情報処理学会東北支部研究報告
IPSJ Tohoku Branch SIG Technical Report

Vol.2025-1 No.2-2
2025/2/22

拡散モデルを用いた歌声のジャンル変換の検討

柏倉直樹† 小坂哲夫†

従来，歌声変換とは歌詞や音程は変えずに入力歌手の声質を目標歌手の声質へ変換する技術であった．これまでの歌声変換では GAN や拡散モデルを用いて変換音声の自然性や性能向上について研究されてきた．本稿では歌声変換の新たな可能性として，拡散モデルを用い，同一歌唱者を前提とした any-to-many の歌声のジャンル変換を行う．拡散モデルをベースとした DDSP-SVC により，ポピュラー歌唱⇔演歌歌唱の変換を行い，自然性や変換性能の検討を試みた．

A study on genre conversion of singing voices using diffusion model

Naoki kashiwagura† and Tetsuo Kosaka†

Traditionally, singing voice conversion has been a technology that converts the voice quality of an input singer to that of a target singer without changing the lyrics or pitch. In previous singing voice conversion studies, GANs and diffusion models have been used to study the naturalness and performance of the converted voice. In this paper, as a new possibility for singing voice conversion, a diffusion model is used to perform any-to-many singing voice genre conversions assuming the same singer. Using DDSP-SVC based on a diffusion model,we converted popular singing to “Enka” singing and attempted to examine the naturalness and conversion performance.

1. はじめに

近年の深層学習の進展に伴い,音声合成や音声変換技術は広範な応用分野で注目を集めている．その中でも,声質変換技術[1]は目標話者を定めて入力話者の声質を目標の話者の声質に変換する．この技術は特に，話者の特徴を維持しながら柔軟に音声特性を操作する点で多くの研究が行われている．多言語翻訳やエンターテイメント分野での応用が進んでいる．

一方で,歌声変換(SVC: Singing Voice Conversion)[2]は声質変換の応用技術である．声質変換と歌声変換は多くの点で類似しているが,音高(ピッチ),リズム,強弱という点では異なる．歌声変換は音程や歌詞は維持したまま声質を変換する技術である．歌声変換の応用例として,音楽制作,仮想アーティストの創作,障害を持つ方の歌唱支援,さらにはヤマハが開発した「なりきりマイク」[3]といったような音楽の分野で多岐にわたる分野での活用が期待されている．これまで主に敵対的生成ネットワーク(GAN)[4][5]や拡散モデルを用いた,異なる歌唱者間の歌声変換の音質向上や自然性の向上,性能向上を目的とした研究が行われてきた．

しかしながら,同一歌唱者の歌声変換に関する研究事例は少ない．その要因の一つとして,日本語の歌唱データセットの不足が挙げられる．特に,同一歌唱者が異なる歌唱スタイル（ジャンル）で歌ったデータセットはほとんど存在しなかった．しかし,近年,「jaCappella コーパス」[6]が公開され,同一歌唱者による多ジャンルの日本語アカペラデータが利用可能となった．

拡散モデルを使用することで,高品質な音声かつ自然な音声を生成しやすく,多様な音声分布を学習できるため,変換度合いを変更しながら音声変換を行え,かつ柔軟性が高い．本研究では,歌声変換技術の新たな可能性として,同一歌唱者のジャンル変換(歌唱スタイル変換)を試みる．すなわち,入力歌手と出力歌手の声質は同一でありながら,歌唱スタイルのみを異なるジャンルへと変換することを目的とする．音程や歌詞を保持したまま,歌い方を変換することで,より表現力豊かな歌声変換につながると考えられる．

以上を踏まえ,本研究では拡散モデルを用いた any-to-many の同一歌唱者間でのジャンル歌声変換での品質についての検討を行った．any-to-many では任意の歌唱者の歌声を複数の音楽ジャンルの歌い方に変換することができる．

† 山形大学
Yamagata University.

Vol.2025-1 No.2-2
2025/2/22

2. 音声変換モデル

2.1 拡散モデル(Diffusion Model)

拡散モデル[7][8]とはデータをノイズに変換し(前向きプロセス),そのノイズから元データを復元する(逆拡散プロセス)することで,新しいデータを生成する．このプロセスはマルコフ連鎖と呼ばれる確率的な遷移を利用している．図 1 に拡散モデルの概要図を示す.

