人は歌声合成ソフトと歌えるか
：歌唱時における音声の協調ダイナミクスの検討
Can a human sing with an unseen artificial partner?
: Coordination dynamics of when singing with an unseen human or artificial partner

西山 理奈†，野中 哲士†
Rina Nishiyama, Tetsushi Nonaka
†神戸大学大学院人間発達環境学研究科
Graduate School of Human Development and Environment, Kobe University
239d106d@stu.kobe-u.ac.jp

概要
本研究では，視覚情報なしで，人が歌声を聴きながらそれに合わせて一緒に歌うとき，その歌声が人間か歌声合成ソフトかによって，歌声の協調パターンが異なるのかどうかを調べた．相互相関分析で，相手と参加者の各歌声の振幅包絡線の時系列相関を比較し，グレンジャー因果性検定で，予期的なダイナミクスを調べた．その結果，視覚情報がなくても，参加者は，人間と歌う方がより先読みして歌唱行動を同期させることが示された．

キーワード：個人間協調，予期的同期，相互作用，歌声合成ソフト

1. はじめに

音楽の演奏において，演奏者同士が音を同調させる能力には，他者の行動や音に対して知覚と予測を行うだけでなく，継続的なタイミング調整と，予測制御の重要性も含まれている（Proksch et al., 2022）．このことから，音楽は，複雑な相互作用のダイナミクスについて調べる点で，理想的なモデルとみなされている（Proksch, et al., 2022）．また，音楽は，社会的な関わりに関連する知見を提供できる（McCallum and McOwan, 2018）ことから，人間と機械の相互作用を理解するための潜在的な意味を持つと言える．

近年の情報科学技術における急速な進歩とともに，人間と機械の相互作用問題に関する重要性も指摘されている（Mara and Appel, 2015）．そこで，音声などの，音を介した相互作用に焦点を当てた研究が行われているなか，人間のような音響的特徴を生成する技術も飛躍的に発展している（Umbert et al., 2015）．その例の 1 つが歌声合成技術である．しかし，協調的な演奏というのは，身体動作を介して生じる音に対し，演奏者同士で互いに適応することで，相互作用が成り立つ（Konvalinka et al., 2010），という風に，身体を使った複雑な運動協調と同期を要するため，人間と機械の協調的な音楽演奏は，依然として難題である（Wang et al., 2024）．そのため，機械を用いた楽器演奏での協調に関する研究は多く行われている．一方で，歌声合成ソフトと「一緒に歌う」ことが実社会で増加しているにもかかわらず，そのことをテーマとする研究はほとんど進んでいない．そこで，人は，人間の身体を持たない人工的なパートナーが作り出す歌声と協調し，一緒に歌うことができるのかどうか，という問題を提示した．

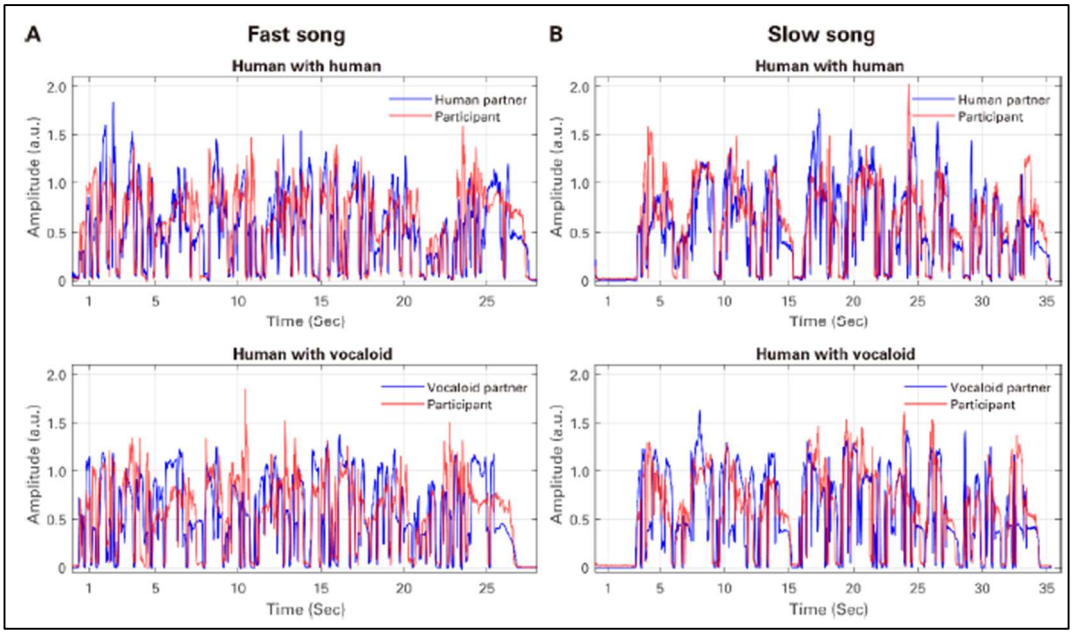
本研究では，視覚情報なしで，歌唱者が，人間，あるいは歌声合成ソフトと歌う場合，歌声の振幅包絡線に関する時系列変化パターン間の類似性と同期性は，パートナーによって異なるのかどうかということについて，歌声合成ソフトウェア「VOCALOID6」（ヤマハ株式会社，2022）を用いて，検証した．また，歌声合成ソフトは，視覚手がかりである身体表現を持たないため，視覚情報は無しで統一し，歌声の協調ダイナミクスに焦点を当てた．

