため，本研究でもこの値を使用することとした．また，std は人間歌声と合成歌声の F_0 差の標準偏差のことを指す．

4.3　評価結果

本研究では，人間歌声と合成歌声の斉唱（人間–AI 斉唱）における一体感を評価した．評価者は 39 名であり，各斉唱歌声に対し斉唱の一体感の度合いを 5 段階で評価してもらった．全体的な評価結果を表 1 に示す．

人間–AI 斉唱の自然さについて，5 段階 MOS による主観評価を実施した．MOS は順序尺度であり正規性を仮定しにくいことから，ノンパラメトリック検定を用いた．まず 5 条件間の差の有無を検討するため，Kruskal–Wallis 検定 [25] を適用した．さらに事後比較として，Baseline と各提案手法（4 条件）との二群比較を Mann–Whitney の U 検定（両側検定）[26] により行った．複数の事後比較に伴う第一種過誤の増大を抑えるため，多重比較補正として Holm 法 [27] を用いた．有意水準は $\alpha = 0.05$ とした．

Kruskal–Wallis 検定の結果，5 条件間に有意差は認められなかった（$H = 3.84$, $p = 0.428$）．

- H：Kruskal–Wallis 検定の検定統計量
- p：その検定統計量に対応する p 値

また Baseline と各提案手法の事後比較（Mann–Whitney の U 検定，両側）においても，Holm 補正後の p 値はいずれも有意水準に達しなかった．

- Exact-static：$p_{\text{adj}} = 0.625$
- Approximate-static：$p_{\text{adj}} = 0.885$
- Exact-progressive：$p_{\text{adj}} = 0.625$
- Approximate-progressive：$p_{\text{adj}} = 0.625$

以上より，本実験条件下では，いずれの提案手法も Baseline に対して統計的に有意な MOS の改善は確認できなかった．

4.3.1　楽曲差の影響

提案手法の効果が特定の楽曲に依存するのかを考

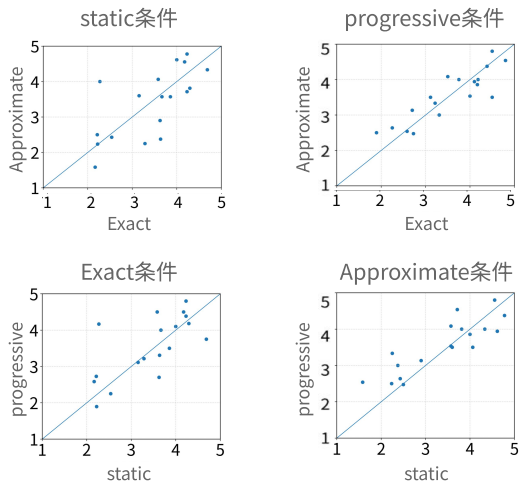

図 4 パラメータを変えたときの MOS 値の変化

察する．表 1 を見ると，ふるさとの評価が他 2 つの楽曲，赤とんぼ，春が来たと比べて顕著に低い評価となった．

“ふるさと” について，提案手法の評価のほうがベースライン手法の評価よりも上回っている．隣接音符間のピッチ変動が他 2 曲より小さいため，ピッチやタイミングの微小な差が知覚されやすく，その結果，提案手法による補正効果が相対的に強調されたからであると考えられる．一方で，“ふるさと” はロングトーンが多く曲長も長いことから，ビブラート等を含む持続区間におけるピッチの細かなゆらぎの不一致が目立ちやすい．このため，斉唱としての一体感を成立させる難易度自体が他 2 曲より高く，総合的な評価が低めにとどまった可能性がある．

4.3.2 性別差の影響

提案手法の効果が両性別の人間歌手に対し同様に機能するかどうかを考察する．表 1 を見ると，男性歌唱はベースライン手法が全ての提案手法と比べて高評価であり，女性歌唱は全ての提案手法がベースライン手法と比べて高い評価であった．