- 前向きプロセス

音源データに少しずつノイズを加えていくプロセスである．最終的にデータは完全なランダムノイズに変換される．確率的な変換を何ステップも繰り返していく．

- 逆拡散プロセス

ノイズの状態から元のデータを復元するプロセスである．ステップごとにノイズを除去していき,最終的に元のデータと復元したデータの誤差が最小になるように学習をする.

2.2 DDSP-SVC

DDSP-SVC(Differentiable Digital Signal Processing Singing Voice Conversion) [9]は拡散モデルをベースとした音声変換モデルである．図 2 に DDSP-SVC の構造図を示す．歌詞や音素に関する情報を持つ HuBERT 特徴量,声の高さに関する基本周波数,音量,話者 ID を条件としてメルスペクトログラムを生成する．音声波形の生成には事前学習済みの HiFi-GAN を使用している．訓練時は時刻を t = 1, … , k_{max}≤1000 までステップごとにランダムに選び,時刻に応じてノイズを加え,それを予測するように学習を行う．音声変換時は時刻 k を選び,1, … ,k のノイズ除去ループを実行し最終的には時刻 t=0 の状態を得る(元のデータの復元)．DDSP-SVC では,変換強度を柔軟に調整することができる．入力した音声にどの程度ノイズを加えるかにより,変換強度を調整することが可能となる．さらに,話者 ID(正の整数)で話者ごとの情報を管理するため,複数話者の同時学習が可能となる．DDSP-SVC では 4 つの条件から生成したメルスペクトログラムに対し,ノイズを与える．ノイズが加わったメルスペクトログラムをモデルに通すことでどのようなノイズが加わったのかを予測する(予測ノイズ).

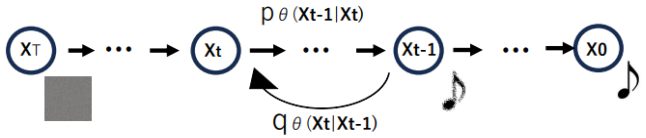

図 1　拡散モデルの概要

加えたノイズと予測ノイズの平均二乗誤差(MSE)が最小となるように学習を行う．変換したい音声のメルスペクトログラムに一定のノイズを加え,そこからノイズを除去することで新たなメルスペクトログラムを生成する．このメルスペクトログラムを事前学習済みの HiFi-GAN ボコーダーに通すことで変換音声が生成される.

