旋律の歌唱可能性の定量化に向けた歌唱の正確さの一分析

川原 未波[1,a)]　北原 鉄朗[1,b)]

概要：
本研究は, AI 技術を用いた歌唱旋律自動生成において, 人間が「歌いやすい」旋律を生成する基盤構築を目的とする. 本稿では発声の音高における歌唱の正確さに影響を与える要素を解析していく. 実験では, 異なる音の跳躍や BPM(Beats Per Minute), 歌詞の組み合わせによる 18 種類の旋律を用い, 100 名の参加者に歌唱実験を行なった. そして, その録音データを pYIN アルゴリズムで分析し, 二乗平均平方根誤差（Root Mean Square Error, RMSE）を用いて音高の一致率を評価した. 結果として, 跳躍パターンが RMSE に影響を与える一方で, BPM の変化による影響は限定的であることが示された. ただし, 慣れが精度に関与している可能性がある. 今後は特に影響が大きい音の跳躍パターンを中心に見つつ, 歌詞やリズムが歌唱の正確さに及ぼす影響も含めたさらなる分析が必要である.

1. はじめに

昨今, AI 技術を用いた音楽生成の分野において, 歌唱旋律の自動生成は大きな関心を集めている. 近年の音楽制作やエンターテインメントの分野では, AI を活用して高品質な楽曲や歌声を生成する技術が発展しつつあり, その成果はバーチャルシンガーや AI ボーカロイドとして広く普及し始めている. しかし, これらの歌唱旋律自動生成技術は, 人間の歌い方を出発点として設計されることが肝心であるにもかかわらず, 従来存在している膨大な楽曲データを学習データとしているのみに留まっている都合上, 生成された歌唱が「人にとって歌いやすいもの」であるかどうかについては未だ明確ではない.

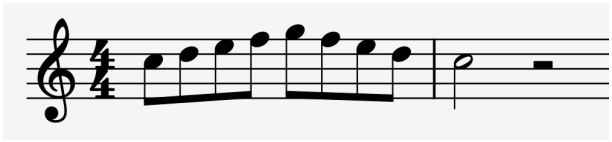
「歌いやすさ」という要素は, 単に楽曲の旋律が美しいかどうかだけではなく, 人間の音域や発声特性, 歌唱時のリズム処理など, 身体的・心理的な側面とも密接に関連する. しかし, 現存する自動生成技術ではこの歌いやすさの観点が十分に考慮されておらず, 生成された旋律や歌唱データは必ずしも人間が自然に, あるいは快適に歌えるものではない可能性がある.

本研究の目的は, 被験者による歌唱行為を実際に分析し, そのデータを基に人が歌うことができるかどうかの「歌唱可能性」を定義する試みを行うことである. さらに, この結果を反映させた歌唱旋律の自動生成ツールを構築し, 人間がより自然に歌える旋律の生成を目指す. 本研究が実現すれば, 生成される歌唱旋律が単なる機械学習の成果に留まらず, 人間の歌唱の身体的・心理的特徴に基づいた, より実用的で親しみやすい音楽表現へと発展することが期待される.

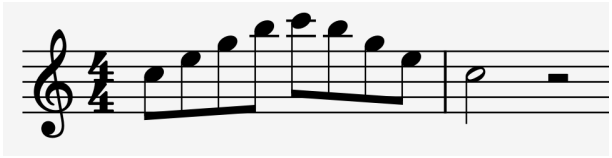
関連研究として, 歌いやすさに着目した楽曲検索システムの先行研究 [1] が挙げられる. この研究では, 主観評価を基に楽曲の難易度を算出しているが, 実際の歌唱行為における正確さや生理的な側面には十分踏み込んでいない. また, 調音運動に基づく発声難易度の指標化や歌詞の歌いやすさ評価に関する研究 [2] も存在するが, これらは主に歌詞のテキスト情報を対象としており, 旋律や歌唱行為そのものに関する検討は限定的である. 歌いやすさ (Singability) について触れている研究も多く存在するが, 翻訳した際の歌いやすさについて考慮したものが大半である [3][4]. ただ実践や指導を行うことにより歌唱性を高めるといった趣旨の内容 [5] や, 既存曲の歌いやすさを検証した研究 [6][7] は存在するが, これらは歌唱経験が少ない人などを対象とした分析ではなかった. また, 楽曲の難易度を演奏速度やリズムの不規則性から考慮する研究 [8] も存在したが, 歌いやすさに着目したものではない. 歌唱生成 AI に歌いやすさを考慮させる研究 [9][10] も存在したが, 歌詞と曲の旋律の整合性をとり, 音の強弱や長短を揃えるに留まっていた.