この差異は，男女の F_0 帯域や声質の違いにより，F_0 帯域に応じたパラメータ調整や適応的制御の必要性を示唆している．

4.3.3 パラメータ $\alpha_{\max}$ の影響

補正強度を制御するパラメータ $\alpha_{\max}$ が斉唱の自然さに与える影響を考察する．T_{start} および T_{shift} を同一に設定した条件下で, Exact-static と Approximate-static，ならびに Exact-progressive と Approximate-progressive を比較した結果を，図 4 上段の 2 つの図に示す．図 4 の各点はそれぞれの人間–AI 斉唱の MOS の値を意味し，対角線はある人間–AI 斉唱 2 つの手法の MOS が一致する線を表している．

上段の横軸，すなわち Exact 条件の MOS が比較的低いデータでは，Approximate 条件において MOS が向上する点が多く観測されるのに対し，Exact 条件の MOS が高いデータでは，MOS が低下する点も見られる．これは，人間歌声と合成歌声のピッチ系列が完全に一致する条件（$\alpha_{\max} = 1$）による強いピッチ同調が必ずしも常に有効ではなく，特に斉唱の一体感が十分に形成されていない条件においては，許容範囲内のばらつきを残すことで自然性が改善される可能性を示している．一方，Exact 条件である横軸で，すでに高い一体感が得られている斉唱では，補正割合を 1.0 から小さくすることが斉唱としての一体感を損なう要因となり得る．これらの結果から，$\alpha_{\max}$ の最適な値は曲・歌唱者依存で、固定の最適値が一意に定めにくく，一体感に大きな差がないことが示唆される．

4.3.4 パラメータ T_{start} と T_{shift} の影響

補正時間を制御するパラメータ T_{start} と T_{shift} が斉唱の自然さに与える影響を考察する．$\alpha_{\max}$ を同一に設定した条件下で，Exact-static と Exact-progressive，ならびに Approximate-static と Approximate-progressive を比較した結果を図 4 の下段の 2 つの図に示す．

この 2 つの図の多くのデータ点が対角線より上側に位置しており，この傾向は static 条件と比較して progressive 条件において強く見られる．これは，補正開始時間 T_{start} および補正漸近時間 T_{shift} を導入することで，歌唱開始直後からピッチの同調が実現できているという違和感が緩和され，人間同士の斉唱に近い時間的同調挙動が再現されたためであると考えられる．すなわち，最終的なピッチの一致度そのものよりも，どのような時間構造で同調が進行するかが，斉唱の自然性に大きく影響している可能性が示唆される．

一方，MOS が低い static 条件の斉唱については，時間的に漸近的な同調が有効に機能し，斉唱の一体感を改善する効果が確認される．この結果は，人間–AI 斉唱において常に同一の同調パラメータを適用するのではなく，斉唱の一体感の状態に応じて補正強度や時間構造を調整する適応的な制御が重要であることを示している．これは，開始直後に聴覚上の違和感が集中しやすく，その時点でのピッチ差が不自然さを抑制していることを示唆している．

5. まとめ

本研究では，人間–AI 斉唱において，合成歌声が人間歌声にピッチを同調されることが斉唱の一体感

に与える影響を検討した．具体的には，人間斉唱の先行研究で指摘されているピッチのばらつきに着目し，合成歌声のピッチを人間歌声に段階的に接近させる手法を提案した．

結果として，提案手法はベースラインと比較して，人間–AI 斉唱の一体感を向上させる傾向が確認された．特に，ピッチを完全に一致させる場合だけでなく，許容範囲内のばらつきを残した同調や，時間的に漸近的な同調を導入することで，一体感が改善される条件が存在することが示された．

今後の課題としては，提案手法がオフライン処理を前提としている点が挙げられ，リアルタイムな人間–AI 斉唱への拡張が求められる．また，ピッチ以外のタイミングや音色といった歌唱要素の同調が斉唱の一体感に与える影響についても，統合的に検討する．

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] 三. 美稲子, “声楽曲の音楽鑑賞教材化に関する研究,” **教材学研究**, vol. 21, no. 0, pp. 225–232, 2010.

