Vol.2025-NL-266 No.23
Vol.2025-SLP-158 No.23
2025/12/17

日本語ラップスタイル合成歌唱を対象としたスウィングの統計分析

山本 泰我[1,a)]　森勢 将雅[1,b)]

概要：歌唱スタイルの多様性は，その登場から著しく発展を遂げている歌声合成において，未だ人間による多彩な表現の実現が困難な要素の一つである．そこで本研究では，ラップスタイルに焦点を絞り，ラップ特有の音響的要素であるスウィングに関する分析と再現を試みた．実音声と合成音声間において，表拍と裏拍の持続時間の比率（以降，スウィング比率とする）に有意な差が得られた．主観評価実験の結果，スウィング比率の調整によって相対的にラップらしさが有意に変化することが示された．また，アンケート調査により，スウィング比率と歌声の自然性とがトレードオフの関係にあることが示唆された．

1. はじめに

機械学習の発展とともに歌声合成技術は著しい進化を遂げており，その高い音質から，アマチュアによる音楽創作からプロフェッショナルな音楽制作に至るまで幅広く活用されている．品質が向上する一方で，歌唱スタイルの多様性においては依然として人間の歌唱に及ばない点も多い．人間の場合，練習や意識によって多様な歌唱スタイル及びそのスタイル固有の表現技法を再現可能である．歌声合成の場合，歌唱スタイル固有の表現技法の制御を目的として，歌唱スタイルごとに細分化したアプローチが提案される [1][2]．これはラップスタイルにおいても同様であり [3]，故にラップスタイル含め特定の歌唱スタイルに特化した歌声合成研究は，歌声合成全般を対象とした研究と比較すると希少である．

ラップに関する既存の研究としては，ラップバトルにおけるバース（ラップバトルにおける返答フレーズ）生成を対象とした自然言語処理的アプローチ [4] や，Text-to-Speech (以降，TTS) 技術を応用した対話型ロボットによるラップバトルの実現 [5] などが挙げられる．音響的アプローチの先行研究としては，ラップスタイルに特化した記譜法の定義と隠れマルコフモデル（Hidden Markov Model; HMM）ベースの歌声合成を行った事例 [6] が存在する．これは，深層学習ベースの手法が主流となった 2025 年現在においては改善の余地がある．また，大規模ラップ歌唱データセットの作成から行い，入力伴奏に合わせたリズムのラップ歌唱の生成を可能にした研究 [7] も存在するが，スウィングやフォールといったラップ特有の歌唱技法の制御は考慮されておらず，より自由度の高いインターフェースへと設計を改善する余地がある．

本研究では，ラップスタイル歌唱における特徴的な要素であるスウィングに関する統計分析を試みた．具体的には，実音声及び合成音声におけるスウィング比率に関する比較を行い，その後，スウィングとラップらしさの相関を明らかにすることを目的に主観評価実験を行った．その結果，スウィング比率の調整によって相対的にラップらしさが有意に変化することが示された．また，アンケート調査により，スウィング比率と歌声の自然性とがトレードオフの関係にあることが示唆された．

2. 関連研究

2.1 歌声合成

歌声合成（Singing Voice Synthesis; SVS）とは，楽譜及び歌詞を入力し，それに対応した歌声を生成する技術である．2003 年に発表され [8]，2004 年にヤマハ株式会社より発売された VOCALOID [9] は歌声合成を文化的に広めた象徴的存在である [10]．当製品は単位選択型の合成方式が採用された楽音合成に近い技術コンセプトであり，専用の歌声コーパスに含まれる音声の素片サンプルを直接使用し逐次的に合成することが特徴である [11]．この技術の発展として，ブレッシー（breathy）やボーカルフライ（vocal fly）といった非モーダルな歌声の合成を可能にした手法も提案されており，単位選択型の合成方式を利用した高品質な合成が実現されている [12]．

単位選択型とは異なり，統計モデルを活用するのが統計

[1] 明治大学
[a)] cs242038@meiji.ac.jp
[b)] mmorise@meiji.ac.jp

的パラメトリック手法である．当初，統計的パラメトリック歌声合成で採用されていた HMM [13][14] は，自然性向上や柔軟性向上などの観点からディープニューラルネットワーク（Deep Neural Network; DNN）へと置き換わった [15][16] が，いずれも「歌詞と楽譜情報から歌声を一括で生成する」という点は一貫している．この手法の利点としては，歌声モデルとして扱うことでダイナミクスや奏法の指示といった柔軟な合成が可能になることや，波形を直接管理しないためモデルサイズが小さくなることが挙げられる．一方で，生成時に使用するボコーダ起因の品質面での制約があるという欠点も存在する．

