声質変換による合成音声話者の拡張

髙村 健也* 原田 楓* 小髙 知宏** 黒岩 丈介** 白井 治彦*** 諏訪 いずみ****

Increasing Variation of Synthetic Speech Speakers by Voice Conversion Technology

Kenya TAKAMURA*, Kaede HARADA*, Tomohiro ODAKA, Jousuke KUROIWA**, Haruhiko SHIRAI***, Izumi SUWA*****

(Received September 30, 2022)

In this paper, we used voice conversion techniques on synthetic speech to increase the variation of synthetic speech speakers. We performed voice conversion using 'CycleGAN-VC2' which is the non-parallel voice conversion model that does not impose significant restrictions on training data and evaluated output voices. The data set for the experiment was a synthetic voice with four different speakers (two female and two male) from the 'Google Cloud Text-to-Speech' service for source voices. For target voices, we prepared narration voices by the one male speaker.

As a result of the experiment, the voice conversion process was properly performed except for one type of female speaker, and it was possible to change the speaker. The problem with this method was that the conversion process loses the human-like voice quality found in source voices. The conclusion of our method is that it is useful for increasing variation of synthetic voice speakers if the transformation process can be made to sufficiently satisfy quality preservation.

***Key Words*:** Voice Conversion, Synthetic Speech, CycleGAN

1. 緒言

コンピュータによって音声を生成する音声合成技術は人工知能の根幹をなすディープニューラルネットワークモデルにより発展しており，最新の合成音声は人が話す音声に近い品質であることが示されている[1]. このような合成音声は, 電化製品を始めとした様々な物をインターネットに接続して活用する現代社会において，コンピュータからの動的な応答を人に伝える役割を持つ．その最たる例が，スマートフォンやスマート家電に搭載されている音声アシスタントであり，具体的には Google LLC による「Google アシスタント」や Amazon.com, Inc.による「Amazon Alexa」が挙げられる．このような合成音声の話者に着目すると，出力言語を日本語と設定した場合に選択可能な話者は女性 1 種類・男性 1 種類となっていることが多く，利用者はせいぜい 2 種類程度の話者を選択することに限られている現状がある．もしも合成音声話者のバリエーションが豊富であるならば，利用者はそれぞれの好みに合わせた話者を自由に選択し，コンピュータとのやり取りを自由に楽しむことができると考えられる．

そこで，本研究では合成音声に対して声質変換技術を用いることで，合成音声話者のバリエーションを拡張することを試みた. ここで声質変換技術とは，声の高さや抑揚などの音響的特徴を変換し，変換前の音声とは異なった声質を持つ音声を生成するものである[2].

* 大学院工学研究科知識社会基礎工学専攻
** 知能システム工学講座
*** 工学部技術部
**** 仁愛女子短期大学
* Fundamental Engineering for Knowledge-Based Society, Graduate School of Engineering
** Department of Human and Artificial Intelligent Systems
*** Technical Division
**** Jin-ai Women's College

2. 合成音声に対する声質変換の適用

2.1 声質変換の手法

声質変換を行う方法として，パラレル変換とノンパラレル変換の 2 種類がある．これらの方法は，声質変換モデルを構築するために必要な，変換の元とする音声データ（Source 音声）と変換によって再現したい声質の音声データ（Target 音声）の関係性によって分類される．パラレル変換手法では 2 種の音声データに対して，同一の文章を読み上げさせて発話内容を合わせた後に，音素レベルで発話タイミングを同期させることが必要となる．同期の精度を良い状態にするには手動での綿密な調整が不可欠であり，パラレル変換はデータセットの用意が困難である．一方でノンパラレル変換手法では，2 種の音声データに対して制約を課さないため複雑な前処理が不要となる．

合成音声話者の拡張を目的とする本研究においては，合成音声による Source 音声と再現目標である Target 音声の組み合わせは様々であり，複数種類の声質変換モデルを生成するための手法はより簡単な方が好ましい．そのため，本研究ではノンパラレル変換の手法を採用する．