[2] S. Ternström, “Choir acoustics: An overview of scientific research published to date,” *International Journal of Research in Choral Singing*, vol. 1, no. 1, pp. 3–12, 2003. [Online]. Available: https://acda-publications.s3.us-east-2.amazonaws.com/IJRCS/volumeone/ijrcs1_1_ternstrom.pdf

[3] P. Chandna, H. Cuesta, and E. Gómez, “A deep learning based analysis-synthesis framework for unison singing,” in *ISMIR*, 2020. [Online]. Available: https://arxiv.org/abs/2009.09875

[4] H. Daffern and A. J. Gully, “Assessing articulatory perturbations during a sung vowel-matching exercise using articulography,” *Biomedical Signal Processing and Control*, vol. 67, p. 102546, 2021.

[5] W. L. Goodwin, “Select acoustic and perceptual measures of choral formation,” *Journal of Research in Music Education*, vol. 28, no. 3, pp. 173–184, 1980.

[6] S. Ternström, “Perceptual evaluation of voice scatter in unison choir sounds,” *Journal of Voice*, vol. 7, no. 2, pp. 129–135, 1993.

[7] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Singing voice synthesis based on deep neural networks,” in *Interspeech*, 2016.

[8] K. O. Y. N. K. T. Yukiya Hono, Kei Hashimoto, “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.

[9] Y. R. F. C. Z. Z. Jinglin Liu, Chengxi Li, “Diffsinger: Singing voice synthesis via shallow diffusion mechanism,” in *AAAI, vol. 36, no. 10*, 2022, p. 11 020– 11 028.

[10] T. T. Ryuichi Yamamoto, Reo Yoneyama, “NNSVS: A neural network-based singing voice synthesis toolkit,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[11] K. Tokuda and H. Kuwahara, “Synchronization analysis of choir singing,” *Nonlinear Theory and Its Applications, IEICE*, vol. 3, no. 1, pp. 18–30, 2012.

[12] J. Dai and S. Dixon, “Singing together: Pitch accuracy and interaction in unaccompanied unison and duet singing,” *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 663–675, 2019.

[13] A. Grell, J. Sundberg, and S. Ternström, “Rapid pitch correction in choir singers,” *The Journal of the Acoustical Society of America*, vol. 126, no. 1, pp. 407–413, 2009.

[14] Y. Nishizawa, R. Yamamoto, and S. Takaki, “Investigating factors related to the naturalness of synthesized unison singing,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[15] O. Perrotin and C. d’Alessandro, “Target acquisition versus expressive motion: Dynamic pitch warping for intonation correction,” *ACM Transactions on Computer-Human Interaction*, vol. 23, no. 2, pp. 1–31, 2016.

[16] C. Molina Villota, O. Perrotin, and C. d’Alessandro, “Dynamic pitch warping for expressive vocal retuning,” in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2023, pp. 1–8.

[17] J. Hai, R. Huang, Y. Zhang *et al.*, “Diff-pitcher: Diffusion-based singing voice pitch correction,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[18] 三. 露風 and 山. 耕筰, “赤とんぼ,” 日本歌曲, 1927, 作詞: 三木露風，作曲: 山田耕筰.

[19] 高. 辰之 and 岡. 貞一, “ふるさと,” 日本唱歌, 1914, 作詞: 高野辰之，作曲: 岡野貞一.

[20] ——, “春が来た,” 日本唱歌, 1901, 作詞: 高野辰之，作曲: 岡野貞一.

[21] T. Shirani, “nnsvs_natsume_singing,” https://github.com/taroushirani/nnsvs_natsume_singing, 2023, accessed: 2025-12-17.

[22] T. Ogawa and M. Morise, “Tohoku kiritan singing database: A singing database for statistical parametric singing synthesis using japanese pop songs,” *Acoustical Science and Technology*, vol. 42, no. 3, pp. 140–145, May 2021.

[23] MyEdit, “Myedit,” https://myedit.online/jp/create?type=image, 2024, accessed: 2025-01-10.

[24] ——, “Myedit,” https://www.audacityteam.org/.

[25] W. H. Kruskal and W. A. Wallis, “Use of ranks in one-criterion variance analysis,” *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[26] H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” *The Annals of Math-*

ematical Statistics, vol. 18, no. 1, pp. 50–60, 1947.
[27] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.