読み上げ音声を用いたCycleGAN-VC2 によるオペラ歌唱音声合成

菅原, 碧斗 ; 岸本, 宗真 ; 足立, 優司 ; 田井, 清登 ; 高島, 遼一 ; 滝口, 哲也

(Citation)
神戸大学都市安全研究センター研究報告,27:51-56

(Issue Date)
2023-03

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/0100482519

(URL)
https://hdl.handle.net/20.500.14094/0100482519

神戸大学都市安全研究センター
研究報告，第27号，令和５年３月

読み上げ音声を用いた CycleGAN-VC2 によるオペラ歌唱音声合成

Opera singing voice synthesis based on CycleGAN-VC2 using speaking voice

菅原 碧斗[1)]
Aoto Sugahara
岸本 宗真[2)]
Soma Kishimoto
足立 優司[2)]
Yuji Adachi
田井 清登[2)]
Kiyoto Tai
高島 遼一[3)]
Ryoichi Takashima
滝口 哲也[4)]
Tetsuya Takiguchi

概要：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。また、近年ではより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。本研究では、オペラ歌唱未経験ユーザーの読み上げ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。一般に声質変換では、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読み上げた音声を変換先音声として変換元音声とのパラレルデータとすることにより、特徴量間のギャップを減らすことを提案する。主観評価実験より、提案手法を用いることで品質と歌詞了解度の向上が示されたため、パラレルデータを用いることの有効性が確認できた。

キーワード：オペラ歌唱、声質変換、テキスト音声合成、歌声合成

1. はじめに

歌声合成技術は一般利用者でも利用可能なツールとして広く知られており、CPU やスマートフォン上で手軽に利用可能なモデルやソフトウェアが提供されている。また近年では声質変換技術に注目が集まっており、カラオケやライブ配信サービスといった娯楽分野において、リアルタイムに声質変換が可能なツールも開発されている。しかしこのような声質変換ツールにおいて、変換先音声の声質と入力されたユーザーの声質が大きく離れていると変換音声の品質や変換精度の低下が見られる。また歌声を用いる場合、変換音声に変換元の歌唱音声の歌唱特徴を十分に反映させることができないといった問題も存在する。そのため、本研究では一般の歌唱と比較して歌唱特徴が異なるオペラ歌唱音声に着目し、オペラ歌唱未経験ユーザーの読み上げ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ことで、変換音声の表現力拡大を目指す。

また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が、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読み上げた音声を変換先音声として変換元音声とのパラレルデータにすることによ

り、特徴量間のギャップを減らすことを提案する。

2. パラレル歌詞朗読音声を用いた声質変換

本研究では、変換元音声としてプロのオペラ歌唱音声、変換先音声としてユーザーの読み上げ音声を用いて声質変換モデル(VC モデル)を学習し、学習した VC モデルに変換元音声であるプロのオペラ歌唱音声を入力することで、ユーザーの声質のオペラ歌唱音声を合成することを目的とする。また、変換先音声をプロのオペラ歌唱音声と同一歌詞内容を読み上げた音声である、パラレル歌詞朗読音声を作成し、それを VC モデルの学習に用いることで声質変換音声の品質向上を目指す。図 1 に本研究におけるパラレル歌詞朗読音声の作成と声質変換モデルの学習手順を示す。