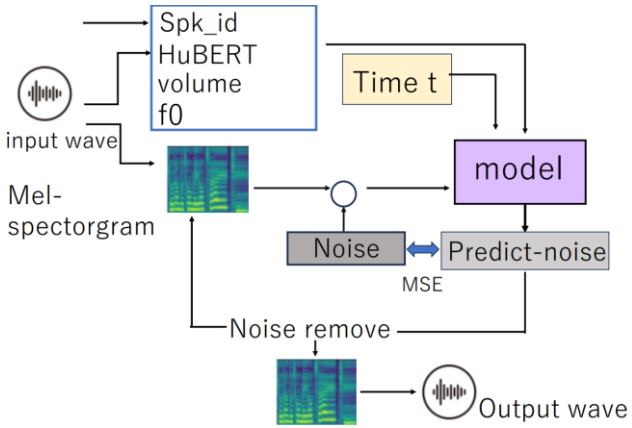

図 2　DDSP-SVC の構造図

2.3 提案手法

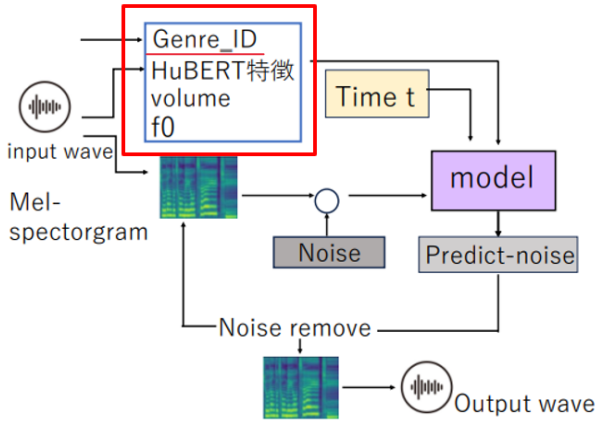

図 3　提案手法

Vol.2025-1 No.2-2
2025/2/22

DDSP-SVC においては話者 ID を利用して複数話者の学習が可能であった．話者 ID で条件付け学習をするため,音声生成時に話者 ID を与えることによって,話者の切り替えができる．よって 1 つのモデルで複数話者の音声生成が可能になる．本研究ではこの話者 ID をジャンル ID として使用する．図 3 に提案手法を示す．異なるジャンルを ID で管理し,そのジャンルの歌唱データを与えて学習する．音声生成の際には入力音声にジャンル ID を与えることで様々なジャンルの歌声を生成することができる．これにより任意の歌唱者の歌声を複数ジャンルの歌声へ変換できる any-to-many 変換が可能になる．多数話者の学習データが使用できれば,ジャンルと話者の組み合わせを「ジャンル×話者 ID」として与えて学習することで,様々な歌唱者に対応することも可能と考えられる．本研究ではコーパスの関係上,歌唱者 1 名を使用して歌声変換を行っている．

3. 使用コーパス

3.1 jaCappella コーパス

jaCappella コーパスは日本語のアカペラボーカルアンサンブルのコーパスである．ボーカルアンサンブル曲の楽譜,個々のボーカルパート(リードボーカル,ソプラノ,アルト,テナー,ベース,ボーカルパーカッション)の個別オーディオ録音で構成されている．著作権が切れた日本の童謡を編曲したものである．ジャンルは 10 のジャンルで構成されていて,それぞれ 5 曲ずつ収録されている．ジャンルは Jazz, Punk rock, bossa nova, ポピュラー, reggae, 演歌, neutral, ballad, EDM, Soul funk である．総時間に関してはそれぞればらつきがある．サンプリング周波数は 48kHz で収録されている．音声ファイルはモノラル WAVE 形式で提供されており,歌手は全員日本語を母国語としている．本研究でボーカルパートのうち,リードボーカルの音源を使用する．同一歌唱者が存在しないジャンル,またジャンルとしての特徴が分かりづらい EDM 歌唱や reggae 歌唱などといったジャンルは使用しない．使用するジャンルは演歌とポピュラーの 2 ジャンル間の双方向変換の分析および実験を行う．

3.2 データ拡張

jaCappella コーパスは 1 ジャンルに 5 曲とかなり少ないデータしか存在しない．音声変換ツール sox を用いて,セントを±50,±100 にピッチシフトを行い,データ量を 5 倍に拡張した．拡張後のデータ量については表 1 に示す．

表 1 jaCappella コーパスのデータ拡張後の詳細

ジャンル	データ拡張前	データ拡張後	歌唱者
演歌	361.1 秒(約 6 分)	1805.5 秒(約 30 分)	野村美和
ポピュラー	352.5 秒(約 6 分)	1762.5 秒(約 29 分)	野村美和

4. 音響的ジャンル特徴

4.1 演歌

演歌は「こぶし」が特徴的な歌唱方法である．「こぶし」は母音を強調する歌い方である．短時間に声の高さの急激な上昇・下降で F0 に特徴が表れる．スペクトログラムにおいて,こぶし付近には倍音構造とともに摩擦部分が観察される．これは声の高さを上げようとして声帯を引き延ばしていることの現れである(10)．また演歌歌唱はビブラートが強いのも特徴である．

4.2 ポピュラー

ポピュラーという歌唱には様々なテクニックが使用されている．一番多くみられる特徴としてビブラートである(11)．スペクトログラムに波線のような推移が見られる．これがビブラートの特徴になる．また,ブレスもみられる．ブレスは男女に関わらず,1.6kHz~1.7kHz の周波数帯域にパワーのピークが存在することが多い(12)．