2.2 使用する声質変換モデル

本研究ではノンパラレルな声質変換手法として，「CycleGAN-VC2」声質変換モデルを使用する[3]．このモデルは深層学習モデルである CycleGAN をベースとしている[4]．CycleGAN のアーキテクチャを図 1 に示す．CycleGAN は異なる 2 つのドメイン間のマッピングを行うモデルであり，4 種のニューラルネットワーク（Generator 2 種，Discriminator 2 種）から構成される．Generator は一方のドメインに属するデータをもう一方のドメインに属するデータに似せるようにマッピングを行う役割を持ち，Discriminator は Generator が生成したデータとドメインに属するデータとを分類する役割を持つ．CycleGAN では，これらのニューラルネットワークを適切に制御するために以下の 3 種類の損失を用いて学習を行う．

- Adversarial Loss
- Cycle Consistency Loss
- Identity Mapping Loss

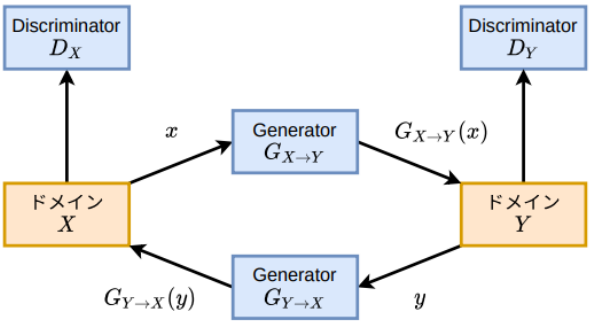

図 1　CycleGAN のアーキテクチャ

Adversarial Loss は，Generator と Discriminator が敵対するように機能させるための損失であり，Generator のマッピングにおけるデータの劣化を軽減する役割を持つ．Cycle Consistency Loss は，一方向の Generator の出力を逆方向の Generator の入力とし，最終的な出力を最初の入力と比較する損失であり，データの構造情報の保持を促進する役割を持つ．Identity Mapping Loss は，あるドメインに属するデータを，所属ドメインが変わらないように Generator でマッピングし，その出力を元の入力を比較する損失であり，声質変換タスクでは音声データに含まれる言語情報の保持を促進する役割を持つ[5]．

2.3 声質変換の流れ

学習済みの CycleGAN-VC2 モデルによる声質変換の流れを図 2 に示す．まず，音声合成分析器 WORLD（D4C edition）を使用して変換の対象とする音声から以下の 3 つの特徴量を抽出する[6] [7]．

- メルケプストラム（MCEP）
- 基本周波数
- 非周期性指標

MCEP は音の音色，基本周波数は音の高低，非周期性指標は有声音における雑音成分に対応する．

次に抽出した特徴量に対してパラメータ変換を行う．変換の対象とするのは MCEP と基本周波数であり，非周期性指標には何の処理も行わない．MCEP には，変換モデルを学習する際に使用するトレーニングデータから得られる統計情報を元にした線形的な変換処理と，Generator によるマッピング処理を組み合わせて適用する．基本周波数には，自然対数をとった後に同様の線形変換処理のみを適用する．ここで扱う線形変換は次のように表される．

$$f_{conerted} = \frac{f - \mu_s}{\sigma_s} * \sigma_t + \mu_t \quad (1)$$

ここで，fはある特徴量，μ_s, σ_sは Source 音声話者のトレーニングデータから得られる特徴量fの平均と標準偏差，μ_t, σ_tは Target 音声話者のトレーニングデータから得られる特徴量fの平均と標準偏差を意味している．

最後に，パラメータ変換処理を行った各特徴量を WORLD によって合成し，声質変換を適用した音声を出力する

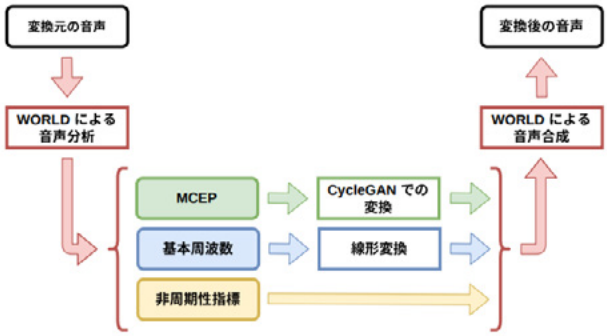

図 2　声質変換の流れ

3.　実験

3.1　音声データの取得と処理

実験で使用する音声データの種類は Target 音声と Source 音声の 2 種類であり， Source 音声に対して変換処理を行うことで，まるで Target 音声の話者が話しているような音声を生成する．