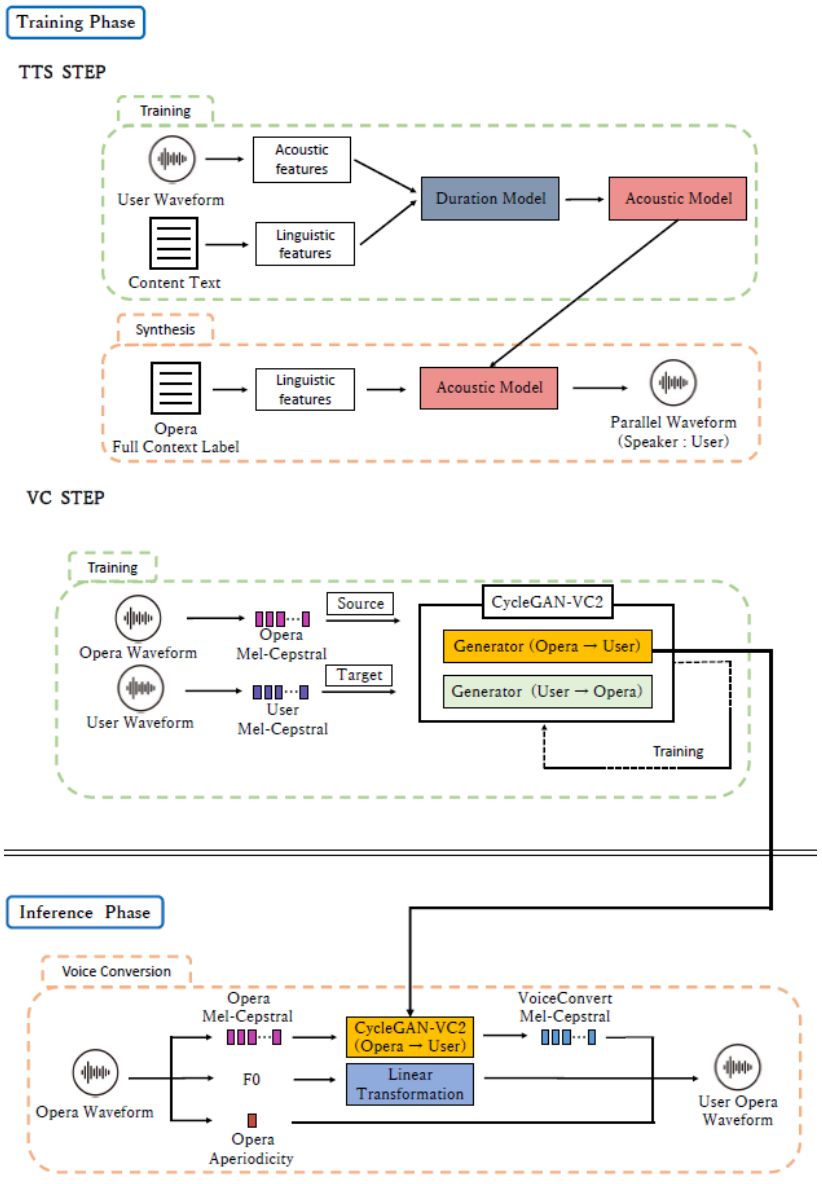


図 1.　パラレル歌詞朗読音声の作成と声質変換モデルの学習手順

学習フェーズでは TTS を用いてパラレル歌詞朗読音声を生成するステップ（TTS ステップ)と、パラレル歌詞朗読音声とオペラ歌唱音声を用いて声質変換モデル(VC モデル)を学習するステップ(VC ステップ)で構成されている。合成音声を生成するための TTS モデルとして、DNN に基づいた統計的パラメトリック音声合成の構造を使用する。また、音声信号から声質変換を行う VC モデルとして、CycleGAN-VC2[1)]を使用する。

TTS ステップでは、まず従来のテキスト音声合成と同様に、ユーザーの読み上げ音声とその音声に対応す

るフルコンテキストラベルからそれぞれ音響特徴量と言語特徴量を抽出する。抽出した特徴量を用いて音素継続長を推定する継続長モデルと音響特徴量を推定する音響モデルの学習を行う。次にプロのオペラ歌唱のフルコンテキストラベルから抽出した言語特徴量を入力として、学習済の音響モデルのみを用いて音響特徴量を推定する。最後に推定した音響特徴量をもとに声質がユーザーのパラレル歌詞朗読音声を作成する。VC ステップでは、変換元音声として、プロのオペラ歌唱音声と変換先音声として、TTS フェーズで作成した声質がユーザーのパラレル歌詞朗読音声を用いる。まず、変換元音声と変換先音声から抽出したメルケプストラムを用いて、CycleGAN-VC2 の学習を行う。学習した 2 つの Generator のうち変換元音声であるオペラ歌唱から変換先音声であるユーザーのパラレル歌詞朗読音声への声質変換を行う $G_{Opera \rightarrow User}$ を以降の声質変換で用いる。

推論フェーズでは、変換元音声から抽出したメルケプストラム、F0、非周期性指標を用いて、一般的な統計的声質変換と同様に声質変換を行う。ここでメルケプストラムの声質変換には $G_{Opera \rightarrow User}$ を用い、F0 は線形変換を行う。また、非周期性指標はオペラ歌唱音声のものをそのまま用いる。そして変換を行った 3 つの音響特徴量に基づいて合成音声を作成する。

3. 評価実験

(1) 使用するデータ

本実験では、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 [2]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。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。