5. 評価実験

5.1 実験概要

ポピュラーと演歌の双方向に歌声を変換して主観評価実験を実施した．主観評価ではポピュラーから演歌及び,演歌からポピュラーへの変換を行い,品質の比較検証を行った．評価データのポピュラー歌唱の音源と演歌歌唱の音源を 2 小節に分割して使用した．評価データについてはポピュラー歌唱の音声,ポピュラーから演歌の変換音声,演歌歌唱の音声,演歌からポピュラーの変換音声,それぞれ 9 曲ずつ使用している．実験は被験者 15 名に対して行った．音声 A,B を比較してもらい,それぞれの音声の品質(MOS),及び音声がどのジャンルに感じたかを相対評価で行った. MOS とは評価対象の品質を測定するため, 5 段階評価で音声の自然性について評価を行う。数値が高いほど高品質な音声となる．音声 A,B のジャンルは必ずポピュラー or 演歌であり,同一曲である．なお,ジャンルに関しては「ポピュラー」「演歌」「わからない」の 3 択の中から選択してもらった．DDSP-SVC の学習条件はサンプリング周波数を 44.1kHz,バッチサイズ 8,エポック数 100000,学習率 5.0×10^{-4} に設定した。学習データに演歌歌唱音源を約 27 分,ポピュラー歌唱音源を 26 分使用した．

5.2 実験結果

自然性の評価結果を表 2 に示す．Ground truth(GT)とは、正解データのことであり、この場合は変換前の人間の歌声のことである．結果より,ポピュラー(GT)と演歌(GT)の品質は良好である．ポピュラー(GT)とポピュラー→演歌を比較すると品質が低下した．ポピュラーを変換した際に韻律が崩れてしまったことが原因と考えられる．演歌(GT)

Vol.2025-1 No.2-2
2025/2/22

と演歌→ポピュラーを比較すると,こちらは比較的自然性は保たれている．演歌へ変換するよりもポピュラーへ変換したときの方が,ポピュラーを反映した変換が行えたことで,韻律などの不自然さがなかったと考えられる．被験者からは「全体的に品質はいいように感じた．しかしながらポピュラーと判断した音声にはボーカロイドのような機械音やノイズがやや入っているような音声に聞こえた．韻律が崩れているように感じる箇所がある」などの意見が寄せられた．

図 4 にジャンル認識率についての結果を示す．ポピュラー(GT)とポピュラー→演歌を比較すると両者ともに 5 割程度が正しく認識できているが,残り 5 割は「分からない」を含め不正解していることから正しく認識できているとは言えない．演歌とポピュラーの中間のような音声になってしまい演歌としての特徴が上手く反映できている変換音声ではなかったのが要因と考えられる．一方で演歌(GT)とポピュラー→演歌について比較すると,両者ともにかなりの割合で正しい認識率を得ることができた．演歌(GT)にあった強いビブラートを効かせた歌い方がポピュラーへ変換したときには,ビブラートが抑えられている音声になっていたことが一番の要因であった．被験者からも「ビブラートの強弱で演歌なのかポピュラーなのかを判断した」という意見がほとんどであった．

