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 *

☆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っていたが，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, 4] を対象とする．また，任意の歌詞付き楽譜から歌唱音声を合成する研究や歌唱音声から歌唱音声を合成する研究は存在するが，任意の発話をしている音声から歌唱音声の合成を行う研究はほとんどない．本研究では，オペラ歌唱未経験ユーザーの発話音声から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 [5]．本研究では，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する．

2 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 [6] 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．

2.1 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ュニットを出力する．

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

* A mid-high frequency enhancement network for opera-singing voice synthesis using Diff-SVC. 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.)

2.2 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．

2.3 Diff-Singer

Diff-Singer 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る．学習においては，t 番目の拡散ステップにおけるメルスペクトログラム M_t を取り込み，t と楽譜から抽出した情報 x を用いてランダムノイズを予測する．推論においては，デノイザーを用いてノイズを予測し，t ステップ目のメルスペクトログラム M_t と予測したノイズを用いて $t-1$ ステップ目のメルスペクトログラム M_{t-1} を求める．この反復過程を $\mathcal{N}(\mathbf{0},\mathbf{I})$ からサンプリングしたガウシアン白色ノイズから，T 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．

更に，以上の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．

Diff-SVC においては，楽譜から抽出した情報 x の代わりに，FastSpeech2 から出力された中間特徴量を用いてユーザーの声質のメルスペクトログラムを合成する．

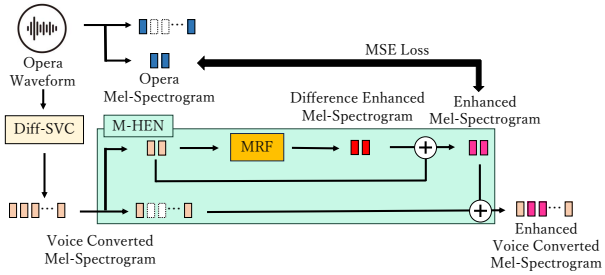

Fig. 1 Training procedure of mid-high frequency enhancement network(M-HEN).

3 中高域強調ネットワーク

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 [5] 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．そこで，以上の問題を解決するために，2.2kHz〜4.0kHz 帯の中高域を強調するネットワークの導入を検討する．

本研究では，HiFi-GAN [9] で提案されているマルチ受容野混合層 (MRF) を用いた中高域の強調ネットワーク検討する．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，ユーザー音声を用いて 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，ユーザー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，ユーザーの強調メルスペクトログラムを得る．

4 声質変換の概要と学習手順

Fig. 2 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，

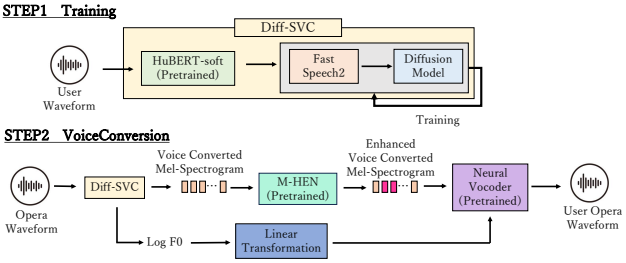

Fig. 2 Training procedure using Diff-SVC with mid-high frequency enhancement network.

変換先のユーザー音声を 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ー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ーに入力することで波形を生成する．

5 評価実験

5.1 実験条件

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ーのオペラ歌唱音声，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 43 曲 (約 85 分) のメルスペクトログラムを用いた．また，中高域強調ネットワークの有効性を確認するため，プロのオペラ歌唱音声とユーザーのオペラ歌唱音声のメルスペクトログラムから得た次元ごとの平均パワーの差分を用いてユーザーのオペラ歌唱音声の中高域のパワーを強調した音声との比較も行う．ここで変換音声の 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ーは 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．

本研究で用いる変換元音声，変換先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．

[2]https://sites.google.com/site/shinnosuketakamichi/publication/jsut-song

[3]https://github.com/vtuber-plan/NSF-HiFiGAN

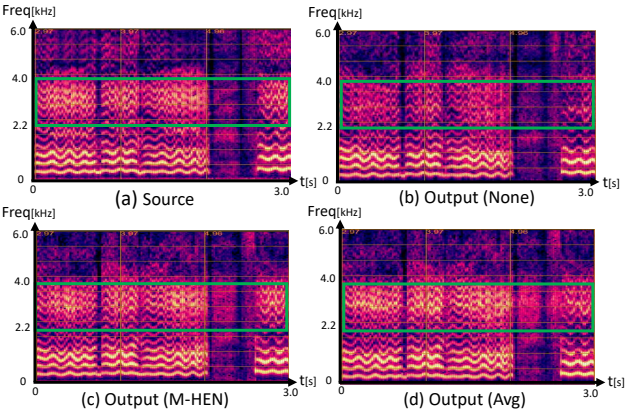

Fig. 3 Comparison of spectrograms in mid-high frequencies range.