本研究では, 実際に被験者に歌唱してもらう実験を通じて, 歌唱の正確さを客観的に計測し, 正確さに影響を与える要素を明らかにする. その第一弾として, 本稿では発声の

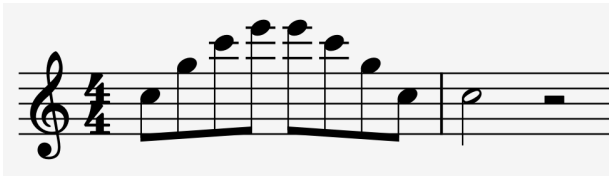
[1] 日本大学文理学部情報科学科
[a)] minami@kthrlab.jp
[b)] kitahara@kthrlab.jp



(フレーズ 1) 順次進行



(フレーズ 2) 跳躍 2, 3 度の進行



(フレーズ 3) 跳躍 3, 4, 5 度の進行

図 1 実験に際して使用した歌唱フレーズ

音高における歌唱の正確さを解析する．具体的には, 順次進行や跳躍進行などの異なる旋律パターン, BPM, 歌詞の組み合わせに基づく 18 種類の旋律を用いた歌唱実験を行い, その歌唱データを基本周波数推定を行う．そして, その推定値から歌唱の音高の正確さを分析する．

2. 実験

2.1 使用データ

本実験において, 参加者には本研究で準備した指定の旋律に沿って歌唱してもらう形式とした．実験条件は以下の通りである．

- 順次進行, 跳躍 2, 3 度進行, 跳躍 3, 4, 5 度進行
- 歌詞 A(*ダダダダダダダダー*),
  歌詞 B((ダビダバダビダバダー)
- BPM80, 120, 150

上記の跳躍進行別 3 種, 歌詞別 2 種, BPM 別 3 種を全て組み合わせた計 18 旋律をそれぞれ歌唱してもらう．用意した旋律における跳躍別の例について, 図 1 にて記す．

また, 18 種類ある旋律を音源 1-1-1〜音源 3-2-3 で分類する．分類の仕方としては, 先頭の数字が順に順次進行, 跳躍 2, 3 度, 跳躍 3, 4, 5 度を, 真ん中の数字が歌詞 A, B を, 最後の数字が BPM 別 80, 120, 150 を表す．

なお, 以降において, 実験参加者が収録した音声データを「録音データ」, 本研究で用意した歌唱旋律を「正解データ」と呼称する．

正解データでは, 歌声合成ソフト「Synthesizer V」にて機械音声の重音テトの歌声を利用して, 図 1 の通りに歌唱したものを使用する．

2.2 実験参加者

本研究の実験には, クラウドソーシングサイト「ランサーズ」を通じて募集した 100 名が実験に参加した．

2.3 実験手順

2.3.1 実験フォーム

本研究の実験では, 歌唱旋律の再生および録音機能を備えたフォームを Google Apps Script を用いて作成した．実験参加者は, 当該フォームの録音機能を利用して音声を順番に収録した．

2.3.2 実験用旋律

実際に歌ってもらう音声では, まず最初にお手本として機械音声で旋律を歌唱する．次に, 音高を掴むために音楽制作ソフトウェア「GarageBand」にて電子ピアノで旋律を追加した．実験参加者には, この電子ピアノが流れているタイミングで機械音声のお手本に沿って歌唱してもらう．

また, 1 つの旋律につき, 機械音声と電子ピアノを 3 回交互に繰り返す．そのため, 実験参加者には同じフレーズを 3 回歌唱してもらう必要がある．旋律録音の回数としては, 余程の録音ミスではない限り一度きりの録音で行っている．

