発話音声を用いたオペラ歌唱音声合成の初期検討 *

☆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，医療分野においては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 [1]。

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 [2, 3, 4, 5] 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。さらに，声質変換で変換先音声として用いるオペラ歌唱未経験ユーザーの発話音声をオペラ歌唱音声とのパラレルデータにすることにより，変換精度の向上を検討する。

2　CycleGAN-VC2

本研究では，プロのオペラ歌唱音声を変換元音声，オペラ歌唱未経験ユーザの発話音声を変換先音声として，CycleGAN-VC2 [6] により話者変換を行う。CycleGAN-VC2 は画像変換分野において用いられている CycleGAN [7] を声質変換分野に拡張した CycleGAN-VC [8] の改良モデルである。CycleGAN は一対一の画像生成ネットワークであり，ノンパラレルデータを用いて画像中の人物を男性から女性に変換するというように，一つの種類の画像集合から別の種類の画像集合に変換するといったドメイン変換が可能である。CycleGAN-VC2 は一対一の声質変換手法であり，CycleGAN と同様にノンパラレルデータでの声質変換が可能である，5 分程度の少量データでの声質変換が可能である，一度の学習で変換元から変換先と変換先から変換元への双方向の声質変換が可能である，という特徴がある。

3　パラレル歌詞朗読音声の作成

異性間の声質変換よりも同性間の声質変換の方が精度が高いように，一般に変換元と変換先の音声特徴量が類似している方が声質変換モデルの学習がしやすい。一方本研究では，変換元がオペラ歌唱音声，変換先が通常発話音声であるため，特徴量間のギャップが大きい。そこで，オペラ歌唱音声の歌詞を朗読したものを通常発話音声としてパラレルデータとすることで，特徴量間のギャップを減らすことを検討する。

パラレル歌詞朗読音声を作成する方法として，以下の二種類が挙げられる。一つ目はオペラ歌唱未経験ユーザーの歌詞朗読音声に対して DTW (Dynamic Time Warping) [10] を用いてオペラ

* Initial study of opera-singing voice synthesis using speaking voice. 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.)

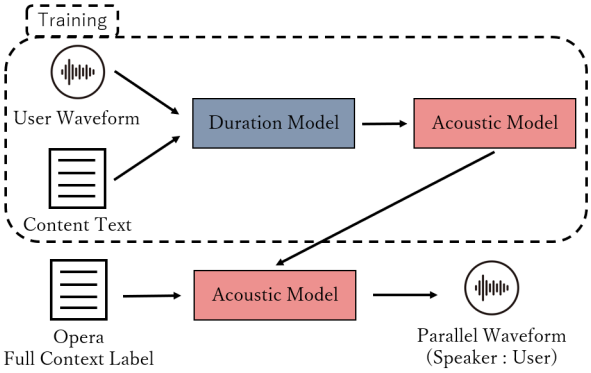


Fig. 1　Procedure for creating parallel data using TTS model.

歌唱音声とのアライメントを一致させる方法である。この方法では，事前にオペラ歌唱未経験ユーザーの歌詞朗読音声に用意する必要がある。また，Sangeon [11] らの研究で，歌唱音声を用いてパラレルデータを作成する際，ビブラートやピッチの変動が大きい箇所でアライメントのエラーが起きることが示されている。また，歌詞朗読音声だけでなくオペラ歌唱音声も DTW による引き延ばしが生じるため，学習データの品質への悪影響が懸念される。

二つ目は Fig. 1 に示すようにユーザの発話音声により学習された TTS (Text-to-Speech) モデルを用いて歌詞朗読音声を合成する方法である。合成の際，音素継続長モデルの出力の代わりに，オペラ歌唱音声から求めた音素継続長を用いることで，DTW をせずともアライメントが一致したパラレル歌詞朗読音声が作成可能である。この方法では変換先音声が TTS による合成音声であるため，TTS による学習データの品質劣化が懸念される一方，DTW を使用しないため，前述のような問題は起こらない。

