様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 ２３ 日現在

機関番号：１３９０１
研究種目：若手研究
研究期間：2019～2021
課題番号：１９Ｋ２０２９５

研究課題名（和文）超短遅延音声変換システムの実現に関する研究

研究課題名（英文）Implementation of super low-delay voice conversion system

研究代表者
小林　和弘（Kobayashi, Kazuhiro）

名古屋大学・情報基盤センター・研究員

研究者番号：５０８１５６０２

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：音声変換は、入力話者が発話した音声を異なる話者の音声へと変換するシステムである。さらにストリーミング変換処理技術と組み合わせる事で、入力された音声をリアルタイムに変換する事が可能である。一方で、遅延量と変換品質はトレードオフの関係にあり、遅延量を短くする事で多くの品質劣化が生じる事が確認されている。本研究課題では、これらの課題を解決に向けて、パラレルデータを用いた音声変換やノンパラレルデータを用いた音声変換などの研究開発に取り組んだ。

研究成果の学術的意義や社会的意義
音声変換技術は、人と人のコミュニケーションで使われる音声を対象とした技術である。声優などの卓越した話者を除き、個人が発話可能な声色の表現範囲は狭く、多くの人にとって他者の声色を完全に模倣する事は困難である。音声変換技術は、声色の表現範囲の壁を超え、誰もが多種多様な声色で発話する事を可能とする技術として期待されている。とりわけ、短遅延音声変換は入力された音声を逐次的に変換できるため、人と人とのコミュニケーションを大きく拡張する事が期待されている。一方で、高品質かつ短遅延な音声変換は未だ困難であるため、その実現に向けた研究成果や知見は重要であると考えられる。

研究成果の概要（英文）：Voice conversion is a technique to convert one's speech into another speaker's speech. It is possible to implement a low-delay voice conversion system combined with streaming conversion techniques. However, since there is a tradeoff between delay and conversion quality, it has been observed that setting a slight delay tends to degrade conversion quality. To alleviate these problems, we aimed to implement low-latency voice conversion systems using parallel or non-parallel utterances in this research.

研究分野：人間情報学

キーワード：音声変換　話者　深層学習　リアルタイム　主観評価実験

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

音声変換は、入力話者の声質を異なる目標話者の声質へと変換する技術である。音声変換を実現する古典的な変換手法として、パラレルデータを用いた音声変換がある。まず、入力話者および目標話者が同一発話内容文を発話した音声データから音響特徴量を抽出する。これらの音響特徴量の対応関係を統計モデルによりモデル化する事で、新たに入力された入力話者の音響特徴量を目標話者の音響特徴量へと変換する事が可能となり、目標話者の声色を持つ変換音声を得ることが出来る。

2018 年頃には、深層学習を音声変換へと適用する事で、古典的な音声変換技術に比べより高品質な音声変換が実現可能となった。代表的な手法として、WaveNet を用いた音声変換システムがある。WaveNet とは、深層学習を用いた音声波形生成モデルであり、信号処理技術を利用した波形生成に比べて、高い音質での音声波形生成が可能である。また、WaveNet を音声変換に適用する事で、非常に高い品質での音声変換が可能となった。一方、WaveNet による音声波形生成には、膨大な計算時間がかかり、そのままの形態で短遅延音声変換を実現する事は困難であった。

他にも、代表的な枠組みとして、変分オートエンコーダを用いたノンパラレル音声変換システムがある。ノンパラレル音声変換は、学習データに入力話者と目標話者の同一発話内容文を用いるという制約を取り除き、任意の発話から変換モデルの学習を可能とする。一方で、ノンパラレル音声変換を短遅延音声変換に適用した研究は存在せず、短遅延音声変換での変換品質は未確認であった。

２．研究の目的

人は、自身の見た目を変える事で、自身のキャラクター性を変化させる事が可能である。例えば、祭事において、鬼の仮面を装着した人は、見た目と声色の両者で、あたかも鬼であるかの様に振る舞い、本物の鬼がそこに存在するかのように、他者に認識させる事が可能である。近年では、ヴァーチャルリアリティ技術の発展により、仮想空間上でアバターを操り、他者と音声コミュニケーションする手段が確立している。仮想空間では、アバターを変えることで、他者に知覚される見た目を変える事は容易である。一方で、音声に対するアバター技術は、確立していないため、見た目はキャラクター、音声は元ユーザーのままであり、他者に対して知覚的な違和感を度々生じさせる。