2. 方法

2.1. 参加者

7 名（平均年齢：22.9 歳，SD=0.8）の参加者を募集し，実験で使用した合成音声が女性の声だったため，全員女性とした．また，人間が歌っている音源を作成するために，女性歌手を 1 名採用した．

図 1. 2 つの条件（人間のパートナー・人工的なパートナー：青線）と参加者（赤線）の、歌声の強弱における時系列変化を表した振幅包絡線の例. (A) Fast song (B) Slow song（それぞれ参加者 ID P5 の 5 回目の試行から）



2.2. 刺激

使用した楽曲は，J-POP の *Sincerely*（75BPM，4/4 拍子：「Slow song」）と *少女レイ*（150BPM，4/4 拍子：「Fast song」）で，実験では，各曲から抽出した部分を使用した（Slow song：10 小節，37 秒，Fast song：17 小節，29 秒）．また，原曲において，*Sincerely* は TRUE の女性ボーカリストである唐沢美帆，*少女レイ*は VOCALOID の初音ミクが歌っている．各曲で，人間が歌い手の音源と，VOCALOID が歌い手の音源を用意した．VOCALOID は，VOCALOID6 に含まれている VOCALOID AI（HARUKA）の歌声を使用し，譜面ではなく原曲を聴きながら，筆者自身が打ち込みで作成した．両方の録音には，歌が始まる前に 2 小節分のクリック音を含めた．そのため，VOCALOID は一定のテンポで歌うが，人間は最初の 2 小節分のクリック音を聴いた後，彼女自身のテンポ感で歌っていた．

2.3. 手順

参加者には，実験の 1 週間前に原曲で練習してくるように伝えた．実験では，一緒に歌う相手は見えないが，ヘッドフォンから聞こえてくる歌声にできるだけ合わせて歌うように指示した．また，パートナー（録音）ーと参加者の歌い始めのタイミングを合わせるため，2 小節分のクリック音を聴いてから歌ってもらった．各パートナー条件（人間と VOCALOID）で，各曲（Slow Song と Fast Song）を 5 回ずつ録音し，各曲の中で，2 つのパートナー条件の試行順番は無作為化された．

2.4. 分析

本研究では，歌という協調行動における重要な指標の一つとして，振幅（抑揚）の時間変化に着目した．ピッチや発声タイミングも検討対象となり得る．しかし，VOCALOID と人間の構造的な違い（例：ピッチの正確性，発声時の音の立ち上がり方）から，これらの指標は，本研究の目的には即してないと判断した．

比較する歌声の要素としては，ピッチ（音程）や発声タイミングなども可能である．しかし，歌声合成ソフトと人間の，発声タイミングは必然的に合わないため，タイミングのズレを比較しても望ましい結果は得られないだろう．また，ピッチを比較する場合は，歌声合成ソフトの歌声を，人力で打ち込むとはいえ，人間よりもかなり正確な音程であることに変わりはない．そのため，ピッチがどれくらい合っているのか，を比較する際は，「ガイドボーカル」に関する目的が適している．そこで，歌声の振幅包絡線に関する時系列変化パターンを見ることは，「2 人で一緒に歌う」という目的を達成するための最も適している指標だと考える．

