歌声合成と変換を用いた未経験者音声からのオペラ歌唱音声合成*

☆菅原 碧斗 (神戸大)，岸本 宗真 (メック株式会社)，足立 優司 (メック株式会社)，
田井 清登 (メック株式会社)，高島 遼一 (神戸大)，滝口 哲也 (神戸大)

1 はじめに

歌声合成技術は娯楽分野において広く普及し，故人や声を失った患者の歌声を再現する手法として注目を集めている．近年では深層ニューラルネットワーク (Deep Neural Networks; DNNs) による音声合成技術の発展に伴い，歌声合成の分野においても高品質な音声の合成が可能になっている．

また，近年では人間らしい表現をもつ歌声の合成に関する研究が行われている．従来の歌声合成では主に童謡や J-POP といったジャンルの歌声音声を対象として行われていたが，本研究では童謡や J-POP とは，ビブラートやピッチ，母音などの特徴が異なっているアカペラオペラ歌唱音声 [1, 2, 3] を対象とする．本研究では，任意の歌詞付き楽譜を入力とし，オペラ歌唱未経験ユーザの発話音声を用いてオペラ歌唱音声を合成可能なシステムの実現を目的とする．

発話音声を用いてオペラ歌唱音声を合成するためのアプローチとして大きく二つ挙げられる．一つ目は，プロのオペラ歌唱音声を用いて学習したオペラ歌唱音声合成モデルに対して，ユーザの発話音声を用いて話者適応を行うアプローチである．しかし話者適応のアプローチでは，通常発話のコンテキストラベルと発話音声のデータを用いてファインチューニングするため，モデルが発話音声の合成に過適合することが懸念される．二つ目は，声質変換技術を用いて，オペラ歌唱音声の声質をユーザの声質に変換するアプローチである．オペラ歌唱音声特有の特徴と話者依存の特徴が独立なものと仮定すると，プロのオペラ歌唱音声から話者性のみをユーザのものに声質変換できれば，ユーザの声でオペラ特有の性質を備えた歌唱音声が生成可能と期待できる．そのため，本研究では後者の声質変換手法を検討する．我々は以前，Diff-SVC を用いたオペラ歌唱音声合成手法を検討した [4]．また，更なる合成品質の向上を目的として，マルチ受容野混合層 (MRF) を用いた中高域強調ネットワークを検討した [5]．しかし以前の方法では変換元となるプロオペラ歌唱音声が必要であるため，任意の歌を合成することができないという課題があった．本研究では，声質変換で用いるオペラ歌唱音声を歌声合成手法である DiffSinger で合成し，合成した音声を Diff-SVC で声質変換することで，任意の歌のユーザオペラ歌唱合成が可能なシステムを検討する．

2 DiffSinger

DiffSinger [6] は拡散モデルの一つである denoising diffusion probabilistic model (DDPM) をベースとした歌声合成モデルであり，主な構成要素は Variance モデルと Acoustic モデルの 2 つである．Variance モデルでは MIDI と歌詞，音素を入力として音素継続長やピッチ，息遣い等を予測し出力する．Acoustic モデルでは，音素や前述の Variance モデルで出力された音素継続長，ピッチ，息遣い等を入力としてメルスペクトログラムを出力する．

更に，前述の Variance モデルや Acoustic モデル中で用いられている拡散モデルの高速化と音質の改善のため，浅い層の拡散モデルと境界予測器を導入している．推論において，楽譜から抽出した情報 x から真のメルスペクトログラムとの L1 Loss で学習したデコーダーを用いて $\widetilde{M}$ を作成する．ここで t が十分に大きいとき，$\widetilde{M}_t$ と M_t とは一致することから，ガウシアン白色ノイズから逆過程を行うのではなく，$\widetilde{M}_t$ と M_t が一致するような最小のステップ数 $t = k$ を境界予測器によって予測し，ステップ k における中間サンプル $\widetilde{M}_k$ から逆過程を k ステップ繰り返すことで x に対応するメルスペクトログラム $\mathcal{M}$ を求める．