予備実験として，両者の方法で生成されたパラレル歌詞朗読音声を比較した結果，TTS を使用した方が高品質な音声となっていたため，本研究では TTS による方法を採用する。

4　声質変換の概要と学習手順

Fig. 2 に本研究の声質変換の概要を示す。まず，Step1 として，WORLD [9] を用いてオペラ歌唱音声とユーザー発話音声からそれぞれ抽出したメルケプストラムを用いて CycleGAN-VC2 の学習を行う。次に Step2 ではオペラ歌唱音声を入力として声質変換を行う。Step1 と同様にオペラ歌唱音声から WORLD を用いてメルケプストラム，基本周波数，非周期性指標を抽出し，メルケプストラムに対しては Step1 で学習した CycleGAN-VC2 を用いて声質をユーザーに変換する。基本周波数に対しては線形変換を行い，非周期性指標に対してはオペラ歌唱音声のものをそのまま用いる。変換した特徴量に対して WORLD ボコーダーにより波形を生成する。

5　評価実験

5.1　実験条件

変換元音声として女性歌手 1 名による日本語アカペラオペラ歌唱音声 48 曲 (約 93 分) を収録した。この 48 曲のうち，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。変換先音声として，ATR 日本語データベース [12] に収録されている男性話者 1 名 (MHT) と女性話者 1 名 (FTK) の音素バランス文 503 文を使用した。各話者につき 503 文を用いて TTS モデルを学習し，TTS モデルにより男女それぞれ 43 曲分のパラレル歌詞朗読音声の学習データを生成した。またパラレル歌詞朗読音声を用いる有効性を確認するため，TTS を使用せずに音素バランス文 503 文のうち 450 文 (約 40 分) をそのまま学習データとして使用した場合とも比較した。本実験ではこのデータを非パラレル音声と呼ぶことにする。パラレル歌詞朗読音声作成時に TTS モデルの入力として用いるオペラ歌唱音声のフルコンテキストラベルは，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。

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 16kHz，量子化ビット数は 16 である。TTS モデルで使用する音響特徴量は，メルケプストラム 60 次元，帯域非周期性指標，対数基本周波数，有声/無声フラグで構成され，有声/無声フラグ以外に関しては 2 次までの動的特徴量を含んだ計 187 次元からなり，次元ごとに平均 0 分散 1 となるよう標準化を行った。また CycleGAN-VC2 では，メルケプストラム 36 次元，対数基本周波数 1 次元，非周期性指標 1 次元を音響特徴量として用いた。