この問題に対処するため提案された手法が，WaveNet [17] や HiFi-GAN [18] といった波形を直接生成する技術を利用したニューラルボコーダ及び End-to-End 方式の歌声合成である [19][20]．これら手法は，従来の音響特徴量に基づく波形合成を行うボコーダとは異なり，音響モデルによって生成されたメルスペクトログラム（もしくは楽譜やテキスト情報）から歌声波形を生成可能である．主観評価実験では，自然音声とほぼ同等の品質であることが報告されている [21]．ただし，End-to-End 方式で扱う潜在空間は高次元であるため，ユーザの細かな表現の意図をモデル内部の音響パラメータ及び出力生成波形に直感的に反映させることは困難である．したがって，表現の制御性や扱いやすさにおいては統計的パラメトリックな手法が優れているとされている [22]．

2.2 歌唱技法を制御可能な歌声合成

歌声合成分野は品質向上という観点においては実音声に匹敵する水準を達成している [23]．その後は歌声表現技法の制御 [24] や通常歌唱とは異なる歌唱スタイルの再現 [7]，ゼロショット学習を利用した異なる話者性の適用 [25] などが試みられている．このような歌声の表現性に関する要素は，未だ実音声と比較して合成音声が向上の余地のある部分であると同時に，歌声が聴き手に与える印象の重要な要素である [26][27]．

Sinsy [28] は最も基本的な歌声表現技法の一つであるビブラートの付与を可能にした手法である．当研究では基本周波数（以降 f_0）とは別にビブラートを DNN でモデル化することで，HMM ベースのシステムに比べ高い主観評価結果が得られたことが報告されている．SinTechSVS [24] はより広範囲な歌唱技法の制御に焦点を当てた手法である．当研究では，ピッチや音素，持続時間といった基本的な音響特徴量に加え，ピッチや音色に関する歌唱技法ラベルを同時に学習することで，Diffusion ベースの音響モデルにおける歌唱技法の制御性向上が実現されている．また，入力楽譜から歌唱技法列を生成する推薦システムも提案されており，歌唱技法ラベル作成の簡素化を狙っていることも特徴である．

2.3 ラップスタイル歌声合成

ラップスタイル歌声合成は歌声合成（SVS）分野に属する研究区分の一つであり，これまでも様々なアプローチで研究が行われてきた．HMM ベースの歌声合成が主流であった 2010 年代の研究 [6] では，グリッサンドやアクセントといったラップスタイル歌唱特有の歌唱技法に対応した記譜法が提案されている．これら技法は HMM によってモデル化されており，f_0 軌跡の視覚上では再現が確認されているものの，その楽譜を基に HMM 歌声合成で得られた歌声を使用した主観評価実験ではその有効性が示されなかった．

明確なメロディを持たない場合もあるというラップの特性から，SVS ではなく TTS でのアプローチが行われた研究 [5] も存在する．当研究は人間のラッパーとラップバトルが可能なロボットを開発した点が革新的だが，その最終的な出力に相当する音声合成部は Google が提供する TTS がベースである．同じくラップバトルを想定したものとして，バースの生成に焦点を当てた研究 [4] があるが，これはテキストベースの研究であり歌声の合成には至っていない．

2025 年に発表された Freestyler [7] は，条件付きフローマッチング [29] を利用した音楽性と品質の両立が実現しているラップスタイル歌声合成研究である．Transformer の言語モデルでの意味トークンの予測の際に，歌詞に加えて伴奏情報を入力することで伴奏のリズムに対応した歌声の合成が実現されている．ゼロショットでの話者性の付与も可能であり，主観評価実験では実音声に匹敵する品質であることが報告されている．一方で，スウィングやフォール，アクセントといったラップスタイル特有の表現技法の制御性については言及されていない．また，入力情報は歌詞と伴奏に限られており，楽譜を入力とするような意図した音高のラップフレーズの生成には対応していない．

SynthesizerV [30] は，ラップスタイルの歌声の生成が可能な歌声合成ソフトウェア製品の一つである．この製品は，深層学習に基づく音響モデルにより実音声に匹敵する自然な歌声が生成可能であり，2023 年にはラップスタイル歌唱の生成にも対応した [31]．しかしながら，ラップスタイルとは多岐に渡るものであり [32][33]，本製品においても何を以てラップらしい歌声とするかはユーザに任せられている．

2.4 本研究の位置付け

2.3 節にあるように，ラップスタイルの歌声合成に関連した研究はいくつか存在する．しかしながら，ラップらしさに影響を与える要素について分析を行った事例は少ない．本研究では，DNN をベースとした歌声合成フレームワークである NNSVS（Neural network based singing voice synthesis library）[34] を一部改良し，任意のスウィング比率に設定した音源を生成した．それら音源と実音声を比較し，NNSVS による合成歌唱におけるスウィングの再現性