(2) 分析条件及びモデル設定

TTS モデルで用いる音響特徴量は、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、学習時に次元ごとに平均 0 分散 1 となるよう標準化を行った。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、Open JTalk の Python ライブラリである pyopenjtalk 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。音響モデルに関しては中間層が 3 層の全結合で構成され、次元数は 975 次元(フレームレベルの場合はフレーム特徴量が追加されて 979 次元) であり、次元ごとに最小が 0、最大が 1 となるように min-max 正規化を行った。

声質変換モデルに入力する音響特徴量としてメルケプストラム 36 次元、対数基本周波数 1 次元、非周期性指標 1 次元を用いた。TTS モデルと声質変換モデルの両方において、音声波形の生成や、音響特徴量の取得などには WORLD を使用した。

(3) 実験結果

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が歌詞通りの歌声に感じるか(歌詞了解度と呼ぶこととする)についての評価実験も行った。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。

品質、歌詞了解度、話者性についての主観評価結果をそれぞれ図 2、図 3、図 5 に示す。図 2、図 3 より、パラレル歌詞朗読音声で声質変換した音声は品質、歌詞了解度において男女共に 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いため変換音声の品質と歌詞了解度は変換元音声と比較して低くなったと考えられる。

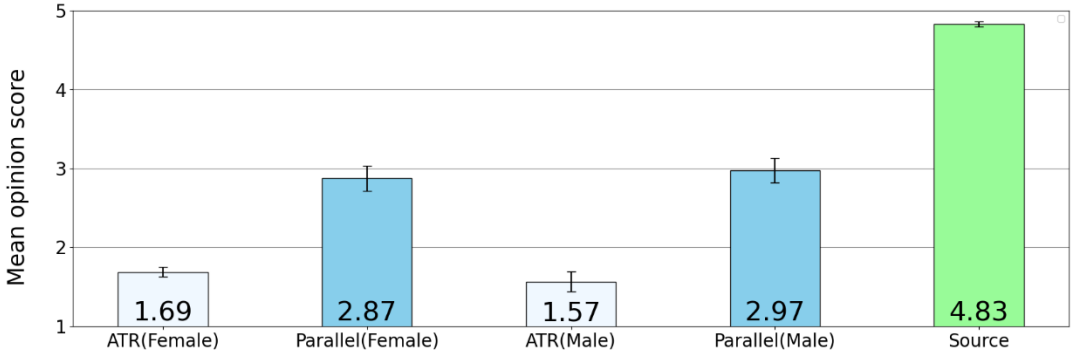


図 2.　MOS 評価(品質)