5.2 実験結果

5.2.1 スペクトル形状の比較

Fig. 3 に女性プロ歌手のオペラ歌唱音声 (Source)，Diff-SVC で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)，平均パワー差分 (Avg) で強調で強調した場合の 0.0kHz〜6.0kHz までのスペクトログラムを示す．プロのオペラ歌唱では緑枠で示す 2.2kHz〜4.0kHz 帯の中高音域のエネルギーが従来の Diff-SVC で作成したオペラ歌唱と比較して強く出ていることが確認できる．また，中高域強調ネットワークを追加した場合では，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．このことから中高域強調ネットワークを導入することで，プロのオペラ歌唱の特徴を付与することが出来ていると考えられる．平均パワー差分で強調を行った場合も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読み取れるが，プロのオペラ歌唱音声と比較して，過度に強調されていることが分かる．

5.2.2 主観評価実験

変換音声の品質，話者性，およびオペラ性の 3 項目について，平均オピニオン指標 (MOS) による主観評価実験を実施した．品質評価においては，1 が非常に悪い音声，5 が非常に良い音声として 5 段階評価を行った．話者性評価においては，1 が変換元音声の話者性に最も近い音声，5 が変換先音声の話者性に最も近い音声として変換音声がどちらに近しいか 5 段階

評価を行った．またオペラ性の評価については，事前に判断基準となる通常歌唱とプロのオペラ歌唱を聴取した上で，1 が通常歌唱に最も近い音声，5 がプロのオペラ歌唱に最も近い音声としてどちらに近しいか 5 段階評価を行った．ここで被験者が事前に聴取した歌唱音声においては，話者性と独立した評価を行いたいため，テスト音声の話者とは異なる話者の音声を用いた．被験者は 9 人で，テストデータからランダムに抽出された 20 フレーズに対して評価を行った．

各主観評価実験の結果を Fig. 4，5，6 に示す．Fig. 4 より，中高域強調ネットワークを用いた場合と強調を行わなかった場合を比較すると，品質に有意な差が見られた．これは，拡散モデルだけでは十分に再現出来ていなかった中高音域の調波構造を再現出来たためであると考えられる．

一方，Fig. 5 より，話者性評価において，中高域強調ネットワークを用いた場合はオペラ歌唱音声を強調しない場合と比較して低いスコアを示している．これは中高域強調ネットワークの学習の教師特徴量として女性プロ歌手のオペラ歌唱音声のメルスペクトログラムを用いたため，中高域では話者性の変化が起こってしまったためだと考えられる．しかし，一般にオペラ歌唱は発声時に顎の開口度を大きくすることで第 1 フォルマントを上昇させ [1]，遠方の聴者へ歌声を聞こえやすくすることから似通った話者性になることがある．そのため今後，このような話者性の変化がオペラの特徴に起因するものなのかを検討することが必要である．

Fig. 6 より，オペラ性評価において，中高域強調ネットワークを用いた場合と平均パワー差分を用いた場合，オペラ歌唱音声を強調しない場合と比較して高いスコアを示している．また，中高域強調ネットワークを用いた場合と平均パワー差分を用いた場合には有意差は確認されなかったが，中高域強調ネットワークを用いた場合の方がわずかに高いスコアを示した．このことから中高域を強調することがオペラ性の知覚に寄与していると考えられる．また，中高域強調ネットワークを用いることで Diff-SVC で作成したユーザーのオペラ歌唱にプロのオペラ歌唱の特徴を十分に付与することができていると考えられる．

6　おわりに

本研究では，オペラ歌唱未経験ユーザーのアカペラオペラ歌唱音声合成の更なる品質向上のために，中高域強調ネットワークの導入を検討した．今後は話者性とオペラ性を独立に評価する指標の検討や，異なるオペラ歌唱の特徴の付与，歌詞の内容が変化してしまうといった問題の改善に取り組む．

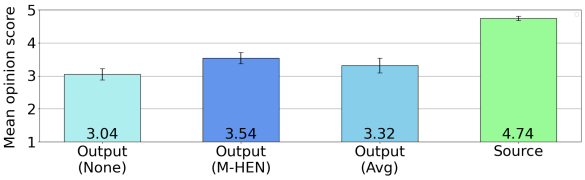

Fig. 4　MOS of quality evaluation.

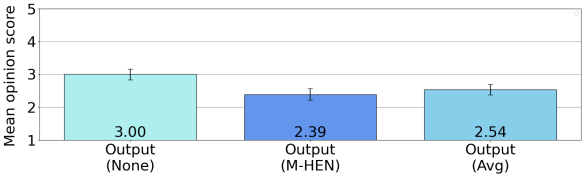

Fig. 5　MOS of speaker individuality evaluation.

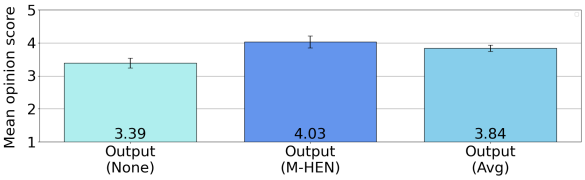

Fig. 6　MOS of operatic evaluation.

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, “歌声の母音変化を考慮した歌声合成の検討,” 音講論秋, pp. 1007–1010, 2019.

[3] 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.

[4] 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.

[5] 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.

[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.