Target 音声を取得するために，書籍をナレーターや西友が朗読した音声データを配信するサービス「audiobook.jp」を利用した[8]．このような音声には目的とする音声以外にノイズや効果音がなどの音が含まれていないためである．Target 音声の話者としてプロの男性声優を設定し，この男性によるナレーション音声のうち，以下の 3 つの作品を購入した．

- 走れメロス（太宰治 著）
- 杜子春（芥川龍之介 著）
- 子供おばさん（山本文緒 著）

これらの音声は，拡張子が mp3，チャンネル数がステレオ，サンプリング周波数が 44.1[kHz]である．

取得した Target 音声に対して行う処理は以下の 3 つである．

- 有音区間の抽出
- 拡張子の変更
- チャンネル数，サンプリング周波数の変更

有音区間の抽出では，単語や文章の間に挟まれる無音の区間を排除する.「走れメロス」以外の音声では，音量の閾値を-24.0[dB]，無音部の最大持続時間を 0.80[s]とするなど，一定条件の設定下においてプログラム的に抽出した.「走れメロス」の音声では，作品の原文を参照しながら，文章の構造に従い手動で抽出を行った．拡張子の変更では，音声解析の際によく利用される wav 形式へと変更し，量子化ビット数を 16[bit]とする．チャンネル数，サンプリング周波数の変更では，後述する Source 音声のものと同期させるために，ステレオからモノラル，44.1[kHz]から 24.0[kHz]へと変更する．

Source 音声の取得では，本研究の目的を考慮して音声合成技術によって生成された音声データでなければならない．そのため，任意の自然言語によるテキストデータを入力することで，そのテキストの読み上げ音声を出力する音声合成サービス「Google Cloud Text-to-Speech」を利用した[9]．このサービスにおけるテキストを読み上げる話者は，高品質な日本語を話す女性 2 種，男性 2 種の 4 種類とした．

- ja-JP-Wavenet-A （女性）
- ja-JP-Wavenet-B （女性）
- ja-JP-Wavenet-C （男性）
- ja-JP-Wavenet-D （男性）

本論文では，これらの話者を「Source A」あるいは「Source 話者 A」といったように表記する．入力とするテキストについては，日本語のテキストとその読み上げ音声のセットである「JUST コーパス（ver1.1）」に含まれる「basic5000」と，Target 音声の処理において分割した「走れメロス」の音声それぞれと発話内容が一致するようにしたテキストである[10]．取得した Source 音声は，拡張子が wav，量子化ビット数が 16[bit]，チャンネル数がモノラル，サンプリング周波数が 24.0[kHz]である．

3.2　データセットの設定

取得した音声データを基にして，声質変換モデルを学習するために利用するトレーニングデータと，学習後の変換モデル性能を検証するために利用するテストデータを構築する．

まず Target 音声について,「杜子春」と「子供おばさん」をトレーニングデータ,「走れメロス」をテストデータとして利用する．Source 音声については,「basic5000」をトレーニングデータ,「走れメロス」をテストデータとして利用する．つまり，Target と Source 音声のテストデータにおいては，発話内容が同期している形式となっている．

次に，トレーニングデータの詳細な設定として，再生時間に着目してフィルタリングを行う．Target 音声のトレーニングデータは一定の条件下で有音区間を抽出したものであり，学習に適さないほど再生時間が短いものが存在しているためである．再生時間が 3.0[s]未満のデータをフィルタリングすることで，Target 音声の総ファイル数は 502 個となった．Source 音声のデータに対しても同様のフィルタリングを行い，残ったものから 502 個のファイルをランダムに選択した．トレーニングデータの詳細を表 1 に示す.

