合成音声に対する声質変換技術の適用

高村健也・小高 知宏・黒岩 丈介 (福井大学大学院工学研究科)
白井 治彦 (福井大学工学部)・諏訪 いずみ (仁愛女子短期大学)

1　はじめに

音声合成技術の 1 つである声質変換技術は、ある音声に対して言語的特徴を保持しながら、音の高さなどの非言語的特徴を変換する手法である。この手法を用いることで、変換元の音声とは異なった声質を持つ音声を生成することが可能となり、話者の特徴変換や発話の補助など応用は様々である。

本研究では、音声アシスタントなどの合成音声にバリエーションを持たせることを目的に、一般に提供されている合成音声に対して声質変換処理を行い、特定の声質を持つ音声を生成する。日常における音声の声質を好みのものに変換することで、人々はその音声を楽しむことができると考える。

2　声質変換の方法

声質変換手法の 1 つとして、Cycle Generative Adversarial Network (CycleGAN) アルゴリズムを用いた手法が提案されている [1]。CycleGAN とは、異なる 2 つのドメインに属するデータ同士を学習することで、それぞれのデータを相互変換したデータを生成するモデルである。この声質変換手法は、変換元と変換先の音声データに対して、発話タイミングの同期などの複雑な前処理が不要であるノンパラレル声質変換に分類され、様々なデータ同士を扱いやすいため本研究で利用する。

3　実験

変換元とする合成音声は、Google 社が提供しているサービス「Google Cloud Text-to-Speech」による WaveNet 音声である。日本語で出力可能な話者は男性 2 種、女性 2 種の合計 4 種である。与えるテキストデータは、日本語テキストと読み上げ音声のセットとして公開されている「JSUT」コーパスに含まれる、「basic5000」のテキストを利用する [2]。変換先の音声データは、ある男性声優が読み上げているナレーション音声で、一般向けに販売されているものを利用する。以上、変換元 4 種類、変換先 1 種類として、それぞれ一対一の変換実験を行う。

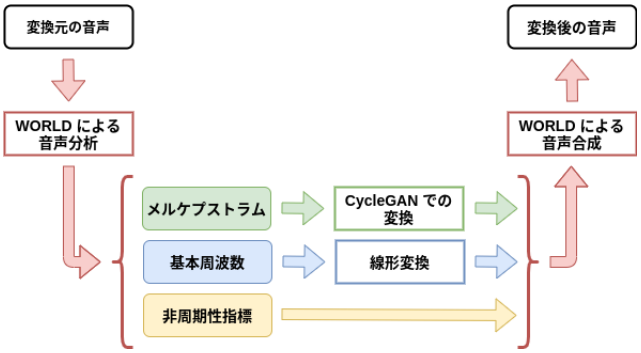

図 1　声質変換の流れ

声質変換において利用する音響特徴量は、メルケプストラム（MCEP）、基本周波数、非周期性指標である。特徴量の抽出には音声分析合成システム「WORLD」を用いる。これらの特徴量のうち、MCEP の変換モデルを CycleGAN によって学習する。そして、学習した CycleGAN モデルを使用して声質変換を行う。声質変換の流れを図 1 に示す。実験の詳細については当日示す。

4　考察とまとめ

既存の合成音声に対して声質変換技術を適用し、新しい音声の生成を行った。本研究の手法では、WaveNet などの Vocoder を使用しない音声生成モデルと比較すると、声質変換に追加の処理時間が掛かる。何らかのアプリケーションに本手法を組みこんだ場合には、この時間を減らす工夫が必要であると考えられる。

参考文献

[1] 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), 2019

[2] Ryosuke Sonobe, Shinnosuke Takamichi and Hiroshi Saruwatari, “JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” arXiv preprint, 1711.00354, 2017.