表 2　自然性の MOS 評価

ポピュラー (GT)	ポピュラー→演歌	演歌(GT)	演歌→ポピュラー
4.18	3.62	4.44	3.98

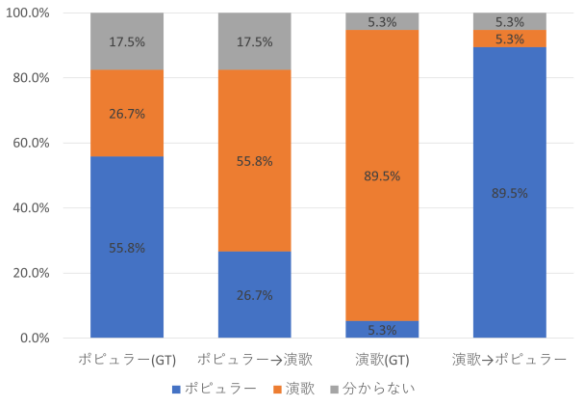

図 4　ジャンル認識率の結果

6.　分析

ポピュラー→演歌,演歌→ポピュラーに変換した音声のスペクトログラム及び基本周波数について分析を行う．

6.1　ポピュラー歌唱から演歌歌唱への変換

入力音声にポピュラー歌唱を使用し,演歌歌唱へと変換した音声を生成した．元音声のスペクトログラム及び基本周波数を図 5 に,変換音声のスペクトログラム及び基本周波数を図 6 に示す．元音声からスペクトログラムに若干の変化が見られるが，これを摩擦部分と呼べるかは微妙なところである．母音間の調音変換によるものだと考えられる．子音と母音の遷移中に摩擦的な要素が生じることがある．例えば、母音間での息漏れが多いと摩擦的なノイズが発生することがある．また 24.8 秒付近の基本周波数について上昇・下降が見られるが，これはいわゆる「半ピッチ・倍ピッチエラー」になっていると考えられる．つまり変換したことで演歌としての要素を反映できたとは言えない．

6.2　演歌歌唱からポピュラー歌唱への変換

入力音声に演歌歌唱を使用し,ポピュラー歌唱へと変換した音声を生成した．元音声のスペクトログラム及び基本周波数を図 7 に,変換音声のスペクトログラム及び基本周波数を図 8 に示す．元音声である演歌歌唱と変換音声であるポピュラー歌唱のスペクトログラムの摩擦部分及び基本周波数を比較すると,変化はなかった．このことから,変換音声のポピュラー歌唱にも演歌歌唱の要素が残存していると考えられる．変化があった部分はスペクトログラムに現れるビブラートである．元音声ではスペクトログラムが濃く波線のように推移していることが分かる．

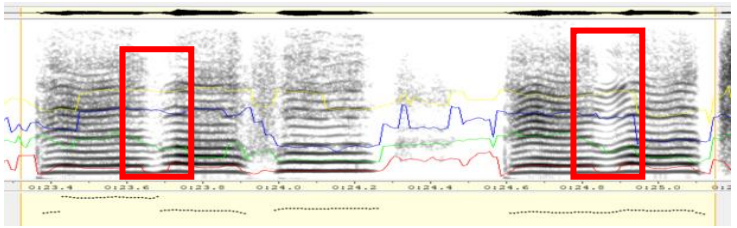

図 5　元音声のスペクトログラム及び基本周波数

Vol.2025-1 No.2-2
2025/2/22

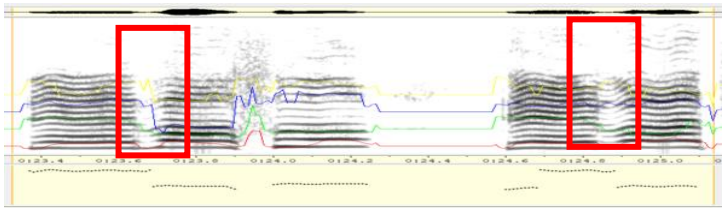

図 6　変換音声のスペクトログラム及び基本周波数

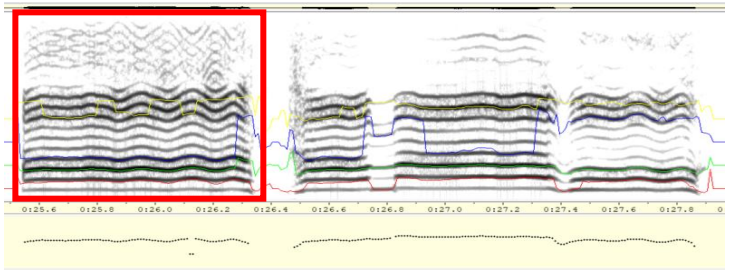

図 7　元音声のスペクトログラム及び基本周波数

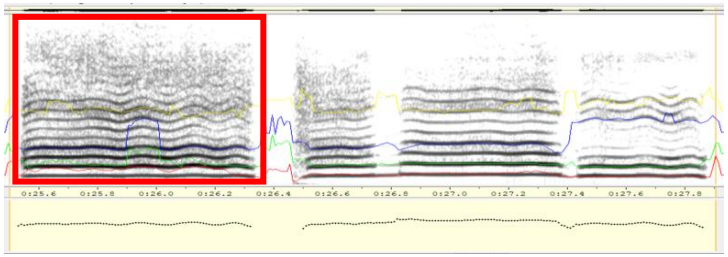

図 8　変換音声のスペクトログラム及び基本周波数

一方で変換音声では,スペクトログラムがやや平坦に推移していることが分かる．このことから,変換音声では元音声の強いビブラートが抑えられて変換されている．ポピュラー歌唱の母音を強調せずに安定した歌い方が反映されていると推測できる．