最後に，テストデータの詳細な設定として，再生時間とセリフ文に対するフィルタリングを行い，各話者間で発話内容が完全に同期するように除去を行う. 具体的には再生時間のフィルタリングは 1.5[s]未満とし，Target と Source を比較して読み方が大きく異なっているセリフ文を除く．これに加え，ある発話内容の音声が除去された際にはその発話内容を持つ音声を全体から除外する．これらの結果，各テストデータのファイル数は 266 個となった．テストデータの詳細を表 2 に示す.

表 1　トレーニングデータの詳細

話者	ファイル数	総再生時間 [s] (平均±標準偏差)
Target	502	2,850 (5.7±2.7)
Source A	502	3,107 (6.2±2.6)
Source B	502	2,892 (5.8±2.4)
Source C	502	2,765 (5.5±2.3)
Source D	502	2,805 (5.6±2.4)

表 2　テストデータの詳細

話者	ファイル数	総再生時間 [s] (平均±標準偏差)
Target	266	1,054 (4.0±2.6)
Source A	266	1,263 (4.7±2.6)
Source B	266	1,180 (4.4±2.4)
Source C	266	1,129 (4.2±2.3)
Source D	266	1,143 (4.3±2.4)

3.3　声質変換モデルの学習

はじめに，各話者のトレーニングデータから WORLD を用いて音響特徴量を抽出する．抽出する特徴量は基本周波数と MCEP の 2 つであり，音声の再生時間に対して 5.0[ms]を 1 つのフレームとして，それぞれの特徴量を抽出する. この時 MCEP は，512 次元のスペクトル包絡を取得して 36 次元の MCEP に変換することで抽出を行う.

次に，それぞれの特徴量から統計情報として平均と標準偏差を求めて保存する．基本周波数は自然対数をとった後にそのまま算出する．MCEP は平均と標準偏差を算出した後に，MCEP に対して，平均を 0，標準偏差を 1 とする正規化を行う．変換モデルの学習は，この処理によって得られた MCEP を用いて行う.

次に，それぞれの特徴量から統計情報として平均と標準偏差を求めて保存する．基本周波数は自然対数をとった後にそのまま算出する．MCEP は平均と標準偏差を算出した後に，MCEP に対して，平均を 0，標準偏差を 1 とする正規化を行う．変換モデルの学習は，この処理によって得られた MCEP を用いて行う.

モデルの学習では，1 人の Source 話者と Target 話者に対応するデータ間で行われるため，CycleGAN-VC2 における 2 つのデータドメイン(X, Y)の組み合わせは(Source A, Target)，(Source B, Target)，(Source C, Target)，(Source D, Target)の 4 つとなる．学習時に与える入力データは 36 次元の MCEP におけるランダムな個所の連続した 128 フレーム分のセグメント

表 3　実験環境

OS	Ubuntu 20.04.3 LTS
CPU	AMD Ryzen 7 1800X
GPU	NVIDIA Geforce 1080Ti （CUDA 11.3）
開発言語	Python 3.8.10
機械学習ライブラリ	PyTorch 1.10.0

とする．502 回のイテレーションを 1Epoch とし，学習回数を 20 から 200Epoch の範囲で行う．この際，Cycle Consistency Loss とトレードオフの関係にある Identity Mapping Loss を制御するために，学習の進行具合に応じて重み付けを定めるパラメータの値を 0 へ変更して複数種類のモデル学習し，それぞれのモデルの性能を評価する．

本実験における環境を表 3 に示す．モデルの学習には，機械学習ライブラリ PyTorch と並列計算を行うためのプラットフォームである CUDA を利用する．

3.4　声質変換モデルによる合成音声の評価

声質変換は 2.3 節で述べた流れに沿って行う．WORLD によって音響特徴量を抽出する手法は，モデルを学習する時と同様である．モデルによって変換される 36 次元の MCEP は，512 次元のスペクトル包絡に変換され，変換後の基本周波数と無変換の非周期性指標と共に合成される．

