人間または歌声合成ソフトとの歌唱における協調ダイナミクス －視覚情報の役割－

Coordination Dynamics in Singing with Human and Artificial Partners: The Role of Visual Information

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
Kobe University, Graduate School of Human Development and Environment
239d106d@stu.kobe-u.ac.jp

概要

本研究では，人が，人間または人工パートナーと歌う場合で，歌唱における協調ダイナミクスがどう異なるのかを検証し，これらの協調に対する視覚情報の影響も調べた．参加者と相手の各歌声の振幅包絡線について，相互相関係数で類似性，グレンジャー因果性検定で予測性を評価した結果，人間同士でより強い予期的同期と振幅包絡線の類似性が見られ，視覚情報により，予期的同期がさらに強まった．これらから，人間の歌声の音響的特性と微細な身体運動が，歌唱行動に対する予測と同期を支える知覚情報を提供していることが示唆された．

キーワード：inter-personal coordination, anticipatory synchronization, visual information, self-other integration

1. はじめに

音楽演奏では，他者の行動や音に対する予測を行うだけでなく，タイミング調整と予測制御を行う重要性も含まれている（Proksch et al., 2022; Keller et al., 2007; Konvalinka et al., 2010）．このことから，音楽での相互作用は，複雑な協調ダイナミクスを調べるうえで理想的なモデルである（Proksch et al., 2022; D'Ausilio et al., 2015）．より広く言えば，人間と機械あるいはロボットとの相互作用に関する研究を推進するための有望な枠組みを提供する．

情報科学技術の進歩により，人間らしい音響合成がより可能になってきた（Kühne et al., 2020; Robinson et al., 2022; Wang et al., 2024）．特に歌声合成技術によって，機械は人間の声に極めて近い歌声を生成できるようになった．先行研究では，音や声が人間と機械の相互作用において中心的な役割を果たすことが示されており，機械による演奏の音色，表現力，および協調能力の進展が報告されている（Sunardi & Perkowski, 2017; Robinson et al., 2022; Wang et al., 2024）．しかし，人間と機械による協調的な音楽演奏の実現は未だ困難であり，その主な理由は，精微な運動協調と正確な時間的同期を必要とするためである（Wang et al., 2024）．さらに，人間との円滑で双方向的な協働を支えることができる人工パートナーの開発は，依然として大きな課題である（Dotov et al., 2024; Wang et al., 2024）．

人間の音楽演奏の協調では，身体の動きが重要な役割を果たし（Klein et al., 2022），演奏者の動きを視覚的に捉えることは，タイミングや協調に極めて有益な情報を提供する（Bishop et al., 2019）．また，身体の動きは，他者の行動の予測を支える（Keller et al., 2007; Burger et al., 2014）．しかし，相手が人間ではなく人工物である場合，視覚情報が同等の影響を及ぼすかどうかは未解決の問題である．

我々の研究（Nishiyama and Nonaka, 2024）では，視覚情報がなくても，人工パートナーと歌う場合と比較して，人間パートナーと歌う場合の方が，より強い予期的同期が生じ，歌声の振幅包絡線の展開ダイナミクスにおける類似性が高いことが示された．この知見に基づき，本研究では参加者が相手を視認できる条件を加えた．この拡張により，視覚情報の有無にかかわらず，人間パートナーとの歌唱が引き続きより強い先行同期をもたらすのか，あるいは視覚情報の追加によって，人工パートナーとの協調が人間同士の歌唱で観察される協調ダイナミクスに近づくのかを検証する．

これらの問題に取り組むため，VOCALOID 6（ヤマハ株式会社，浜松市）の VOCALOID AI を用いて歌声を生成し，さらに「Hedra Character 3」（Hedra Studio; https://www.hedra.com）を用いて，合成音声と顔や上半身の動きが同期した人工パートナーを生成した．