まず，全ての歌声を，MATLAB に読み込み，16kHz でリサンプリングした．各歌声の強弱を表す振幅の時間変化を明確にするため，ヒルベルト変換した波形に

ついて，振幅の大きさを計算した．32 Hz のカットオフ周波数を持つ 3 次バターワース IIR フィルタを，各時系列データに 2 回適用し，歌声の振幅における遅い時間変調（32 Hz 未満）を分離した．これにより，声質などに関わる高周波数成分を除去することが可能となった．その後，200 Hz にダウンサンプリングし，連続する，重ならない 5ms のウィンドウ内の時点を平均化した．

次に，相互相関（CC: Cross correlation）分析を行った．この分析では，参加者とパートナーの歌声の強弱における時間的変調パターン間の類似性を評価した．そのため，各試行中の各パートナーの振幅包絡線時系列と各参加者のパフォーマンス間の CC 係数を計算した．CC 係数は，-0.1～0.1 秒のラグについて計算し，最大 CC 係数時のラグも算出することで，2 つの時系列間の時間差も示した．また，ラグが 0 のときの CC 係数も算出し，2 つの時系列データ間の同期性も評価した．

そして，グレンジャー因果性（GC: Granger causality）検定を行った．GC とは，時系列データに適用される統計的因果関係の概念であり，「ある変数の履歴が他の変数の予測に有用か」を評価するための手法である．この分析を行うために，多変量グレンジャー因果性ツールボックス（Barnett and Seth, 2014）の手順に従い，歌声における振幅包絡線の時系列変化パターンについて，パートナーから参加者へ，およびその逆方向の GC 値を計算した．相手から参加者の GC 値が高いときは，参加者が相手の歌声を聞いてから，それに合わせるようにして歌っていることを示唆し，逆方向の場合は、参加者が相手の歌声を予期的に予測して歌っていることを示唆する．

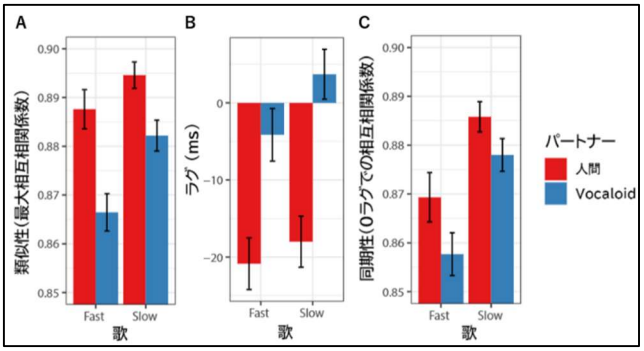
統計分析では，前述の 3 つの CC 分析の結果を結果変数として線形混合効果モデルでモデル化し，固定効果は，パートナーとテンポとした．同様に，パートナー，テンポ，GC の方向を固定効果因子とする線形混合効果モデルでは，GC 値を結果変数としてモデル化した．また，参加者因子を，両モデルの切片のランダム効果として含めた．有意な効果の α 値は 0.05 とした．

3. 結果

参加者は，人間と歌う条件で（$F_{(1,130)} = 91.30$，$p < 0.0001$），また，Slow song で（$F_{(1,130)} = 41.87, p < 0.0001$），歌声における強弱の時間的変調パターンの類似度が高いことが示された（図 2A）．さらに，参加者は，人間が相手だと，相手に先行する形で協調して歌っていたが，VOCALOID と一緒に歌った場合，参加者はあまり先行しなかった（$F_{(1,130)} = 184.62, p < 0.0001$）（図 2B）．

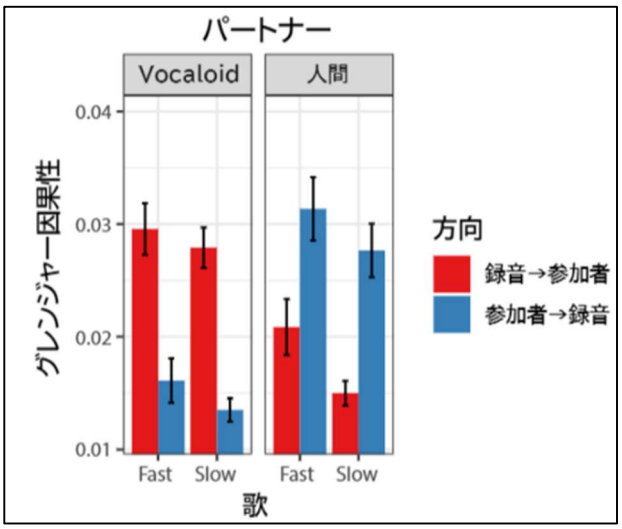
同様の傾向は，パートナーと参加者の，歌声の振幅包絡線に関する時系列変化パターン間の同期程度を示す 0 ラグの CC 係数でも観察され（図 1C），人間と歌う時の方が，同期していた（$F_{(1,130)} = 16.62, p = 0.0001$）．