3 Diff-SVC

音声認識モデルの HuBERT，音声合成モデルの FastSpeech2，拡散モデルを用いた音声合成手法の DiffSinger を組み合わせた声質変換手法である Diff-SVC[1]に，オペラ歌唱未経験ユーザの発話音声を学習させ，推論時にはプロのオペラ歌唱音声を入力とすることで話者変換を行う．

3.1 HuBERT

HuBERT は BERT と同様の masked prediction タスクと，iterative training を組み合わせた自己教師あり学習により事前学習された音声認識モデルであり，CNN，Transformer，Projection 層の 3 つの主要部分から構成される．まず，入力音声からフレームご

*Opera-singing voice synthesis from inexperienced voice using singing voice synthesis and conversion. by Aoto Sugahara (Kobe Univ.), Soma Kishimoto (MEC Company Ltd.), Yuji Adachi (MEC Company Ltd.), Kiyoto Tai (MEC Company Ltd.), Ryoichi Takashima (Kobe Univ.), Tetsuya Takiguchi (Kobe Univ.)

との音響特徴量を抽出し，その音響特徴量の系列から k-means 法により離散ラベル系列を生成する．次に，音声を CNN エンコーダに入力することで音声表現 $X = [x_1, \ldots, x_T]$ を抽出する．この抽出した音声表現 X はランダムにマスクされ，Transformoer に入力することで文脈全体の音声表現 $Z = [z_1, \ldots, z_T]$ を得る．最後に，生成した離散ラベル系列を用いてマスクされた時刻の文脈表現 z_t がどの離散ラベルに属するかを Projection 層にて予測する (masked prediction タスク)．さらに，学習済みの Transformer エンコーダの出力を用いて再度 k-mean 法を適用することで離散ラベル系列を生成し直し，これを教師ラベルとして前述の学習を行うことで音声認識精度を向上させる (iterative training)．

Diff-SVC においては，音声から抽出した音響特徴量の系列から離散ラベル系列を抽出した後，学習済の Transformer エンコーダの出力に対して線形射影を用いて通常の HuBERT では離散的に表現されていたラベル系列をソフトスピーチユニット と呼ばれる連続値で表現するソフトコンテンツエンコーダを導入した HuBERT-soft [7] を用いる．また推論時は，音声を入力としてソフトスピーチユニットを出力する．

3.2 FastSpeech2

FastSpeech2 は，テキストを音素に変換した後に，音素を入力としてメルスペクトログラムを出力する，End-to-End の非自己回帰型音声合成モデルである．主な構成要素はエンコーダ，バリアンスアダプタ，デコーダの 3 つである．エンコーダは音素埋め込み層，自己注意機構，1 次元畳み込み層からなり，音素の離散表現を連続表現に変換する．次にバリアンスアダプタはエンコーダの出力から，音素継続長，ピッチ，エネルギーを予測し，エンコーダ出力に加える．最後にデコーダはバリアンスアダプタの出力からメルスペクトログラムを予測する．

Diff-SVC においては，前述の HuBERT-soft から出力されたソフトスピーチユニットを入力とし，バリアンスアダプタ，デコーダを通じて，中間特徴量と f_0 を出力する．

4 中高域強調ネットワーク (M-HEN)

従来研究 [3, 8] より，プロのオペラ歌唱において，口頭母音/a/においては 2.2〜3.7 kHz 帯，全音素では，2.8〜4.0kHz 帯のエネルギーがアマチュアのオペラ歌唱と比較して強く出ることが示されている．しかし，我々の先行研究 [4] より，Diff-SVC では前述したようなプロのオペラ歌唱音声の特徴を十分に保持できていないという問題があった．

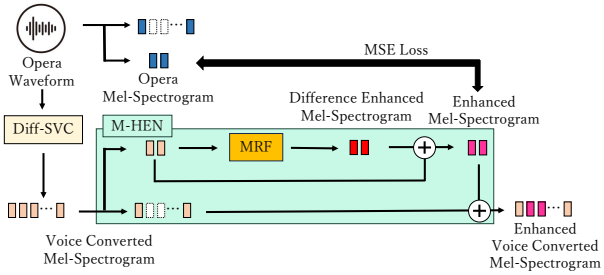

Fig. 1 Training procedure of mid-high frequency enhancement network (M-HEN).