を検証した．また，スウィングとラップらしさとの相関調査を通し，スウィングのラップスタイル歌声合成パラメータとしての妥当性の考察を行った．

3. 実験準備

3.1 ラップスタイル歌唱データセット

本研究では，クリプトン・フューチャー・メディア株式会社より提供されたラップスタイル歌唱データベースを使用した．当該データベースは，ラップ歌唱に精通した日本語話者の女性 1 名により収録された 231 のラップフレーズ（48 kHz/24 bit）及び，各フレーズに対応するモノフォンラベルから構成されている．これら音源とラベルを基に，Harvest [35] を用いてピッチ推定された f_0 系列から XML 形式の楽譜を作成した．なお，楽譜は BPM・歌詞・音符情報のみを含み，それ以外にスウィングやフォールといった表現記号は含まれていない．以上 3 点（音源・ラベル・楽譜）を統合しラップデータセットを構築した．

3.2 ラップスタイル歌唱モデル作成

ラップスタイル合成音声の学習・生成には，既存の歌声合成フレームワークである NNSVS [34] を使用した．NNSVS の学習部は順にタイムラグモデル，持続時間モデル，音響モデルの 3 つのモジュールから成り，各モジュール詳細は付録 A·1〜 A·4 に記載の通りである（記載外の学習パラメータは NNSVS のデフォルト設定に準拠）．生成時には WORLD [34] ボコーダを使用した．

3.1 節に述べたデータセットのうち，24 フレーズを検証用，3 フレーズをテスト用，204 フレーズを学習用とした．

4. スウィング比較実験

スウィング感は，演奏されるリズムのタイミング変動に起因するものである．特に標準的な 2 分割形式の記譜（4 分音符・8 分音符・16 分音符等）と三連符的分割の対比がその代表例であり [37]，ラップ音楽においても用いられる（図 1）．本研究では，合成音声のスウィングの再現度を検証するため，実音声と合成音声間のスウィング比率の比較を行った．

歌声の生成には 3.2 節で作成した歌唱モデルを使用し，3.1 節で作成した歌唱データセット内の楽譜を入力として得られた音声を実験に使用した．すなわち，歌唱データセットの 231 の歌詞フレーズに対し，実音声と合成音声がそれぞれ存在する．

4.1 スウィング定義・算出

本研究では，表拍と裏拍で対を成すペアの 8 分音符を対象とし，その持続時間の比を「スウィング比率」と定義した．音素ごとの持続時間を Julius [38] による自動音素ラベリングによって取得し，それを基に各拍の長さを算出した．すべてのフレーズは 12 の 8 分音符を含むため，1 フレーズあたり 6 つのスウィング比率が存在する（図 2）．

4.2 実験結果

実音声と合成音声の対応箇所におけるスウィング比率を比較するため，ウェルチの t 検定を実施した．算出した全てのスウィング比率のうち，表拍・裏拍のいずれかの持続時間が 16 分音符未満であるものは検定の対象外とした．12 あるスウィング比率の集合のうち，検定の対象が最も少なかった集合の標本数は 196 であり，いずれの集合も等分散性はなかった．検定の結果を図 3 に示す．横軸は各表・裏拍ペア，縦軸はスウィング比率を示し，エラーバーは 95%信頼区間を示す．検定の結果，すべてのペアの 8 分音符において統計的に有意な差（$p < 0.05$）が認められた．

一般に，学習データを推論時の入力とすると，モデル本来の性能に比べ楽観的な結果が得られることが知られており [39][40]，これは歌声合成においても同様である．本実験において，使用した音源の生成に学習セット内の楽譜が使用されている事実は，楽観的な結果が期待できることを示す．したがって，この結果は，NNSVS による合成音声が実音声に比べ，スウィングリズムを適切に再現できていない可能性を示唆している．

5. スウィングとラップらしさの相関調査

本研究では，スウィング比率がラップらしさに与える影響を評価するため，主観評価実験を実施した．本実験では，被験者は異なるスウィング比率で作成した複数の音源を聴取し，そのラップらしさを評価した．

5.1 実験概要

実験に使用する音源は，NNSVS の持続時間モデルの直後にスウィングポストフィルタモジュールを新たに追加し

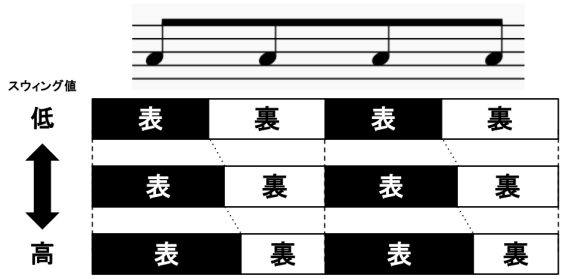

図 1: スウィングの概略図.