2. 方法

女子大学生 13 名（平均年齢 25.2 歳，標準偏差＝5.8）が参加し，人間パートナーの歌声を録音するために女性歌手が起用された．

実験では,『きよしこの夜』(70 BPM, 3/4 拍子)の一部 (23 小節, 約 55 秒) を使用した. 相手の歌声として, 人間と, VOCALOID AI (VOCALOID 6:「HARUKA」) の 2 種類を用意した. 各録音には, テンポ提示のために歌唱開始前に2小節分のメトロノーム音が含まれた. そのため, VOCALOID AI は一定のテンポを維持したが, 人間の歌手は最初のメトロノームを聴いた後, 自らのテンポで歌唱した. ただ, 人間パートナーの歌唱テンポについて, 各小節の第 1 拍オンセット間隔から算出したところ, 平均 70.05±2.59 BPM で, 人工パートナーのテンポとの大きな差は見られなかった (図 1 参照).

視覚刺激として, 人間パートナー条件では, 歌手の約 1.5 メートル前方からビデオカメラで撮影した. 人工パートナーの条件では,「Hedra Character 3」を用いて, AI で歌唱中の動画を生成した. 両条件で, パートナーは正面を向いており, 鎖骨から上部の動きが記録された.

参加者には事前に曲を通知し, 歌詞の暗記を指示した. 実験では, 参加者にイヤフォンから流れる相手の歌声に合わせてユニゾンで歌うことと, 2 小節分のクリック音を聴いてから歌うよう指示した. 視覚条件では相手の歌唱映像を提示し, 非視覚条件では提示しなかった. 各参加者はパートナー条件ごとに各視覚条件で 5 試行ずつ行い, 計 20 試行を実施した. パートナー条件の順序は無作為化し, 6 名は視覚条件から, 7 名は非視覚条件から開始した. 実験後には, パートナーの識別や, 歌いやすさに関する主観評価のアンケートを実施した. 回答は多様であったが, 一部の参加者は人工パートナーの動きや歌声が AI 生成あるいは人間であることに気づかなかったと報告した.

分析では, まず録音データを 16 kHz にリサンプリングし, 各音声信号の振幅包絡線は, ヒルベルト変換の振幅を計算して抽出した. 3 次バターワース IIR フィルタ (カットオフ周波数 32Hz) を適用し, 声質などに関わる高周波数成分を除去した. 得られた振幅包絡線の時系列データは, 200 Hz にダウンサンプリングした.

上記で得られた歌声の波形 (図 2 参照) を用いて, 参加者と相手の各歌声の, 振幅包絡線の時系列変化パターンの類似性について, 相互相関 (CC) 係数で評価した. 各試行について, 相手と, 参加者の各歌声における振幅包絡線の時系列変化間の CC 係数を, ±0.1 秒の範囲の時間遅れに対して算出した. CC 係数から, 各試行および参加者ごとに 3 つの指標が抽出された; (1) 最大 CC 係数は, すべての時間遅れにおける相関係数の最高値であり, 相手と参加者の各歌声の振幅包絡線ダイナミクスについて全体的な類似性の指標として用いた. (2) 最大相関時の時間差は, 最大 CC 係数に対応する時間遅延として算出された. 正の時間差は参加者の歌声が相手より遅れていることを示し, 負の時間差は参加者の歌声が相手より先行していることを示す. (3) ゼロラグ CC 係数は, ラグ 0 における CC 係数を評価することで算出され, 相手と参加者の各歌声の振幅包絡線間の時間的同期または位相整合の尺度を提示した.

さらに, 各参加者および各試行について, パートナーと参加者の各歌声の振幅包絡線の時系列変化間で, 双方向のグレンジャー因果性 (GC: Granger causality) 検定を行った. 解析は, Multivariate Granger Causality Toolbox (Barnett & Seth, 2014) を用いた. 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, 3 つの CC 指標について線形混合効果モデルでモデル化した. 各モデルにおいて, パートナーと視覚条件を固定効果とした. GC 値については, パートナー, 視覚条件, 方向を固定効果とした線形混合効果モデルを用いて分析した. すべてのモデルにおいて参加者をランダム効果とし, 事後比較はボンフェローニ補正を適用した. 有意な効果の α 値は 0.05 とした.

