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　６ 年　 ９ 月 １２ 日現在

機関番号：１３３０２
研究種目：基盤研究(B)（一般）
研究期間：2020～2023
課題番号：２０Ｈ０４２０７

研究課題名（和文）非並行型学習法にもとづいた多言語間多話者属性変換システム

研究課題名（英文）Multi-lingual multi-speaker voice conversion system by non-parallel learning method

研究代表者

赤木　正人（Akagi, Masato）

北陸先端科学技術大学院大学・先端科学技術研究科・名誉教授

研究者番号：２０２４２５７１

交付決定額（研究期間全体）：（直接経費）　　13,400,000円

研究成果の概要（和文）：本研究では，音声変換（Voice Conversion: VC）による多言語音声へのパラ言語・非言語情報付加を最終目標として設定し，その中心課題の一つである話者性操作を目指して，多言語間での非並行型学習法の提案およびこの学習法にもとづいた多数話者間の属性変換システムの構築を検討する。具体的には，(A) VCのソース言語とターゲット言語が異なる場合の話者情報の扱い方，(B) 多話者対多話者属性変換，(C) 未学習話者を想定した場合の話者特徴の記述法，(D) 変換後の音声の品質・了解度保証である。これらすべてを深層学習の枠組みで検討し，適切な目的関数を設定することにより全体を最適化する。

研究成果の学術的意義や社会的意義
話者のパラ言語および非言語情報を抽出し合成音声に付加することができる音声-音声翻訳のための多言語間音声変換システムを開発するために，その第一歩として，非言語情報の一つである話者属性（性別，年齢，声質等）の自由な変換操作を目指して，多言語間での音声変換のための非並行型学習法を提案し，これにもとづいた変換システムを検討する。これにより，ある言語で話をした話者の声と同じ声質で別の言語の音声を合成できる，しかも使用言語および使用話者を選ばないシステムの構築が可能となり，入力音声に含まれる話者属性を出力音声でも維持できることで，コミュニケーションの質を向上させることができる。

研究成果の概要（英文）：This study aims to enhance paralinguistic and non-linguistic information in multilingual speech through Voice Conversion (VC), with the manipulation of speaker identity in speech as one of its central objectives. To achieve this, we propose a non-parallel learning method for cross-lingual VC and explore the construction of a multi-speaker attribute conversion system based on this learning approach. Specifically, the issues addressed include (A) handling speaker information when the source and target languages of VC are different, (B) achieving multi-speaker-to-multi-speaker attribute conversion, (C) describing speaker characteristics when considering the use of unseen speakers, and (D) ensuring the quality and intelligibility of synthesized speech after conversion. By addressing these challenges within the framework of deep learning and optimizing the entire process through appropriate objective functions, we attempt to achieve comprehensive optimization.

研究分野：音声情報処理

キーワード：パラ言語情報　非言語情報　音声変換　非並行型学習

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

本研究の最終目標は，図 1 に示すような，話者のパラ言語および非言語情報を抽出し合成音声に自由に付加することができる音声 - 音声翻訳（Speech-to-Speech Translation: S2ST）のための多言語間音声変換（Voice Conversion: VC）システムを開発することである。本提案では，非言語情報の一つである話者属性（性別，年齢，声質等）の自由な変換操作を目指して，多言語間での VC のための非並行型学習法を提案し，これにもとづいた変換システムを検討する。

S2ST は，ある言語の音声に対して，音声認識（Speech-to-Text），機械翻訳（Text-to-Text），音声合成（Text-to-Speech）を通して，別の言語の翻訳済み音声を出力するシステムであり，現在精力的に研究が進められている。現有の S2ST では，言語情報は図中の上の経路を通して伝達されるが，効果的な音声コミュニケーションに不可欠である話者の個人性などの非言語情報や強調などのパラ言語情報は伝達されない。

本研究では，話者の属性などの情報も伝達できるシステム，具体的には，図中の下の経路のように，ある言語で話をした話者の声と同じ声質で別の言語の音声を合成できる，しかも使用言語および使用話者を選ばないシステムを開発する。これにより，入力音声に含まれる話者属性を出力音声でも維持できることで，コミュニケーションの質を向上させることができる。言語および話者を選ばないシステムとするためには，多言語間での多話者対多話者の音声変換を可能とするシステムを検討する必要がある。このために，言語が異なる非並行のデータセットを用いて VC システムをどのように学習するか，また，多数話者間の自然な属性変換をどのように行うかに焦点をあて，研究を実施する。

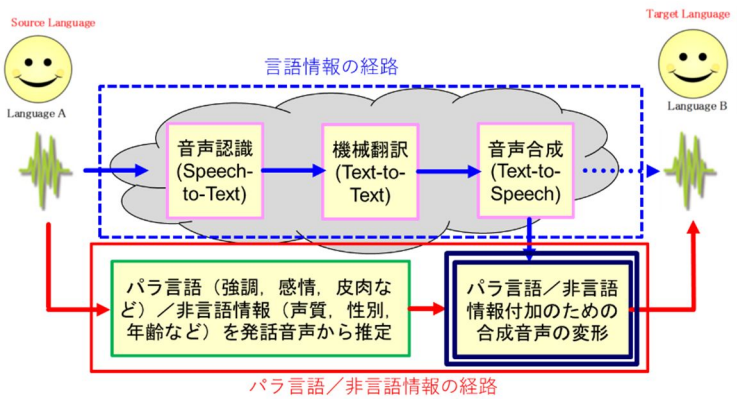

図 1：Speech-to-Speech Translation System

VC は，音声信号内の言語情報を保持しながら話者特性を操作すること，たとえば，男声から女声へ，若者から老人への音声変換など，を目的としている。VC アプローチを分類するにはさまざまな方法があるが，一つの要因は，Source 音声と Target 音声の間で 1 対 1，1 対多，多対 1，または多対多の変換を実行するかどうかである。もう一つの要因は，VC システムの学習に，並行（話者は異なるが同じ言語情報）あるいは非並行（話者も言語情報も異なる）のどちらのデータセットを使えるかということである。VC のための従来の方法は，Source 音声と Target 音声の音響特徴の同時確率をモデル化するためのガウス混合モデル（GMM）にもとづいている。最近，深層学習を適用して VC を End-to-End で学習する方法も提案されている。しかしながら，現在の VC 法のほとんどは、並行データを用いることができる環境で，1 対 1 の変換を達成