収録ではお手本等はイヤホン・ヘッドフォンの着用を必須にしているため, 基本的には実験参加者の歌声しか入っていないことを想定している．

2.4 分析手法

本研究では, pYIN 基本周波数推定アルゴリズムを用いて, 楽曲全体の基本周波数を分析する．

歌唱の正確さを分析するにあたり, 「歌唱データ全体の音高の一致率」の観点から数値的評価を行う．

2.4.1 相互相関による録音データと正解データの時間同期

歌唱データ全体の音高の一致率および一音ごとの音高の一致率を算出するには, 録音データと正解データの時間軸を正確に一致させる必要がある．このため, 拡張モジュール「numpy」の correlate 関数を利用し, 正解データを基準として録音データの時間軸を調整する処理を実施した．

2.4.2 ノイズ調整

本実験では, 参加者が各自の環境で録音を行う形式を採用しているため, 録音中にノイズが混入する可能性が考えられる．この問題に対処するため, 録音データの振幅値を最大値が 1.0 になるように正規化し, 振幅が 0.01 以下のフレームは記録対象から除外するフィルタリング処理を実施した．

2.4.3 音高調整

録音データと正解データを比較する際, 参加者ごとの音域の違いにより正解データとの音高の差が生じ, 解析結果に大きな変動が見られる場合がある．しかし, この変動は単に音高が取れていない可能性も含まれるため, より正確な解析を行うための補正手法を導入した．具体的には, 正

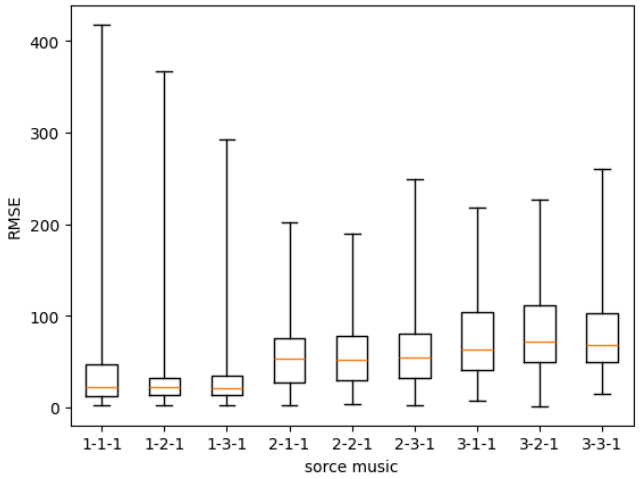


図 2 旋律ごとの録音データと正解データの RMSE の分布
(x-y-z;x = 音の跳躍, y = BPM, z = 歌詞)

解データの音高を 4 倍, 2 倍, 等倍, 0.5 倍, 0.25 倍に変換したデータを参照し, それぞれの誤差を算出した上で, 最も誤差が小さいデータを解析対象として採用する方法を用いた.

2.4.4 歌唱データ全体の音高の一致率

楽曲全体の基本周波数に着目し, その一致率を評価するための手法として, 得られた基本周波数の推定値を基に正解データとの誤差を RMSE を用いて評価する. 具体的には, 以下の式で誤差を算出する.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - p_i)^2}{n}}$$

ここにおける変数は下記の通りである.

$\boldsymbol{x_i}$ フレーム数 i における録音データの基本周波数推定値
$\boldsymbol{p_i}$ フレーム数 i における正解データの基本周波数推定値
$\boldsymbol{n}$ 歌唱データ全体の総フレーム数

RMSE が小さいほど, 推定結果が正解データに近く, 一致率が高いことを示す.

3. 実験結果・考察

3.1 歌唱データ全体の音高の一致率

今回は実験参加者 100 名のうち, pYIN 基本周波数推定アルゴリズムにて有効な数値が取得できなかった 9 人分を除いた 91 名分のデータを提示する.

また今回の解析では歌詞における歌唱の正確さの変化に関しては省略することとし, 歌詞 A の歌唱のみを分析対象とする.

図 2 にて録音データと正解データの RMSE を旋律ごとに分布をとったものを箱ひげ図にて掲示する. この図は, 横軸が歌唱による変化の要素を除いた跳躍 3 種と BPM3 種の計 9 種類の旋律であり, 縦軸は RMSE の大きさを示す.

3.2 考察

グラフから, 順次進行である音源群と跳躍が 3 度以上である音源群を比べると, RMSE の中央値および四分位範囲が変化していることが確認できる. この結果は, 音の跳躍が歌唱の正確さにおける分析結果に影響を与えている可能性を示唆している. ただ, 跳躍が 2, 3 度の音源群と跳躍が 3, 4, 5 度の音源群の分布を比較してみると, 順次進行の音源群ほどの違いは無いようにも見える.