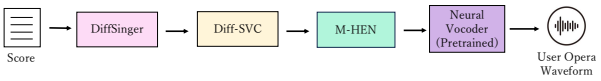

Fig. 2 Overview of the proposed the user's opera-singing synthesis system.

我々は，HiFi-GAN [9] で提案されているマルチ受容野混合層 (MRF) を用いた中高域強調ネットワーク (M-HEN) を検討した [5]．Fig. 1 に中高域強調ネットワークの学習の概要を示す．まず，ユーザ音声を用いて Diff-SVC を事前に学習しておく．この Diff-SVC にプロのオペラ歌唱音声を入力することで，ユーザ音声に変換されたオペラ歌唱音声のメルスペクトログラムが出力される．そのメルスペクトログラムを中高域帯に対応する部分とそれ以外のメルスペクトログラムに分け，その内，中高域帯に対応する部分のメルスペクトログラムを MRF に入力することで，差分強調メルスペクトログラムを得る．差分強調メルスペクトログラムと MRF の入力を加算することで，強調後のメルスペクトログラムが得られる．そして強調後メルスペクトログラムと対応するプロのオペラ歌唱音声の部分メルスペクトログラムとの MSE Loss を取ることで，MRF の学習を行う．なおこの際，事前に学習した Diff-SVC のパラメータは更新せず，MRF のパラメータのみ更新する．また推論の場合，前述の操作で得た強調部分メルスペクトログラムと最初の操作で分割した中高域以外のメルスペクトログラムを結合することで，ユーザの強調メルスペクトログラムを得る．

5 ユーザオペラ歌唱音声合成の概要と学習手順

Fig. 2 にユーザオペラ歌唱音声合成システムの概要を示す．まず，楽譜を DiffSinger に入力して，プロのオペラ歌唱を合成する．次に合成したオペラ歌唱を Diff-SVC に入力し，その後 M-HEN に入力することで，ユーザのオペラ歌唱を合成する．最後に，ニュー

[1] https://github.com/prophesier/diff-svc

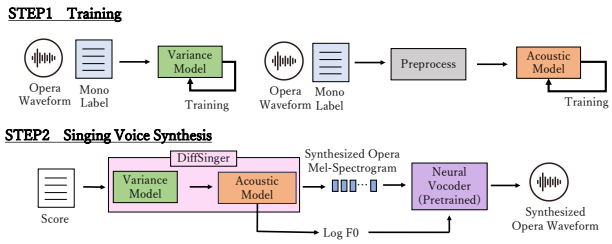

Fig. 3 Procedure of opera-singing voice synthesis using DiffSinger.

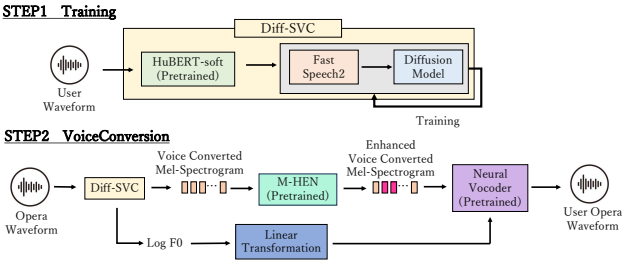

Fig. 4 Procedure of voice conversion using Diff-SVC with mid-high frequency enhancement network.

ラルボコーダに入力することで波形を生成する．

Fig. 3 に DiffSinger を用いたオペラ歌唱合成の概要を示す．まず，Step1 として，プロのオペラ歌唱音声と対応するモノフォンラベルを用いて，DiffSinger 内の Variance モデルと Acoustic モデルの学習を行う．Step2 では楽譜を入力として，オペラ歌唱音声合成を行う，まず楽譜を Variance モデルに入力することで，音素継続長などの予測を行う．次に，Variance モデルの出力を Acoustic モデルに入力することで，オペラ歌唱のメルスペクトログラムと $logf_0$ を予測する．そして，そのメルスペクトログラムと $logf_0$ を事前学習したニューラルボコーダに入力することで波形を生成する．