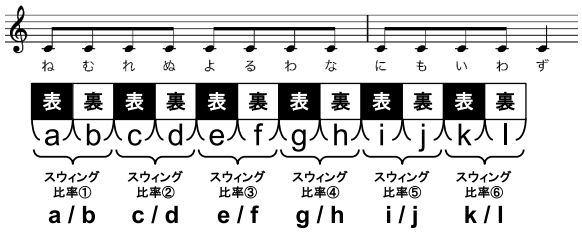

図 2: 各フレーズに存在する 6 つのスウィング比率例.

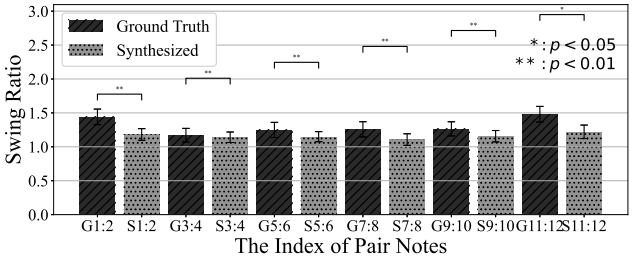

図 3: 各表・裏拍ペアにおけるスウィング比率.

表 1: 実験使用音源詳細

項目	詳細
音源数	**243 音源**
- フレーズ数：9 種類	
- スウィング比率：9 段階（1.0〜3.0, 0.25 間隔）	
- BPM：3 段階（160, 200, 240）	
サンプリングレート	96 kHz
量子化ビット数	16 bit
フレーズ長	2 小節
ファイル形式	WAV

表 2: 実験環境詳細

被験者	日本語話者男女 20 名
再生環境	騒音レベル 20-30 dB の防音室
ヘッドフォン	HD 660S (Sennheiser)
Audio I/O	ADI-2 Pro FS R (RME)
A/D 変換	96kHz/16bit

作成した．スウィングポストフィルタモジュールは，設定したスウィング比率に基づき表拍と裏拍の持続時間を調整することで，任意のスウィング比率に対応した音源を生成可能にするモジュールである．スウィング比率は 1.0（最小）から 3.0（最大）まで，0.25 刻みの 9 段階で設定した．実験に使用した音源の詳細については，表 1 に示す．なお，歌詞は極力韻を踏まないように作成し，入力楽譜のメロディの音高はすべて C3 に固定することで，メロディの影響を排除した．

被験者は日本語話者の男女 20 名で，各音源に対してラップらしさを 1 から 5 の 5 段階（Mean Opinion Score; MOS）で評価した．評価に際しては，メロディとリズムのみに着目するよう指示し，歌詞や韻に起因する影響が出ないよう配慮した．また，聴取開始前に，評価の軸となるリファレンス音源を提示することで，被験者間でラップらしさの共通認識を形成し，評価軸のずれを防ぐ工夫を行った．リファレンス音源は，歌唱データセットのうちテスト用に該当する 3 つの実音声である．実験環境の詳細は表 2 に示す．

5.2 実験結果

本実験の結果を図 4 に示す．横軸は音源生成時の指定スウィング比率，縦軸は評価値を示し，エラーバーは 95%信頼区間を示す．スウィング比率 1.0 から 1.75 の範囲において，各比率間（0.25 刻み）で統計的に有意な差が確認された．一方で，最も高い評価を得た音源はスウィング比率 2.25 のものであったが，スウィング比率 2.0 以降の比率間において有意な差は確認されなかった．

被験者を属性別に分類し，評価値の平均を比較した結果を図 5 に示す．属性グループは，2 年以上の音楽経験がある（N=11），音楽経験がない（N=7），日頃のラップ聴取習慣がある（N=6），ラップ聴取習慣がない（N=12）の 4 つである．各属性グループに該当する全ての回答に関してマンホイットニーの U 検定を行った結果は以下の通りである．

- 音楽経験の有無：何らかの音楽経験があるグループ（最短経験期間は 2 年）と比較して，音楽経験がないグループの平均 MOS 評価値は 0.63 ポイント高く，有意な差が確認された．
- ラップ聴取習慣の有無：日頃からラップを聴く習慣があるグループは，そうでないグループに比べ平均 MOS 評価値が 0.15 ポイント高く，有意な差が確認された．

これらの結果は，ラップらしさの知覚が被験者の音楽的な背景や日常的な聴取習慣によって影響を受けることを示唆している。

5.3 事後アンケートによる定性的意見

実験後に実施したアンケート結果に基づき，定性的分析を行った．その結果，特に「語尾がフォールしている（下降している）音源がラップらしく感じられた」という意見が多く確認された．これは，スウィング比率以外の音響的特徴もラップらしさの認知に寄与している可能性を示唆する．その他の具体的な意見については，表 3 にまとめて示す．

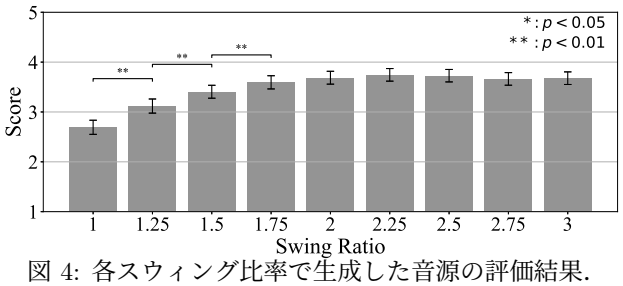

図 4: 各スウィング比率で生成した音源の評価結果.