生成した音声を評価するために，本研究では以下の 2 つの点に着目した．

- 声質的評価
- 言語的評価

声質的評価では，変換モデルによって Target 話者の声質が適切に再現されているかどうかを評価する．声質の変換性能を客観的に評価する指標として，Mel-cepstral distortion (MCD)を使用した．MCD は次のように表される．

$$MCD[dB] = \frac{10\sqrt{2}}{\ln 10}\frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{d=1}^{D-1}\{mc_d^X(t) - mc_d^Y(t)\}^2} \quad (2)$$

ここで，Dは MCEP の次元数，Tは MCEP のフレーム数であり，$mc_d^X(t)$と$mc_d^Y(t)$はt番目のフレーム，d次元の MCEP を表している．MCD は比較する MCEP の違いを算出するものであり，値が小さいほど違いの程度が小さくなることを意味する．MCD の算出時には，比較する MCEP に対して Dynamic Time Warping を適用してパラレルな状態とする．

言語的評価では，声質変換適用の前後において言語情報が保持できているかどうかを評価するために，音声データから読み上げられているテキスト情報を取得し，そのテキストを比較することで評価を行う．テキスト情報の取得には，Google 社が提供する「Google Cloud Speech-to-Text」を使用する[11]．取得した変換前と変換後のテキスト情報の比較は文字単位で行い，変換によって失われた言語情報を算出する．評価指標として失われた言語情報の割合[%]を，変換によって失われた語数をもとの文章全体の語数によって除算することで算出する．

各評価に使用するテストデータについて，声質的評価では全てのデータを使用する．一方で，言語的評価では Source 話者のテストデータから再生時間[s]が[1.5, 3.0), [3.0, 6.0), [6.0, 9.0)の各範囲にある音声データかを 5 つずつランダムに選択し，合計 15 個のデータを対象とした．この際に，全 Source 話者の間で 15 種類の各発話内容が同一となるようにする．

3.5　声質変換の処理時間計測

合成音声に対して声質変換を適用し話者を変更するという本手法において，元の合成音声を取得してから最終的な音声を再生するまでの間に処理の時間が必要となる．この時間は，声質変換を適用せずに音声を再生する場合を基準とすると，追加で掛かる時間である．この時間の程度を計測するために，学習モデルや統計情報をロードした状態の声質変換システムにおいて，変換する音声データファイルを読み込むタイミングで計測を開始し，変換を適用した音声データファイルを出力したタイミングで計測を停止する．

4.　実験結果

4.1　声質変換モデルの性能

声質変換を適用する前の各 Source 話者と Target 話者のデータ間で算出した声質的評価，および声質的評価と言語的評価の観点において最良である声質変換モデルの各評価値を表 4 に示す．表中の値は，平均値と標準偏差を意味する．声質的評価に着目すると，変換前，変換後ともに Source 話者 D の値が最小

表 4　声質変換モデルの評価値

話者	声質的評価		言語的評価
	変換前	変換後	
Source A	28.05±5.65	17.45±2.76	0.55±0.22
Source B	26.31±5.63	16.05±1.84	0.10±0.14
Source C	25.04±5.79	15.52±2.24	0.04±0.04
Source D	24.29±5.82	14.92±2.23	0.05±0.08

表 5　各 Source 話者の変換処理時間

話者	変換処理時間
Source A	0.28±0.01
Source B	0.28±0.01
Source C	0.30±0.01
Source D	0.30±0.01

であり，最も Target 音声を再現している．Source 話者 D の言語的評価も 0.05（5%）であり，変換による影響が非常に小さい．

声質的評価と言語的評価を総合した結果として，声質変換が適切に行われた話者は Source B，Source C，Source D である．一方で，変換後の音声をイヤホンで聞く主観的な評価として，元の合成音声が持つ自然さは少なからず失われており，ノイズが含まれていると感じられ，元の音声の品質を適切に保持できていないといえる．

4.2　声質変換の処理時間

異なる条件のもと学習を行った各モデルによる，声質変換の処理時間を計測し，各データの変換に要した処理時間をそのデータの再生時間で除算することで，1 秒間の音声を変換するのに必要な処理時間[s]の平均値および標準偏差を算出した．各 Source 話者の変換処理時間を表 5 に示す．結果として，学習条件の違いは処理時間に影響を与えず，各 Source 話者の変換モデルによる処理時間に大きな違いはない．