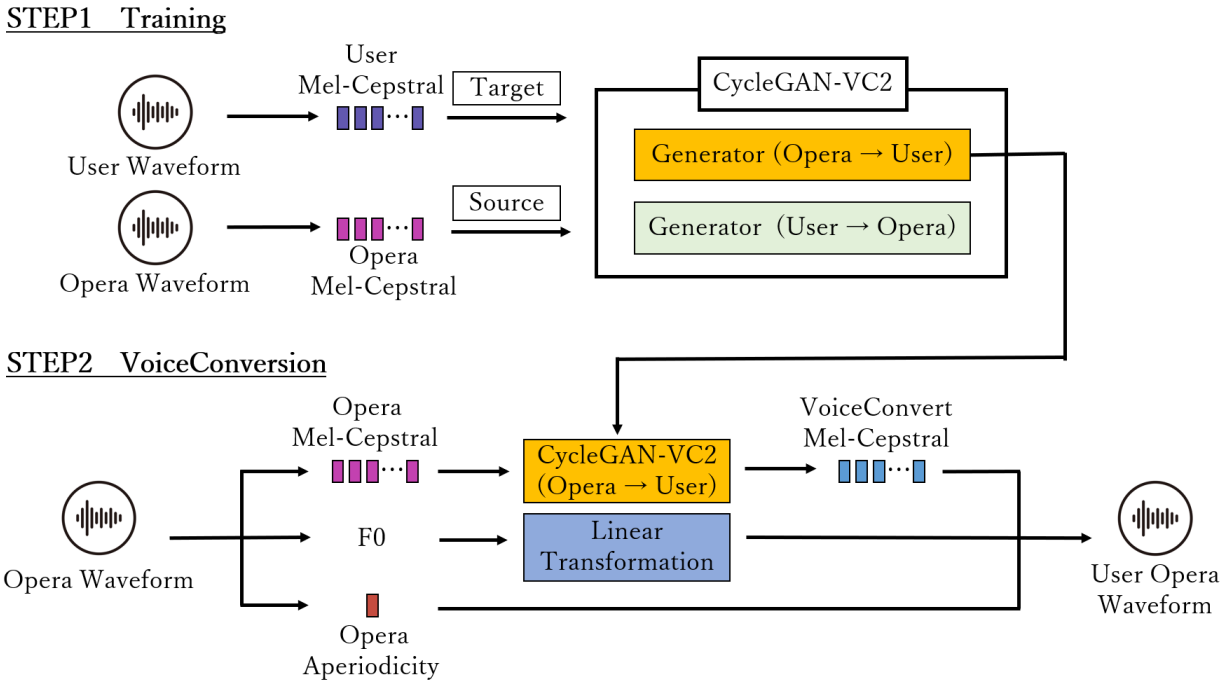


Fig. 2 Training procedure using CycleGAN-VC2-based voice conversion.

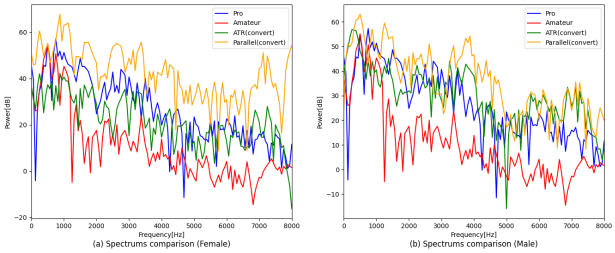


Fig. 3 Comparison of spectrums in opera singing voices.

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 (Pro), 女性アマチュア歌手のオペラ歌唱音声 (Amateur), そして非パラレル音声 (ATR), パラレル歌詞朗読音声 (Parallel) それぞれで学習された CycleGAN-VC2 により変換されたオペラ歌唱音声の平均対数パワースペクトルを示す。平均対数パワースペクトルはテストデータ 5 曲のうち 3 曲分の音声に対するフレーム毎の対数パワースペクトルを全フレームで平均したものである。従来研究 [5] と同様に, プロのオペラ歌唱では 3,000Hz から 4,000Hz 帯の中高音域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ていることが確認できる。また変換音声と比較すると, パラレル歌詞朗読音声で声質変換した音声はどちらの話者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く出ていることが確認でき, このことからパラレル歌詞朗読音声で声質変換することでプロのオペラ歌唱の特徴を保持できると考えられる。

5.2.2 主観評価実験

主観評価指標として変換音声の品質と話者性の評価のために平均オピニオン指標 (MOS) を用いた。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。また変換音声について, 全フレーズのうちいくつのフレーズの歌詞内容が変化していると感じたかを評価実験も行った。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 22 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, Fig. 5 より, パラレル歌詞朗読音声で声質変換した音声は品質, 歌詞内容変化率において男女共に ATR 音素バランス文で声質変換した音声よりも高いスコアを示したが, 変換元音声と比較すると低いスコアを示した。これは前述の通り, 通常発話音声としてパラレル歌詞朗読音声を用いることで, ATR 音素バランス文を用いる場合よりもオペラ歌唱音声との特徴量間のギャップを小さくなったため精度向上し言語情報が保持されたが, まだ特徴量間のギャップが大き

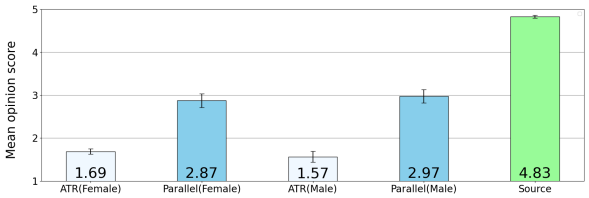


Fig. 4　MOS on quality evaluation.