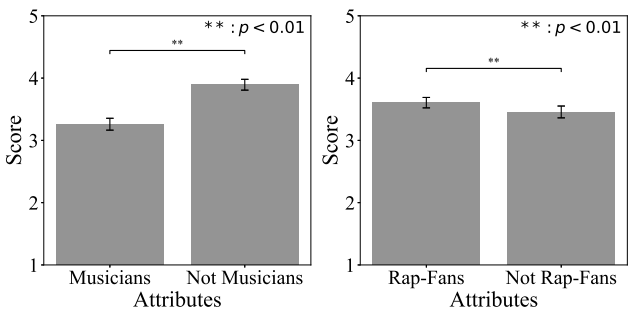

図 5: 属性別結果（左：音楽経験，右：ラップ聴取習慣）.

表 3: 主観評価実験で得られた定性的意見.

評価項目	ラップらしさを感じる要因	ラップらしくなさを感じる要因
速度・リズム	早すぎない，跳ねるリズム	遅すぎる/早すぎる
音高・抑揚	抑揚がある，アクセントがついている	抑揚がない，単調なメロディ
発音・明瞭度	流暢な発音，語尾が伸びている	発話内容が聞き取れない，声にノイズがある
歌詞・音韻	語尾が「よ」で終わる，裏拍が「ん」の歌詞	（記述なし）

6. 考察

本章では，前章で示した主観評価実験の結果に基づき，スウィング比率がラップらしさに与える影響，及び被験者の属性が評価に与える影響について考察する．

6.1 スウィング比率とラップらしさの関連性

スウィング比率 1.0〜1.75 間において有意な差が確認できたことから，スウィング比率が 1.75 以上の音源は 1.75 未満の音源に比べ，相対的に「よりラップらしい」と認識される傾向にあると考えられる．一方で，スウィング比率 1.75 以降では評価値に有意な差が見られなかった．この結果と，実験後のアンケート調査で「スウィング比率が大きくなると，裏拍のモーラが極端に短くなり，自然性が損なわれる」という意見が多く寄せられた点を総合的に考慮すると，スウィング比率の増加と音源の自然性の保持との間にトレードオフの関係が存在する可能性が示唆される．すなわち，過度にスウィング比率を大きくすると，ラップらしさは向上しても自然性が低下し，その結果，評価値の上昇が抑制されたと考えられる．この仮定に基づけば，今回の上限値である 3.0 を超えたさらに大きなスウィング比率の比率間では，自然性の極端な低下によって評価値が再び低下し，相対的にラップらしいと感じられるスウィング比率の上限が存在することが示される可能性がある．

6.2 被験者の属性が評価に与える影響

被験者グループ別の分析結果から，ラップらしさの知覚に対する被験者属性の影響を考察する．

- □ 音楽経験の有無：音楽経験がないグループが，経験があるグループに比べてラップらしいと高く評価する傾向が確認された．これは，音楽経験を持つ人が，詳細なリズムのズレや表現の差異により敏感になることによる批判的聴取を行った可能性を示唆している．彼らはより厳格な基準で音源を評価した結果，相対的に低い評価になったことが考えられる．
- □ ラップ聴取習慣の有無：ラップ聴取習慣があるグループは，ないグループに比べラップらしいと高く評価する傾向が見られた．この結果は，ラップらしさの許容範囲の違いを示唆している．習慣的にラップを聴いている人は，ラップミュージック特有のリズムやグルーヴの多様性に慣れ親しんでいるため，本実験で作成した様々なスウィング比率の音源をより広くラップらしい表現として受け入れやすかったことが考えられる．

これら属性別分析より，「ラップらしさ」の知覚は被験者個人の音楽的背景に影響されることが示唆される．本実験の被験者数が十分ではないことから，これ以上の詳細な分析は困難である．しかしながら，「音楽経験はないがラップ聴取習慣はある」といった，より詳細な属性別のグループに被験者を分類し，その評価傾向を分析することで，属性ごとの更なる認知特性が明らかになる可能性がある．