次に，Fig. 4 に中高域強調ネットワークを導入した Diff-SVC の声質変換の概要を示す．まず，Step1 として，変換先のユーザ音声を Diff-SVC に入力することで Diff-SVC 内の FastSpeech2 と DiffSinger の学習を行う．Step2 ではオペラ歌唱音声を入力として声質変換を行う．Step1 と同様に変換元のオペラ歌唱音声を Diff-SVC に入力することでユーザの声質に変換されたオペラ歌唱音声のメルスペクトログラムと $logf_0$ を出力される．次に，変換したメルスペクトログラムを中高域強調ネットワークに入力することで，強調メルスペクトログラムが得られる．そしてこの強調メルスペクトログラムと線形変換された $logf_0$ を事前学習したニューラルボコーダに入力することで波形を生成する．

6 評価実験

6.1 実験条件

DiffSinger においては，プロ女性歌手 1 名による日本語アカペラオペラ歌唱音声 48 曲 (約 93 分) のうち，43 曲 (約 85 分) を学習データ，5 曲 (約 8 分) をテストデータに用いた．ここで，オペラ歌唱音声のモノフォンラベルは，OpenJTalk のフロントエンド部と HMM ベースの強制アライメントによって生成した 38 種類の音素 (空白含む) からなる HTS 形式のものを使用した．Diff-SVC においては，変換先音声として JSUT コーパス [10] に収録されている女性話者 1 名の Basic5000 (約 4 時間) と JSUT-song[2] (約 25 分) を使用し，線形変換に用いる対数基本周波数の平均と分散を JSUT-song を用いて計算した．中高域強調ネットワークの入力特徴量と教師特徴量にはそれぞれ，Diff-SVC により作成したユーザのオペラ歌唱音声，DiffSinger と同様のプロ女性歌手 1 名によるオペラ歌唱音声 43 曲のメルスペクトログラムを用いた．ここで Diff-SVC において，HuBERT は Librispeech (約 960 時間) で事前学習されたモデルを用い，ニューラルボコーダは HiFiGAN [9] に NSF (neural source-filte) 構造を導入した NSF-HiFiGAN[3]を 96 時間の中国語歌唱音声で事前学習されたモデルを用いた．

本研究で用いるアカペラオペラ歌唱音声，ユーザ音声のサンプリング周波数は 44.1kHz であり量子化ビット数は 16 である．また NSF-HifiGAN の入力してはメルスペクトログラム 128 次元，基本周波数 1 次元を音響特徴量として用いた．また，MRF の入力としては 2.2kHz〜4.0kHz 帯に対応する 21 次元を Diff-SVC で作成したメルスペクトログラムから抽出して用いた．

6.2 スペクトル形状の比較実験結果

Fig. 5 に女性プロ歌手のオペラ歌唱音声 (Source)，DiffSinger で合成したオペラ歌唱音声 (DiffSinger) の 0.0kHz〜6.0kHz までのスペクトログラムを示す．DiffSinger で合成したオペラ歌唱音声はプロのオペラ歌唱と同様に，緑枠で示す 2.2kHz〜4.0kHz 帯の中高音域のエネルギーが強調されていることが確認出来る．このことから，プロのオペラらしさを維持しつつ高品質な音声が合成出来ていると考えられる．しかし，青枠で示すように，プロのオペラ歌唱と比較して細かなビブラートが再現出来ていない．これはプロのオペラ歌唱がアカペラであることから，入力に用いた楽譜と実際の歌唱に差が生じてしまったためで

[2]https://sites.google.com/site/shinnosuketakamichi/publication/jsut-song
[3]https://github.com/vtuber-plan/NSF-HiFiGAN

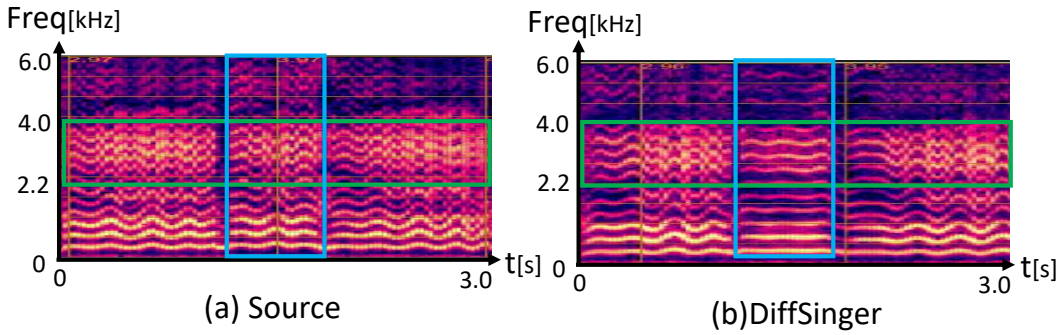

Fig. 5　Comparison of spectrograms of source voice and synthesized voice using DiffSinger.