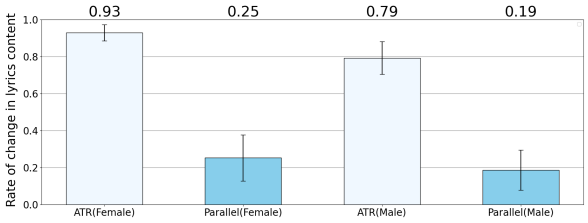
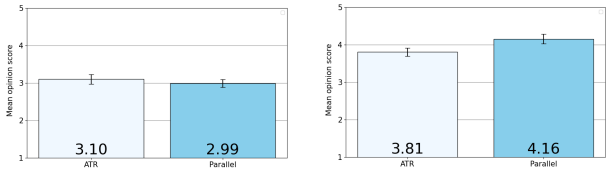


Fig. 5　Rate of change in lyrics content.



(a) Female　(b) Male

Fig. 6　MOS on speaker-similarity evaluation.

いため変換音声の品質は変換元音声と比較して低くなり，歌詞内容変化率が大きくなったと考えられる。

Fig. 6 より，話者性評価において男性話者の場合パラレル歌詞朗読音声で声質変換した音声の方が高いスコアであった。一方，女性話者の場合 ATR 音素バランス文で声質変換した音声の方が高いスコアであったが t 検定による有意差は認められなかった。また，女性話者よりも男性話者の方が高いスコアとなっている。これは同性間の声質変換と比較して異性間の声質変換の方が変換元音声からの変化が大きく判別し易いためであると考えられる。

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ために，プロのアカペラオペラ歌唱音声の声質をオペラ歌唱未経験ユーザーの声質に変換する手法を検討した。また，声質変換で変換先音声として用いるオペラ歌唱未経験ユーザーの発話音声をアカペラオペラ歌唱音声とのパラレルデータにすることによって変換精度の向上が確認できた。今後はよりオペラ歌唱音声の特徴を保持した変換や，歌詞の内容が変化してしまう問題の改善に取り組む。

参考文献

[1] 和田 蒼太 他, “歌声合成におけるニューラルボコーダの比較検討,” IEICE Report, SP2019-42(2019-12), pp. 85-90, 2019.

[2] Johan Sundberg 他, “歌声の科学,” 東京電機大学出版局, pp. 165–177, 2007.

[3] 片平 健太 他, “歌声の母音変化を考慮した歌声合成の検討,” 日本音響学会秋季研究発表会講演論文集, pp. 1007–1010, 2019.

[4] 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,” 日本音響学会春季研究発表会講演論文集, pp. 991–994, 2021.

[5] 北村 毅 他, “深層学習を用いた歌声音声の帯域強調の検討,” 日本音響学会秋季研究発表会講演論文集, pp. 1201-1204, 2018.

[6] Takuhiro Kaneko *et al.*, “CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion,” ICASSP, 2019.

[7] Jun-Yan Zhu *et al.*, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” ICCV, 2017.

[8] Takuhiro Kaneko and Hirokazu Kameoka “Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks,” EUSIPCO, 2018.

[9] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, “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.

[10] Pavel Senin, “Dynamic time warping algorithm review,” Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, vol. 855, pp. 1-23, 2008.

[11] Sangeon Yong and Juhan Nam, “Singing expression transfer from one voice to another for a given song,” ICASSP, 2018.

[12] Akira Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, Vol. 9, No. 4, pp. 357-363, 1990.