6.3 今後の検討課題

アンケート調査において，被験者から「語尾がフォール（f_0 が滑らかに下がる表現技法）している音源は特にラップらしく感じられた」という意見が多く寄せられた．この定性的な意見は，ラップらしさに影響を与えるパラメータとして，スウィング比率だけでなく，ピッチの表現技法（f_0 フォールなど）が重要であることを示唆しており，今後の研究で注目すべき要素である．なお，今回実験に使用した音源はすべて合成音声であり，人間が歌唱した実音声は含まれていない．合成音声と実音声を比較した場合に，ラップらしさの認知がどのように変化するかについては，今後のさらなる議論と検証が必要である．

7. おわりに

本稿では，実音声と合成音声のスウィング比率を比較し，合成音声がスウィングを十分に再現できていない可能性を示した．主観評価実験の結果から，スウィング比率とラップらしさの知覚に一定の相関が示された．また，属性別分析により，音楽的な背景といった被験者の属性によって評価結果に統計的に有意な差があることが示された．

次の試みとして，各音源の該当箇所に新たにフォール記号を定義・付与した楽譜を作成し，再学習を行った．図 6 は，同一楽譜を入力とした 3 つの異なる歌唱モデルによる各生成音源の f_0 軌跡である．横軸は時間，縦軸は基本周波数（f_0）を示す．時刻 2.0 s 辺りに着目すると，フォール付与モデルでは，フォール表現の再現性が向上していることが確認できる（図 6 点線）．このモデルでは，ラベルを one-hot ベクトル化するための質問ファイルにおいて，音素位置に関する質問を削除した．これにより，音素位置に

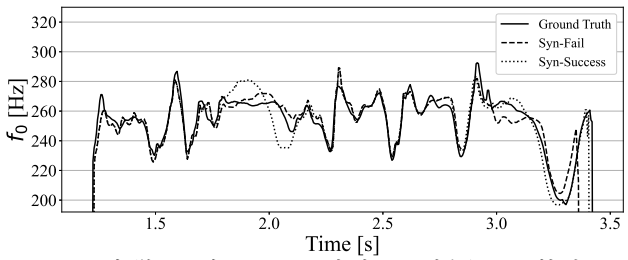

図 6: 同一楽譜より各モデルで生成した音源の f_0 軌跡.
（実線：実音声，破線：質問ファイルの編集なしの合成音声，点線：質問ファイルの編集ありの合成音声）

基づく f_0 バイアスを軽減し，フォール記号による f_0 制御の学習精度を高めた．質問ファイルの編集を行わなかった場合には，記号を与えてもフォールが再現されないことが確認された (図 6 破線)．今後は，このフォール付与モデルの性能評価を行い，複数の表現技法を組み合わせることで，ラップスタイル歌声合成の更なる表現力向上を目指す．

謝辞 本研究は，明治大学科学技術研究所重点研究 (B) による支援を受けて実施された．本実験に用いた歌唱データは，クリプトン・フューチャー・メディア株式会社から提供を受けたものである．