図 1　2 種類のパートナー (人間と VOCALOID AI) の歌唱テンポにおける変動

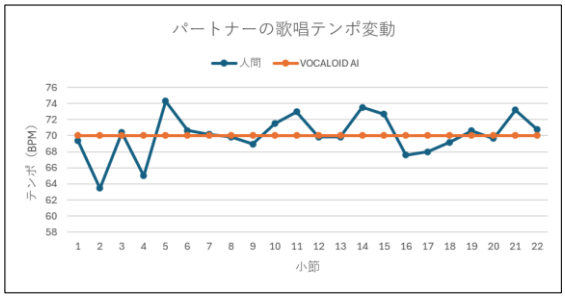
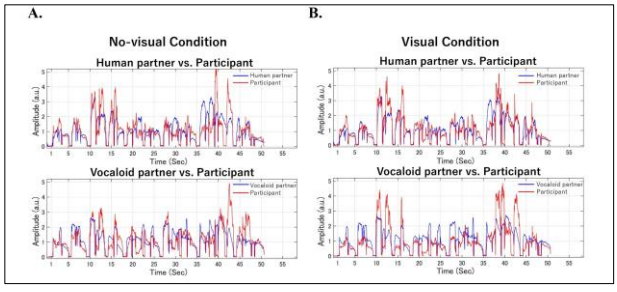


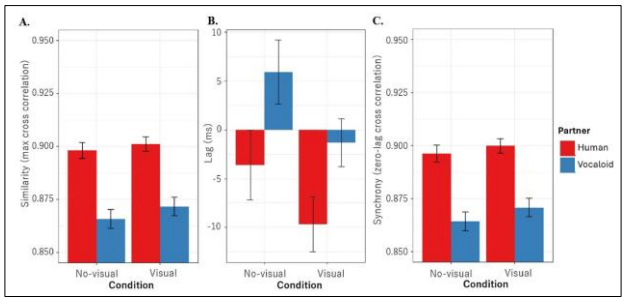
図 2　2 種類のパートナー (人間と VOCALOID AI) と参加者の各歌声の振幅包絡線における時間変化の例. A) 非視覚条件 (B) 視覚条件



3. 結果

参加者は，人間と歌う場合，振幅包絡線の時間的変調についてより高い類似性を示した（図 3A）（$F_{(1, 244)} = 114.18, p < .0001$）．視覚条件の有意な効果は得られなかった．人間が相手だと，参加者はパートナーより先に歌う傾向が見られ，この傾向は，視覚情報がある場合に強まった．対照的に，非視覚条件で人工パートナーと歌う場合，参加者はパートナーより遅れる傾向が見られ，視覚条件ではラグが小さくなったものの，人間パートナー条件に比べれば先行性は低かった（図 3B）．線形混合効果モデルの結果，パートナー（$F_{(1,244)}=32.14, p < .0001$）および視覚条件（$F_{(1,244)}=17.72, p < .0001$）の主効果が有意であり，視覚情報提示時には参加者がより先行して歌う傾向が示された．その他の効果は有意ではなかった．同期性についても，同様の傾向が示され（図 3C），参加者は人間パートナー条件でより強い相関を示した（$F_{(1, 244)} = 108.47, p < .0001$）．

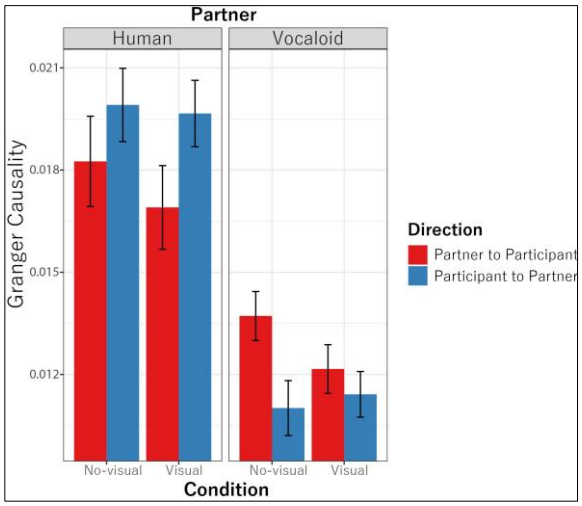
図 3　歌声の強弱における時系列変化パターン間の CC 分析の結果．（A）最大 CC 係数値（B）最大 CC 係数時のラグ（C）0 ラグでの CC 係数値