一方, BPM の変化に伴う RMSE の変動は図から顕著ではないことが示唆される. 例えば, 音源 3-2-1 では BPM の変化により RMSE の中央値および四分位範囲が上昇しているものの, 続く音源 3-3-1 ではこれらの値が低下している. このことから, BPM80 から 150 の範囲におけるテンポの遷移が歌唱の正確さに大きな影響を与えるとは言い難い.

さらに, 音源 1-1-1 の最大値や四分位範囲が 1-2-1, 1-3-1 より大きくなっている. この傾向は, 音源 1-1-1 が実験における最初の旋律であり, 実験参加者がまだ歌唱に慣れていない段階で収録されたことによる影響であると考えられる. 特に, 音源 1-1-1, 1-2-1, 1-3-1 は最大値が特筆して大きいが, 中央値などの分布としてはかなり低い数値でまとまっているため, 一部の実験参加者で慣れない人の誤差が高いことが要因である可能性が高い.

4. まとめ

本実験では, 歌唱の正確さを分析することを目的として, 実験参加者に歌唱実験を実施した. そして参加者の歌唱データから基本周波数を推定し, 音高に着目した解析を行った. 今回の解析で, 音の跳躍が歌唱の正確さに影響を及ぼしている可能性が高く, BPM に関しては 80 から 150 の区間においてはそこまで高い影響を及ぼしていない可能性があることが分かった. また, 歌唱の慣れによっても正確さが変わる可能性があることも確認できた.

今後の研究においては, 本実験で特に変化量が大きかった音の跳躍に注目して解析を進めていく. また, 本研究では測れていなかった歌詞の発音や, 単調でない複雑なリズムを歌唱した際に変化する歌唱の正確さについても解析を行なっていきたい.

謝辞

本研究は, JSPS 科研費 23K24966, 24H00748 の助成を受けた.

参考文献

[1] 山本 雄也, 平賀 譲, 歌いやすさ・歌いにくさに着目した楽曲検索システムのためのポピュラー楽曲の歌唱難易度算出の検討, 情報処理学会 研究報告音楽情報科学（MUS）, vol.2019-MUS-124, No.9, pp.1–6, 2019.
[2] 宋 健智, 齋藤 大輔, 峯松 信明, 調音運動に基づく発声難易度の指標化と歌詞の歌いやすさ評価への応用の検討, 情報処理学会 研究報告音声言語情報処理（SLP）, vol.2018-SLP-122, No.40, pp.1–6, 2018.
[3] Longshen Ou, Xichu Ma, Min-Yen Kan, Ye Wang, Singable and Controllable Neural Lyric Translation,

arXiv:2305.16816, 2023
[4] Haven Kim, Kento Watanabe, Masataka Goto, Juhan Nam, A Computational Evaluation Framework for Singable Lyric Translation, arXiv:2308.13715, 2024
[5] Saijun Chen, A Brief Analysis of Singability in Erhu Fiddle Performance and Teaching, Curriculum and Teaching Methodology, vol.5, Issue10, 2022.
[6] Michael D. Barone, Karim M. Ibrahim, Chitralekha Gupta, Ye Wang, Empirically Weighting the Importance of Decision Factors for Singing Preference, ISMIR, pp.529-536, 2018.
[7] 青野裕司, 片寄晴弘, 井口征士, 世代別歌の歌い易さ評価モデルと音楽コンテンツ制作への応用, インタラクション９９予稿集, pp.67-68, 1999.
[8] Véronique Sébastien, Henri Ralambondrainy, Olivier Sébastien, Noël Conruyt, Automatically Determining Scores Difficulty Level for Instrumental e-Learning, ISMIR, pp.571-576, 2012.
[9] Qihao Liang, Xichu Ma, Finale Doshi-Velez, Brian Lim, Ye Wang, Improving the Singability of AI-Generated Lyrics with Prosody Explanations, IJCAI, pp.7877-7885, 2024.
[10] Longshen Ou, Xichu Ma, and Ye Wang, Joint Learning of Wording and Formatting for Singable Melody-to-Lyric Generation, arXiv:2307.02146, 2024.