参考文献

[1] Pucher, M., Villavicencio, F., Yamagishi, J.: Development of a statistical parametric synthesis system for operatic singing in German, Speech Synthesis Workshop, pp. 68–73 (2016).
[2] Zheng, M., Bai, P., Shi, X., Zhou, X., Yan, Y.: FT-GAN: fine-grained tune modeling for Chinese opera synthesis, Proc. AAAI Conference on Artificial Intelligence, vol. 38, no. 17, pp. 19697–19705, (2024).
[3] Konstantinos, M., Nikolaos, E., Alexandra, V., Myrsini, C., Panos, K., Georgios, V., Georgia, M., J. S. S., Hyoungmin, P., Pirros, T., Aimilios, C.: Rapping-Singing Voice Synthesis based on Phoneme-level Prosody Control, Speech Synthesis Workshop (2021).
[4] 三林亮太，山本岳洋，佃洸摂，渡邉研斗，中野倫靖，後藤真孝，大島裕明：ラップバトルにおける逆向き生成によるライムを含む返答バース生成, 情報処理学会論文誌データベース TOD, vol. 17, no. 2 pp. 28–39 (2024).
[5] Savery, R., Zahray, L., and Weinberg, G.: Shimon the rapper: a real-time system for human-robot interactive rap battles, Proc. ICCC2020 (2020).
[6] 才野慶二郎，大浦圭一郎，橘誠，剣持秀紀，徳田恵一：ラップスタイル歌声合成の検討, 情報処理学会音声言語情報処理研究会，vol. 90, no. 7, pp. 1–6 (2012).
[7] Ning, Z., Wang, S., Jiang, Y., Yao, J., He, L., Pan, S., Xie, L.: Drop the beat! freestyler for accompaniment conditioned rapping voice generation, Proc. AAAI Conference on Artificial Intelligence, vol. 39, no. 23, pp. 24966–24974 (2025).
[8] Bonada, J., Loscos, A., Mayor, O., Kenmochi, H.: Sample-based singing voice synthesizer using spectral models and source-filter decomposition, Proc. MAVEBA2003, 175–178 (2003).
[9] VOCALOID：VOCALOID(ボーカロイド・ボカロ) 公式サイト (online)，入手先 ⟨https://www.vocaloid.com/⟩ (2025.10.28).
[10] Kenmochi, h.: VOCALOID and Hatsune Miku phenomenon in Japan, Proc. INTERSINGING2010, pp. 1–4 (2010).
[11] 才野慶二郎：歌声の合成における応用技術——歌声合成システム——，日本音響学会誌，vol. 75, no. 7, pp. 406–411 (2019).
[12] Bonada, J., Merlijn B.: Generation of growl-type voice qualities by spectral morphing, Proc. ICASSP2013, pp. 6910–6914 (2013).
[13] Saino, K., Zen, H., Nankaku, Y., Lee, A., Tokuda, K.: An HMM-based singing voice synthesis system, Proc. INTERSPEECH2006, pp. 2274–2277 (2006).
[14] Oura, K., Mase, A., Yamada, T., Muto, S., Nankaku, Y., Tokuda, K.: Recent development of the HMM-based singing voice synthesis system-Sinsy, Speech Synthesis Workshop, pp. 211–216 (2010).
[15] Nishimura, M., Hashimoto, K., Oura, K., Nankaku, Y., Tokuda, K.: Singing Voice Synthesis Based on Deep Neural Networks, Proc. INTERSPEECH2016, pp. 2478–2482 (2016).
[16] Blaauw, M., Bonada, J.: A neural parametric singing synthesizer modeling timbre and expression from natural songs, Applied Sciences, vol. 7, no. 12, p. 1313 (2017).
[17] Aaron van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, Proc. 9th ISCA Workshop on Speech Synthesis Workshop, p. 125 (2016).
[18] Kong, J., Kim, J., Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, Advances in neural information processing systems, vol. 33, pp. 17022–17033 (2020).
[19] Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Hassabis, D.: Parallel wavenet: Fast high-fidelity speech synthesis, Proc. ICML2018, pp. 3918–3926 (2018).
[20] Chen, J., Tan, X., Luan, J., Qin, T., Liu, T. Y.: Hifisinger: Towards high-fidelity neural singing voice synthesis, arXiv preprint arXiv:2009.01776 (2020).
[21] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Wu, Y.: Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, Proc. ICASSP2018, pp. 4779–4783 (2018).
[22] Cho, Y. P., Yang, F. R., Chang, Y. C., Cheng, C. T., Wang, X. H., Liu, Y. W.: A survey on recent deep learning-driven singing voice synthesis systems, Proc. AIVR2021, pp. 319–323 (2021).
[23] Chunhui, W., Zeng, C., He, X.: Xiaoicesing 2: A High-Fidelity Singing Voice Synthesizer Based on Generative Adversarial Network, Proc. INTERSPEECH2023, pp. 5401–5405 (2023).
[24] Zhao, J., Chetwin, L. Q. H., Wang, Y.: Sintechsvs: A singing technique controllable singing voice synthesis system, Trans. ASLP2024, vol. 32, pp. 2641–2653 (2024).
[25] Zhang, Y., Huang, R., Li, R., He, J., Xia, Y., Chen, F., Zhao, Z.: Stylesinger: Style transfer for out-of-domain singing voice synthesis, Proc. AAAI Conference on Artificial Intelligence, vol. 38, no. 17, pp. 19597–19605 (2024).
[26] Umbert, M., Bonada, J., Goto, M., Nakano, T., Sundberg, J.: Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges, IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 55–73 (2015).
[27] Scherer, K. R., Sundberg, J., Fantini, B., Trznadel, S., Eyben, F.: The expression of emotion in the