5.　考察

声質変換モデルによる声質的評価に着目すると，女性話者である Source A と Source B，男性話者である Source C と Source D の順で低くなっている．Target 話者の性別が男性であることを考慮すると，Source 話者と Target 話者同士の変換難易度は性別が同じ，あるいは声が似通っている場合に下がりやすいといえる．言語的評価にも着目した場合も，Source C，Source D の値が非常に小さく，ほかの Source 話者よりも声質変換が適切に行われているといえる．

声質変換の処理時間については，再生時間が 1 秒の音声を変換するために約 0.30[s]要する．つまり，対象とする音声が 5 秒間であれば 1.50[s]，10 秒であれば 3.0[s]の時間が必要となる．WaveNet モデルが 1 秒間に 20.0[s]の音声を出力する，すなわち 1 秒間の音声を生成するために 0.05[s]しか必要としないことを考慮すると，本手法の必要時間は少し長いといえる[1]．人とコンピュータが音声でやり取り可能な環境を提供するアプリケーションに本手法を組み込むことを考えると，コンピュータの応答を人に伝えるプロセスにおいて声質変換の時間がボトルネックになり，アプリケーションの応答時間が人に悪い印象を与える可能性は否めない．

声質変換の結果としては，変換自体は適切に行えているモデルを構築することができたが，元々の音声の品質が失われているものとなった．変換処理の後にノイズ除去などの品質向上を目的とした追加の処理を実行することを考えた場合，話者拡張に要する時間がさらに大きくなることは明白である．そのため，声質変換処理のみでこの問題を解決することが望ましく，モデルの改善や他の声質変換手法を検証することが必要となる．

6.　結言

本研究では，スマートフォン等に搭載されている音声アシスタントなどの音声合成を利用したサービスにおいて音声話者の種類が少ないことに着目し，話者のバリエーションを拡張するために声質変換技術を合成音声に適用して，その出力音声の評価を行った．変換元とする合成音声は Google Cloud Speech-to-Text サービスにおける 4 種類の話者（女性 2 種，男性 2 種）による音声とし，変換先とする音声は一般に販売されている 1 種の男性話者によるナレーション音声とした．結果として，1 種の女性話者を除いて，適切に話者性を変更することが可能であった．この結果から，対象とする音声同士の声質変換に対する相性の重要性を確認でき，和者数が 1 種類しか

存在しないような音声合成サービスにおいては本手法が適用できず，話者性を拡張できない可能性が考えられる．

本手法の問題点としては，最新の合成音声にみられる，まるで人が話しているような音声の品質が変換処理によって失われていることであった．そのため，変換処理によって品質の保持を十分に満たすことができれば，話者のバリエーションを拡張するのに有用であると判断できる．

参考文献

[1] Google Cloud 'Introducing Cloud Text-to-Speech powered by DeepMind WaveNet technology', <https://cloud.google.com/blog/products/ai-machine-learning/introducing-cloud-text-to-speech-powered-by-deepmind-wavenet-technology>, (2022/01/10).

[2] 戸田智基．確率モデルに基づく声質変換技術日本音響学会誌，Vol.24，No.1，pp.34-39，2010．

[3] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6820-6824, 2019.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, pp. 2223–2232, 2017.

[5] Takuhiro Kaneko and Hirokazu Kameoka. Parallel-data-free voice conversion using cycle-consistent adversarial networks. arXiv preprint arXiv:1711.11293, 2017.

[6] 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. E99-D, No. 7, pp.1877-1884, 2016.

[7] Masanori Morise. D4c, a band-aperiodicity estimator for high-quality speech synthesis. Speech Communication, Vol. 84, pp.57-65, 2016.

[8] Audiobook.jp, <https://audiobook.jp/ >, (2020/07/13).

[9] Google Cloud Text-to-Speech, <https://cloud.google.com/speech-to-text>, (2021/04/21)

[10] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. arXiv preprint arXiv:1711.00354, 2017

[11] Google Cloud Speech-to-Text, <https://cloud.google.com/speech-to-text>, (2022/01/14)