音声に対する知覚的な違和感は、見た目の情報と連動した場合のみならず、様々な状況において生じる。代表的な例として、非ネイティブスピーカの発音、音高やリズムを誤った歌唱、発声障害者の発話などが挙げられる。ノンリアルタイムで音声を伝達するという条件下では、再発話する事や、手動で音声波形を加工する事で、他者に与える知覚的違和感を解消する事ができる。一方で、対面や電話を用いたコミュニケーションなどの様に、リアルタイムで発話を伝達する場合では、知覚的違和感を解消する事は未だ困難である。これらの課題を解決する音声加工技術が実現すれば、新たなコミュニケーションの実現が期待される。

３．研究の方法

本研究では、音声コミュニケーションの進展を目指して、超短遅延音声変換技術に関する研究を進める。まず、種々のユーザが音声変換を利用できる環境を構築するために、学習データに対する制約が少ないノンパラレル音声変換技術を対象に研究を進めた。ノンパラレル音声変換システムでは、複数の話者の音声データを学習データとして用いる事で、多対多の音声変換を可能とする。これにより、いかなる入力話者の音声に対しても所望の目標話者へと変換する事が可能となる。言い換えると、複数の話者の音声データで音声変換モデルを学習すれば、ユーザ自身の音声データを学習に用いる必要がなくなる。まず、ノンリアルタイムなノンパラレル音声変換システムの構築を行い、その技術的可用性の検討を行う。更に、短遅延音声変換への拡張を目指し、研究を進めた。

次に、より高品質な音声変換を実現するために、パラレルデータを用いた音声変換に関する研究を進めた。本助成事業期間中に、パラレルデータを用いた End-to-End 音声変換システムが、従来の音声変換システムに比べ、非常に高品質な音声変換を実現出来ると他の研究チームにより公表された。本研究課題でも、高品質な短遅延音声変換を実現するべく、End-to-End 音声変

換システムに関する研究を実施した。
さらに、多用な音声信号に対する短遅延音声変換の可用性を検討するために、発声障害者のための音声強調技術に関して研究を行った。発声障害者とは、喉頭がん、咽頭がん、構音障害などにより、自然な発話ができなくなった話者をさす。本研究では、喉頭がんによって喉頭を摘出した話者を対象に、電気式人工喉頭音声に対する音声強調技術の研究を実施した。電気式人工喉頭を用いた音声は、機械的で自然ではないため、健常者が発話する様な自然な発話へと短遅延で変換できる強調技術の実現に向けて研究を進めた。

４．研究成果

ノンパラレル音声変換を用いた短遅延音声変換の実現に向けた研究として、ベクトル量子化変分オートエンコーダ（VQ-VAE: Vector-Quantized Variational Auto-Encoder）を用いた音声変換に関する研究を行った。本手法は、埋め込み空間に連続な確率分布を用いる変分オートエンコーダに対して、埋め込み表現として離散符号を用いるオートエンコーダを用いた音声変換法である。図１の様に、離散コードの埋め込みを階層的に表現する事でより高精度な音声変換を可能とする。また、音声波形生成には、非自己回帰型音声波形生成法である ParallelWaveGAN を活用し、高速かつ高品質な音声変換を実現した。本研究成果は、査読付き国際会議 IEEE ICASSP に採録され発表を行った。また、研究機関や企業に所属する研究者がノンパラレル音声変換システムを利用出来るよう、オープンソースソフトウェアとして GitHub にて公開を行った。本手法は短遅延音声変換技術へと拡張され、任意のユーザによる短遅延な音声変換が可能となった。一方で、因果的な制約のため、ノンリアルタイムな音声変換に比べ、大きく変換品質が低下する事が確認された。