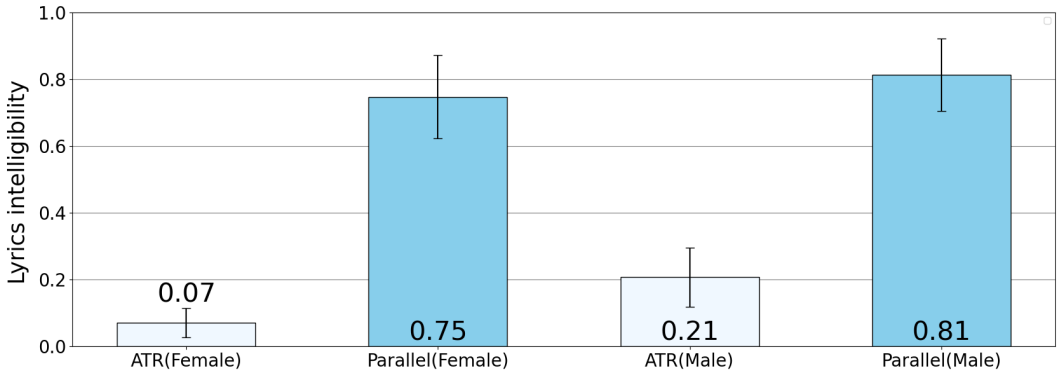


図 3.　歌詞了解度の実験結果

また、例として TTS モデルで合成したパラレル歌詞朗読音声のスペクトログラムを図 4 に示す。音素/u/に着目すると、本来は 1.00 秒から 1.25 秒間の周波数成分のような分布になるはずだが、1.50 秒から 1.75 秒間の周波数成分を見ると特に低周波数成分が小さくなっていることが読み取れる。ここで前述の通り母音の知覚には第一フォルマント周波数と第二フォルマント周波数といった低周波成分が大きな影響を与えているため、1.50 秒から 1.75 秒間では母音の知覚が難しく自然性が低下している。以上の例のように TTS モデルで合成したパラレル歌詞朗読音声において長母音や語尾の母音の周波数成分が小さくなっており、ATR 音素バランス文と比較して自然性が低下している。また通常の統計的音声合成手法を用いて、ノンパラレルな歌詞朗読音声を合成した際も同様に長母音や語尾の母音の部分の周波数成分が小さくなっていることが確認できる。ここで予備実験として、収録したノンパラレル歌詞朗読音声から作成された、一部分の自然性が低下したパラレル歌詞朗読音声と自然性が元の歌詞朗読音声と同程度のパラレル歌詞朗読音声をそれぞれ用いて、オペラ歌唱音声との声質変換実験を行った。この予備実験では、自然性の低下している部分において、品質の低下と歌詞了解度の低下が確認された。このことから、TTS モデルで作成されたパラレル歌詞朗読音声の自然性を向上させることで歌詞了解度のさらなる向上を実現できると考えられる。

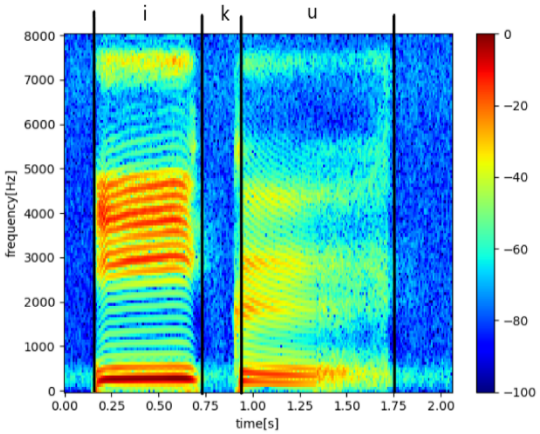
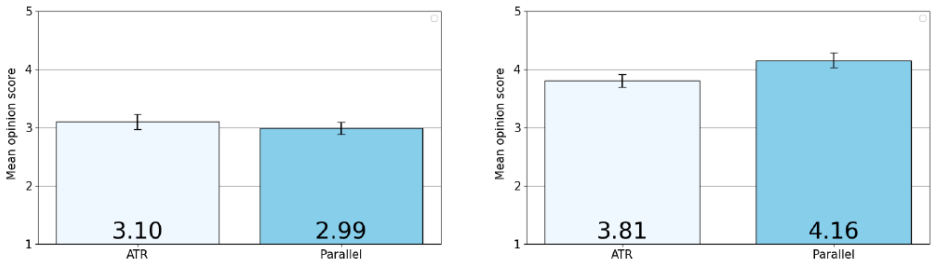


図 4.　自然性の低いパラレル歌詞朗読音声のスペクトログラム例

図 5 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。



(a) Female　　(b) Male

図 5.　MOS 評価(話者性)

4. まとめ

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、音韻の変化により歌詞通りに聞こえない音声が生成されてしまう問題の改善に取り組む。

参考文献

1） Takuhiro Kaneko et al., “CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion,” *ICASSP* (2019).

2） A. 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.

筆者： 1）工学部情報知能工学科、学生； 2）メック株式会社； 3 ）都市安全研究センター、准教授； 4）都市安全研究センター、教授

Opera Singing Voice Synthesis Based on CycleGAN-VC2 Using Speaking Voice

Aoto Sugahara
Soma Kishimoto
Yuji Adachi
Kiyoto Tai
Ryoichi Takashima
Tetsuya Takiguchi

Abstract

Singing voice synthesis technology is widely used in the entertainment field and the medical field, it has attracted attention as a method to reproduce the singing voices of the deceased or patients who have lost their voices. In recent years, research is also conducted on synthesizing singing voices with more human-like expressions. The purpose of this study is to realize a system that can synthesize operatic singing voices from the voices of inexperienced opera singers. In general, it is easier to train voice conversion models when the source and target voice features are similar. In this study, however, the source speech is opera singing and the target speech is normal speech, so the gap between the features is large. Therefore, we propose to reduce the gap between the features by using the source voice and the target voice as parallel data, i.e., the voice from which the lyrics of the opera song are read out. Subjective evaluation experiments showed that the proposed method improved the quality and comprehension of the lyrics, confirming the effectiveness of using parallel data.