また，GC 検定の結果（図 4）から，人工パートナーと歌う場合は，参加者がパートナーに追従する傾向が見られた．対照的に，人間パートナーと歌う場合では，参加者はパートナーより強く先行していた（パートナーの主効果：$F_{(1, 500)} = 101.86, p < .0001$，パートナー×方向の交互作用：$F_{(1, 500)} = 9.02, p = .0028$）．事後比較の結果，両方向と両視覚条件において，人間パートナーの方がより GC 値が高く，特に参加者からパートナーへの方向でその差が顕著であった．非視覚条件では，人間パートナーはより高い GC 値を示し（パートナー→参加者：推定値=0.00454, p=.0006；参加者→パートナー：推定値=0.00890, $p < .0001$），視覚条件でも同様の傾向が認められた（パートナー→参加者：推定値=0.00474, $p = .0003$；参加者→パートナー：推定値=0.00824, $p < .0001$）．その他の効果は有意ではなかった．

また，参加者間で個人差は認められたものの，結果における主要な傾向は，参加者間で概ね一貫していた．

図 4　グレンジャー因果性検定の結果



4. 考察

本研究では，人が，人間のパートナーと，あるいは歌声合成ソフトウェアによって生成された人工のパートナーと一緒に歌う場合とで，歌唱の協調ダイナミクスがどう異なるのかを検証した．さらに，視覚情報の有無がこれらの協調パターンに対してどのような役割を果たすのかを調べた．その結果，参加者は，人間パートナー条件では，人工パートナー条件に比べ，振幅包絡線の時間変調において高い類似性・同期性・予測性を示した．特に，人間との歌唱では，参加者が相手より先行して歌う傾向がみられたが，人工パートナーとの歌唱では，特に非視覚条件で，参加者が遅れる傾向がみられた．GC 検定の結果も同様であり，人間との歌唱では，より強い予測的協調の傾向が示された．これらの結果は，歌唱での「共在」が，表面的な人間らしい音声特性だけで捉えられるものではないことを示唆している．

人間の歌声は，相互に連動する身体的プロセスによって生み出され，それらは歌声と動きの両方に微細な変動をもたらす．これらは，聴覚的（例：吸気）および視覚的（例：上半身の動き）な予測情報として機能する可能性がある．しかし，合成音声や AI 生成の動きでは，これらを再現することは困難である．したがって，人工パートナーは，基本的な同期を支える「十分」な情報を提供できる一方で，強力な予期的同期を支えるような微細な構造は依然として欠如している可能性がある．

VOCALOID AI との類似性と同期性の低下は，人間では依然として達成が困難な音響的特性に起因している可能性がある．たとえピッチの軌跡が密接に一致していたとしても，合成音声には人間の発声に固有の動的変動が欠けている場合があり（Tamura et al., 2015），これは振幅包絡線の構造に反映された可能性がある．

また，視覚情報の提示により協調はより強い予測的傾向へとシフトした．これは，他者の行動を予測する際に，身体の動きが聴覚情報を補完する情報を提供すると示した先行研究（Keller et al., 2007; Burger et al., 2014; Sabharwal et al., 2024）と一致する．しかし，視覚情報は相手の種類による協調の質的な差異を埋められなかった．これは，予測的協調には視覚情報に加え，人間の歌唱に含まれている聴覚・身体運動双方の微細な時間構造が重要であることを示唆している．