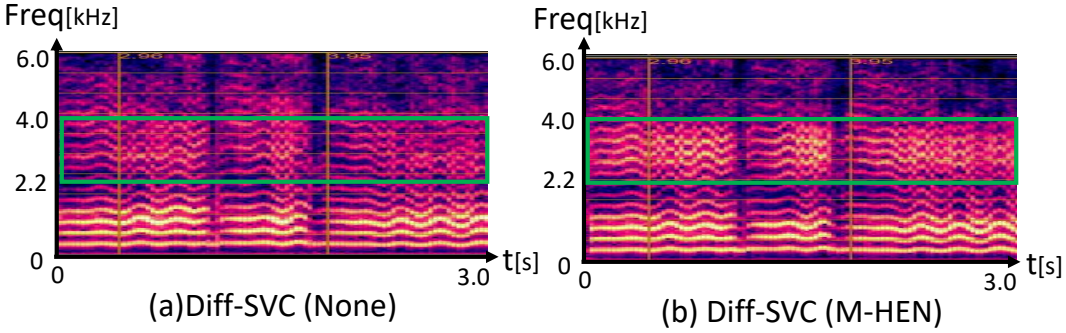

Fig. 6　Comparison of voice spectrograms converted using Diff-SVC with and without M-HEN.

あると考えられる．

次に，Fig. 6 に DiffSinger と Diff-SVC を用いて作成したオペラ歌唱音声を強調しない場合 (None)，中高域強調ネットワークで強調した場合 (M-HEN) の 0.0kHz〜6.0kHz までのスペクトログラムを示す．中高域強調ネットワークを用いない場合，先行研究 [5] と同様に，Fig. 5 の (b) で再現されていた中高音域の強調されたエネルギーが失われているが，中高域強調ネットワークを導入した場合では，プロのオペラ歌唱と同様に中高音域のエネルギーが強調されていることが分かる．このことから，Diff-SVC の入力として DiffSinger で合成したオペラ歌唱音声を用いた場合においても，中高域強調ネットワークの有効性が確認できた．

7　おわりに

本研究では，歌声合成手法の DiffSinger と声質変換手法の Diff-SVC を組み合わせたユーザオペラ歌唱合成システムを検討した．今後は主観評価実験による品質や話者性，オペラ性の評価，楽譜とアカペラ歌唱との差を推定する機構の導入に取り組む．

参考文献

[1] Johan Sundberg 他, “歌声の科学,” 東京電機大学出版局, pp. 165–177, 2007.
[2] 片平 健太 他, “母音の発音と歌唱速度の変化を考慮したアカペラオペラ歌声合成,” 音講論春, pp. 991–994, 2021.
[3] 北村 毅 他, “深層学習を用いた歌声音声の帯域強調の検討,” 音講論秋, pp. 1201-1204, 2018.
[4] 菅原 碧斗 他, “Diff-SVC を用いたオペラ歌唱音声合成,” 信学技報, pp. 30-35, 2023.
[5] 菅原 碧斗 他, “Diff-SVC を用いたオペラ歌唱音声合成における中高域強調ネットワークの検討,” 音講論秋, 2023.
[6] J. Liu *et al.*, “DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism,” AAAI, 36, pp. 11020-11028, 2022.
[7] B. van Niekerk *et al.*, “A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion,” ICASSP, 2022.
[8] S.-H. Lee *et al.*, “The Singer’s Formant and Speaker’s Ring Resonance: A Long-Term Average Spectrum Analysis,” Clinical and experimental otorhinolaryngology, 1, pp. 92-6, 2008.
[9] J. Kong *et al.*, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” Proc. NeurIPS, pp. 17022-17033, 2020.
[10] R. Sonobe *et al.*, “JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” ArXiv, 2017.