7.　結論

主観評価実験及び分析から,ポピュラー→演歌変換では演歌の特徴を確認することができず，その変化はわずかであったためポピュラーと演歌の中間のような音声になった．そうした要因が,韻律の乱れやジャンルとしてどちらかわからないことに繋がったと考えられる．一方で,演歌→ポピュラー変換では入力音声である演歌の特徴が残存していた．しかしながら,入力音声の演歌にあった強いビブラートが変換音声ではビブラートが抑えられていた．そうした要因が,自然性の維持,ジャンル認識の正答率が高かったことに繋がったと考えられる．以上のことから,今回の場合,ジャンルを判断するにはこぶしが影響しているとは考えにくく,ビブラートの強弱によって判断している可能性が大きい．また,被験者の音楽的知識や歌唱テクニックに関する知識があるかないかによってもジャンルを判断するのに影響があった一因でもあった．今後の方針として,今回ジャンル ID を使用したが，これをジャンル×話者 ID として他人間のジャンル変換が行えるかを検証する．また,新たな手法としてアカペラ歌唱音源からジャンル特徴をベクトルとして抽出し,DDSP-SVC へと入力を行うことでジャンル特徴を反映したメルスペクトログラムを生成できると考えられる．このようにジャンルベクトルを与えることにより,歌い方の制御ができる可能性が期待できる．ID ではなくベクトルを使用することにより,演歌らしさの強度を制御したり,演歌とポピュラーの中間の歌い方をさせたりなど多様な変換の実現が期待できる．

参考文献

1)　吉田天哉,田村仁,”機械学習を用いた声質変換手法”,第 82 回全国大会講演論文集 2020(1),pp187-188,2020-02-20
2)　才野慶二郎,”歌声の合成における応用技術”,日本音響学会誌 75 巻 7 号,pp406-411,2019
3)　ヤマハ株式会社: リアルタイムで“あの人”の歌声になれ る AI 歌声変換技術『TransVox™』の実証研究を開始「なりきりマイク」で「Every Little Thing」持田香織さんの声を再現, https://www.yamaha.com/ja/news_release/2022/22082401/, 2023
4)　Ian J,Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, “Generative Adversarial Networks”,arXiv,2014
5)　堀幸央,佐伯和哉,小坂哲夫,”敵対的生成ネットワークを用いた歌声変換の各種検討”,日本音響学会研究発表講演文集,2021
6)　Tomohiko Nakamura; Shinnosuke Takamichi,Naoko Tanji,Satoru Fukayam,Hiroshi Saruwatari, “jaCappella Corpus: A Japanese a Cappella Vocal Ensemble Corpus”,ICASSP,2023
7)　J.Ho et al.,“Denoising Diffusion Probabilistic Models”,arXiv,2020
8)　A.Nichol et al., “Improved Denoising Diffusion Probabilistic Models”,arXiv,2021
9)　Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, Adam Roberts,”DDSP: Differentiable

Vol.2025-1 No.2-2
2025/2/22

Digital Signal Processing”,arXiv,2020
10) 川原繁人,古澤里菜,”城南海の「こぶし」の音声学的特徴と音譜上の分布について:「アイツムギ」と「あなたに逢えてよかった」をもとに”,慶応義塾大学言語文化研究所紀要 第 54 号 pp53-77,2023
11) 川本聡胤,” 音楽学的ポピュラー音楽研究(3) : ジャンルとスタイル”,フェリス女学院大学音楽学部紀要,pp1-22,2022
12) 山本雄也,中野倫靖,後藤真孝,寺澤洋子,平賀譲,” ポピュラー音楽における模倣歌唱を用いた歌唱テクニックの頻度・特徴・生起箇所の分析”,情報処理学会研究報告,Vol. 2021-MUS-132 No20,pp1-8, 2021,