今後は，呼吸や姿勢動揺，歌唱中のテンポ変動と，歌声の協調ダイナミクスの関係も調べることで，予期的同期の要因やそれを支える情報を明らかにしていく．さらに，今回は，参加者が相手に適応する一方向的な協調を対象としたが，実際の共同歌唱場面を踏まえ，歌唱者同士が互いに応答しながら継続的に適応する双方向的な協調についても検証する必要がある．これらの検討は，人間同士の音楽的協調に近い相互作用を実現する人工エージェント設計への応用が期待される．

文献

Barnett, L., & Seth, A. K. (2014). The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *Journal of Neuroscience Methods, 223,* 50–68. https://doi.org/10.1016/j.jneumeth.2013.10.018

Bishop, L., Cancino-Chacón, C., & Goebl, W. (2019). Moving to communicate, moving to interact: Patterns of body motion in musical duo performance. *Music Perception, 37(1),* 1–25. https://doi.org/10.1525/mp.2019.37.1.1

Burger, B., Thompson, M. R., Luck, G., Saarikallio, S. H., & Toiviainen, P. (2014). Hunting for the beat in the body: On period and phase locking in music-induced movement. *Frontiers in Human Neuroscience, 8*, 903. https://doi.org/10.3389/fnhum.2014.00903

D'Ausilio, A., Novembre, G., Fadiga, L., & Keller, P. E. (2015). What can music tell us about social interaction? *Trends in Cognitive Sciences, 19,* 111–114. https://doi.org/10.1016/j.tics.2015.01.005

Dotov, D., Camarena, D., Harris, Z., Spyra, J., Gagliano, P., & Trainor, L. (2024). If Turing played piano with an artificial partner. *arXiv preprint*. https://doi.org/10.48550/arXiv.2402.08690

Keller, P. E., Knoblich, G., & Repp, B. H. (2007). Pianists duet better when they play with themselves. Consciousness and Cognition, 16, 102–111. https://doi.org/10.1016/j.concog.2005.12.004

Klein, L., Wood, E. A., Bosnyak, D., & Trainor, L. J. (2022). Follow the sound of my violin. *Frontiers in Human Neuroscience, 16,* 982177. https://doi.org/10.3389/fnhum.2022.982177

Konvalinka, I., Vuust, P., Roepstorff, A., & Frith, C. D. (2010). Follow you, follow me. *Quarterly Journal of Experimental Psychology, 63,* 2220–2230. https://doi.org/10.1080/17470218.2010.497843

Kühne, K., Fischer, M. H., & Zhou, Y. (2020). The human takes it all. *Frontiers in Neurorobotics, 14,* 593732. https://doi.org/10.3389/fnbot.2020.593732

Nishiyama, R., & Nonaka, T. (2024). Can a human sing with an unseen artificial partner? *Frontiers in Robotics and AI, 11*, 1463477. https://doi.org/10.3389/frobt.2024.1463477

Proksch, S., Reeves, M., Spivey, M., & Balasubramaniam, R. (2022). Coordination dynamics. *Scientific Reports, 12,* 421. https://doi.org/10.1038/s41598-021-04463-6

Robinson, F., Bown, O., & Velonaki, M. (2022). Designing sound for social robots. *International Journal of Social Robotics.* https://doi.org/10.1007/s12369-022-00891-0

Sabharwal, S. R., et al. (2024). Leadership dynamics in musical groups. *PLOS ONE, 19(4)*, e0300663. https://doi.org/10.1371/journal.pone.0300663

Sunardi, M., & Perkowski, M. (2017). Music to motion. International *Journal of Social Robotics, 10(1),* 43–63. https://doi.org/10.1007/s12369-017-0432-9

Tamura, Y., Kuriki, S., & Nakano, T. (2015). Left insula and human voice. *Scientific Reports, 5*, 8799. https://doi.org/10.1038/srep08799

Wang, H., Zhang, X., & Iida, F. (2024). Human–robot cooperative piano playing. *IEEE Transactions on Robotics, 40*, 4650–4669. https://doi.org/10.1109/TRO.2024.3484633

謝辞

本研究は，JST 次世代 AI 卓越博士人材育成プロジェクト JPMJBS2410 の支援および JST ムーンショット研究開発プログラム JPMJMS263O の支援を受けたものです．