図 2. 歌声の強弱における時系列変化パターン間の CC 分析の結果．（A）最大 CC 係数値（B）最大 CC 係数時のラグ（C）0 ラグでの CC 係数値



また，GC 検定の結果（図 3）を見てみると，参加者が VOCALOID のパートナーに合わせて歌う場合，参加者は概ねパートナーに追従し，人間がパートナーの時は，先取りのダイナミクスを示した（$F_{(1,266)} = 86.65$, $p < 0.0001$）．

図 3. グレンジャー因果性検定の結果



4. 考察

本研究の実験では，視覚的情報がなく，パートナーである人間の歌には微妙なテンポの変動があった．しかし，参加者は，自分の歌唱行為を，人間のパートナーの方により予期的に同期させ，歌声の振幅包絡線に

おける時系列変化について，より近い形になるように導いていることが示唆された．

その理由として，人間の歌声には，近い将来において，歌声のダイナミクスの展開がどうなるのかを予測しやすくなるような手がかり（例：ブレス音）が含まれていたことが考えられる．パートナーである人間は，このような聴覚的手がかりが顕著であり，参加者がそれを知覚することで，予期的な同期が実現した可能性がある．一方で，人工的なパートナーであるVOCALOIDには，参加者が歌を予期的に協調することを可能にするような手がかりが欠如しているため，相手を先取りしてあわせる協調ダイナミクスの特徴が失われたと考えられる．なお，発声時における母音と子音の強調の違いや，歌詞の区切り方なども，協調に影響を及ぼす要因として考えられる．そのため，ブレスに限らず他の手がかりについても検討していく必要がある．ただし，本研究における参加者へのフリーインタビューでは，「息づかい」の違いに関する指摘が多かったため，ブレスの影響について，今後優先的に検討を進めたい．

また，今後は，一緒に歌う相手として複数の人間の歌い手を用いる実験や，人間同士で歌う実験も行い，人間特有の協調パターンをより明確に検討していく．さらに，本研究では人間の歌い方に近づけたVOCALOIDの歌声を使用したが，より機械的な歌声も比較対象に加えることで，人間と人工的なエージェントの本質的な違いを明瞭にし，円滑な協調行動を実現するための知見の獲得を目指す．

文献

Barnett, L., and Seth, A. K. (2014). The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. J. Neurosci. Methods 223, 50-68. https://doi.org/10.1016/j.jneumeth.2013.10.018

Konvalinka, I., Vuust, P., Roepstorff, A., and Frith, C. D. (2010). Follow you, follow me: Continuous mutual prediction and adaptation in joint tapping. Q. J. Exp. Psychol. 63(11), 2220–2230. https://doi.org/10.1080/17470218.2010.497843

Mara, M., and Appel, M. (2015). Effects of lateral head tilt on user perceptions of humanoid and android robots. Comput. Hum. Behav, 44, 326–334. https://doi.org/10.1016/j.chb.2014.09.025

McCallum, L., and McOwan, P.W. (2018). Extending Human–Robot Relationships Based in Music With Virtual Presence. IEEE Transactions on Cognitive and Developmental Systems, 10, 955-960. https://doi.org/10.1109/TCDS.2018.2790921

Proksch, S., Reeves, M., Spivey, M., and Balasubramaniam, R. (2022). Coordination dynamics of multi-agent interaction in a musical ensemble. Sci. Rep, 12, 421. https://doi.org/10.1038/s41598-021-04463-6

Umbert, M., Bonada, J., Goto, M., Nakano, T., and Sundberg, J. (2015). Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges. IEEE Signal Process. Mag, 32, 55-73. https://doi.org/10.1109/MSP.2015.2424572

Wang, H., Zhang, X., and Iida, F. (2024). Human–Robot Cooperative Piano Playing With Learning-Based Real-Time Music Accompaniment. IEEE Transactions on Robotics, 40, 4650-4669. https://doi.org/10.1109/TRO.2024.3484633