singing voice: Acoustic patterns in vocal performance, JASA2017, vol. 142, no. 4, pp. 1805–1815 (2017).
[28] Hono, Y., Murata, S., Nakamura, K., Hashimoto, K., Oura, K., Nankaku, Y., Tokuda, K.: Recent development of the DNN-based singing voice synthesis system—sinsy, Proc. APSIPA ASC2018, pp. 1003–1009 (2018).
[29] Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M.: Flow Matching for Generative Modeling, Proc. ICLR2023 (2023).
[30] Dreamtonics: SynthesizerV Studio 2 Pro — AI 歌声合成プラグイン (online), 入手先 ⟨https://dreamtonics.com/ja/synthesizerv/⟩ (2025.11.9).
[31] Dreamtonics: Synthesizer V Studio 1.9.0b1 アップデート: ラップ、広東語への対応ほか (online), 入手先 ⟨https://dreamtonics.com/ja/synthesizer-v-studio-1-9-0b1-update-rap-cantonese-and-more/⟩ (2025.11.9).
[32] Adams, K.: On the metrical techniques of flow in rap music, Music Theory Online, vol. 15, no. 5 (2009).
[33] RedBull: ヒップホップ サブジャンル A to Z(online), 入手先 ⟨https://www.redbull.com/jp-ja/different-types-of-hip-hop-guide⟩ (2025.11.11).
[34] Yamamoto, R., Yoneyama, R., and Toda, T.: NNSVS: a neural network-based singing voice synthesis toolkit, Proc. ICASSP2023, pp. 1–5 (2023).
[35] Morise, M.: Harvest: a high-performance fundamental frequency estimator from speech signals, Proc. INTERSPEECH2017, pp. 2321–2325 (2017).
[36] Morise, M., Yokomori, F., and Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE TRANSACTIONS on Information and Systems, vol. 99, no. 7, pp. 1877–1884 (2016).
[37] Lindsay, K., Nordquist, P.: Pulse and swing: quantitative analysis of hierarchical structure in swing rhythm, The Journal of the Acoustical Society of America, vol. 122, no. 5, pp. 2945–2945 (2007).
[38] Kawahara, T., Lee, A., Kobayashi, T., Takeda, K., Minematsu, N., Sagayama, S., Itou, K., Ito, A., Yamamoto, M., Yamada, A., and others.: Free software toolkit for Japanese large vocabulary continuous speech recognition, Proc. ICSLP2000, vol. 4, pp. 476–479 (2000).
[39] Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection, BMC Bioinformatics, vol. 7, no. 1, p. 91 (2006).
[40] Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization, Communications of the ACM, vol. 64, no. 3, pp. 107–115 (2021).

Vol.2025-NL-266 No.23
Vol.2025-SLP-158 No.23
2025/12/17

付　　録

A.1　各モジュール使用モデル

表 A·1: NNSVS の 3 モジュールにおける使用モデル

Time-lag Module	
Time-lag Model	Variance Predictor
Duration Module	
Duration Model	Variance Predictor
Acoustic Module	
lf0 Model	Bidirection LSTM Residual F0 Non-Attentive Decoder
Shared Encoder	LSTM Encoder
MGC Decoder	Fast Fourier Convolution LSTM
VUV Decoder	Fast Fourier Convolution LSTM
Bap Decoder	Fast Fourier Convolution LSTM

A.2　タイムラグモデル詳細

表 A·2: タイムラグモデル　ハイパーパラメータ

Hyperparameter	**Value**
Input Dimension	82
Output Dimension	1
Hidden Dimension	32
Num Layers	3
Kernel Size	3
Dropout	0.5
Batch Size	8
Using MDN(Mixture Density Network)	True
Num Gaussians Distribution	4
Initialize Type	Kaiming Normal
Num Total Parameters	0.013 milion

A.3　持続時間モデル詳細

表 A·3: 持続時間モデル　ハイパーパラメータ

Hyperparameter	**Value**
Input Dimension	82
Output Dimension	1
Hidden Dimension	256
Num Layers	5
Kernel Size	5
Dropout	0.5
Batch Size	8
Using MDN(Mixture Density Network)	True
Num Gaussians Distribution	4
Initialize Type	Kaiming Normal
Num Total Parameters	1.403 milion

A.4　音響モデル詳細

表 A·4: 音響モデル　ハイパーパラメータ

Hyperparameter	**Value**
Input Dimension	86
Output Dimension	67
lf0 Model	
Input Dimension	86
Output Dimension	1
Embedding Dimension	256
FF Hidden Dimension	256
Conv Hidden Dimension	128
LSTM Hidden Dimension	64
Num LSTM Layers	2
Decoder Layers	1
Decoder Hidden Dimension	256
Pre-net Layers	0
Pre-net Hidden Dimension	16
Pre-net Dropout	0.5
Zoneout	0.0
Reduction Fuctor	4
Input lf0 Index	51
Output lf0 Index	0
Shared Encoder	
Input Dimension	86
Output Dimension	1024
Embedding Dimension	256
Hidden Dimension	512
Num Layers	3
Dropout	0.0
Initialize Type	Kaming Normal
MGC Decoder	
Input Dimension	1024
Output Dimension	60
FF Hidden Dimension	1024
Conv Hidden Dimension	512
LSTM Hidden Dimension	256
Num LSTM Layers	2
Dropout	0.1
VUV Decoder	
Input Dimension	1026
Output Dimension	1
FF Hidden Dimension	256
Conv Hidden Dimension	128
LSTM Hidden Dimension	64
Num LSTM Layers	2
Dropout	0.1
Bap Decoder	
Input Dimension	1026
Output Dimension	5
FF Hidden Dimension	256
Conv Hidden Dimension	128
LSTM Hidden Dimension	62
Num LSTM Layers	2
Dropout	0.0