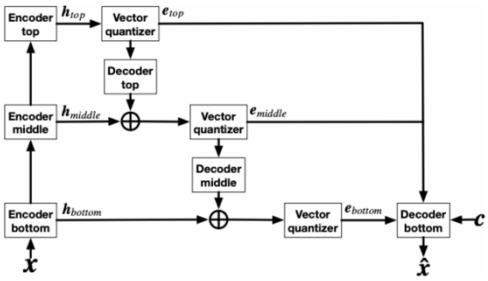

図 1　VQVAE2 に基づく音声変換

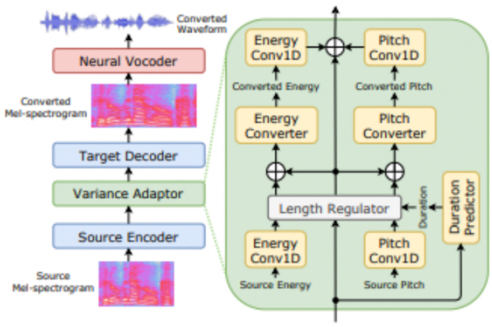

図 2　非自己回帰型音声変換システム

より高品質かつ短遅延な音声変換システムを実現するべく、進展が目覚ましい End-to-End 音声変換システムに関する研究を実施した。本システムは、代表的な End-to-end 音声変換システムである Tacotron2 のように自己回帰的に音響特徴量系列を推定する音声変換システムではなく、図２に示す非自己回帰的に音声特徴量系列を推定する音声変換システムであり、高品質かつ低遅延な音声変換が可能となった。本研究に対する研究成果として、２本の査読付き国際会議での発表を行った。

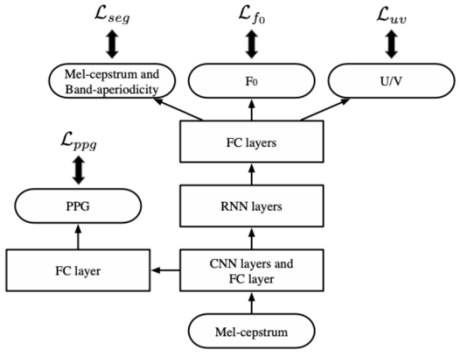

図 3　マルチタスクラーニングを用いた音声強調

発声障害者補助として、電気式人工喉頭音声に対する短遅延音声強調に関する研究を行った。本研究では、電気式人工喉頭音声から抽出された音響特徴量を短遅延で自然発話の音響特徴量へと変換する事で、喉頭摘出者がより自然な音声で発話する事を可能とする技術の研究開発を行った。図３の様にマルチタスクラーニングを用いて、分節的特徴量、韻律的特徴量の両者を一つのモデルで学習する。その結果、低計算量かつ低遅延な音声強調技術が実現する事が出来た。本研究に対する成果として査読付き国際会議で１件の研究発表を行った。また、喉頭を摘出された発声障害者に対して音声強調システムを構築し、実際に利用してもらうなど、今後の研究開発に向けて多くの知見を得ることが出来た。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 T. Hayashi, K. Kobayashi, T. Toda
２．発表標題 An investigation of streaming non-autoregressive sequence-to-sequence voice conversion
３．学会等名 IEEE ICASSP（国際学会）
４．発表年 2022年

１．発表者名 K. Kobayashi, W.-C. Huang, Y.-C. Wu, P.L. Tobing, T. Hayashi, T. Toda
２．発表標題 Crank: an open-source software for nonparallel voice conversion based on vector-quantized variational autoencoder
３．学会等名 IEEE ICASSP（国際学会）
４．発表年 2021年

１．発表者名 T. Hayashi, W.-C. Huang, K. Kobayashi, T. Toda
２．発表標題 Non-autoregressive sequence-to-sequence voice conversion
３．学会等名 IEEE ICASSP（国際学会）
４．発表年 2021年

１．発表者名 K. Kobayashi, T. Toda
２．発表標題 Implementation of low-latency electrolaryngeal speech enhancement based on multi-task CLDNN
３．学会等名 EUSIPCO（国際学会）
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

- 国際会議チュートリアル講演
T.Toda, K.Kobayashi, T.Hayashi, "Statistical voice conversion with direct waveform modeling" Tutorial, INTERSPEECH 2019, Graz, Austria, Sep. 2019.

- オープンソース音声変換ソフトウェア crank の公開
https://github.com/k2kobayashi/ crank

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関