できるのみである。固有声変換法を用いれば，1 対多への拡張も可能であるが，依然として，並行データが存在する条件での手法である。この制限により，S2ST などの複数言語間でのアプリケーションで，VC システムの有用性が制限されている。

２．研究の目的

本研究の目的は,「音声変換（Voice Conversion: VC）による多言語音声へのパラ言語・非言語情報付加を行うために，音声に含まれる話者性の自由な操作が行える手法の提案」である。

上記の目的を達成するためには，次に示す課題を解決する必要がある。

A) 多言語間の S2ST では，VC の Source 音声と Target 音声は異なる言語である。このため，学習時に，並行データセットは利用できない。

B) 誰でも使えるシステムを目指すのであれば，多話者対多話者変換が必要となる。

C) 多話者の中には，未学習話者が含まれる。

D) 変換後の合成音声の品質（自然性，了解度）保証は重要である。

３．研究の方法

この問題を解決するためには,次の要件を満たすシステムが必要であると考えた。本研究では,これらの要件に対して，様々なタイプの深層学習法についての制約条件，目的関数を精査し，それぞれの問題に対して道具として最適な深層学習法を議論することで，問題解決にチャレンジする。ただし，深層学習さえ使えば何でも学習してくれる，というのは幻想である。申請者らが研究してきた変形規則にもとづいた手法で得られた知見をも活かして,適切な入力,適切な出力,適切な評価関数，適切な構造についての議論が必要である。

本研究では，これらの課題を次に説明する手法を用いて順次解決する。なお，文頭の記号は上記課題の記号と一致する。

A) 多言語間の音声変換：多言語間の音声変換は，特殊なタイプの非並行音声変換である。この問題を解決するために，言語情報と話者の属性情報を効率的に分離する音声分解法を検討し，話者情報のみを扱えるようにする。一つの有力な候補は Variational Autoencoder（VAE）である。しかし，現有の VAE で Encode された話者情報は，単純な正規分布で表されることが多く，表現能力は乏しい。VAE を用いて話者の属性を細かく抽出できるようにするには，Encode された情報の新しい表現法を検討する必要がある。解決法として，VAE での Encode 先を話者空間内のベクトルとしてとらえ，その空間（話者空間）での話者の変形を容易にするように仕向ける。

B) 多話者対多話者の音声変換：従来法のように，一人一人の話者の特徴を記述しその間の変換法を構築していては，話者を選ばないシステムには程遠い。このため，話者一人一人の対応関係を記述するのではなく，複数の話者が埋め込まれた話者空間を学習により構築する。一旦話者空間が構築できれば， Source 話者がその空間のどの位置に存在するのかを少量の発話データで推定できる。そして，推定された Source 話者の位置から Target 話者の位置への移動操作を行うことにより，音声変換は可能となる。空間の表現法については，我々のグループで長年研究してきた感情音声認識合成のための感情空間構築法，歌声合成のための歌声空間構築法などの経験を活かし，ヒトが話者を知覚するための複数の因子を用いた話者空間表現法を提案する。

C) 未学習話者への対応：話者空間の構築時に，空間をきめ細かく（できれば連続的な空間として）構築する。これにより，Source 話者の空間内での位置の推定→Target 話者位置への移動

において，未学習入力から未学習出力への拡張も可能となる。これは情報表現として話者空間を用いる利点の一つである。問題は，多くない話者数のデータをもとにどのようにして話者空間を連続的に構築するための学習を実施するかである。解決方法として，少数話者の音声データを使用して話者空間の外郭を構築したのち，Generative Adversarial Network (GAN)あるいは話者間の音声モーフィング操作により話者空間を連続化する。

D) 音声の自然性と了解度の保証：確率モデルに基づく従来の VC は，音響特徴が単純なガウス分布を有することを仮定して，平均二乗誤差を学習の目的関数として使用している。しかしながら，音響的特徴は通常非常に複雑な確率分布を持つため，単純化された仮定はしばしば音響特徴を過度に平滑化し合成波形の劣化をもたらす。考えられる解決策の一つは、変数の変更定理を正規化フローの形式で適用することである。これは，事前の仮定なしに単純なガウス分布から音響特徴の密度を正確に推定できる。さらに，変調スペクトルのような音声の自然さと了解度に関連する音響的特徴を目的関数に組み込むことで，学習のための目的関数の質を向上させる。

４．研究成果

A) 言語情報と話者情報の分離

言語内容が多岐にわたる音声を入力とする話者情報（性別および発話スタイルの分離）の検討をおこなった。階層型の Vector Quantized Variational Autoencoder を用いることにより，性別が異なる複数のプロのアナウンサーと素人の朗読音声を対象として，言語内容によらない性別および発話スタイルの抽出分離に成功した。これは，将来の複数言語を入力とする話者情報の分離への足掛かりとなる。

この成果は，国際ワークショップ Voice Conversion Challenge 2020 および IEEE の雑誌 IEEE Access に発表した。

B) 話者知覚の知見にもとづいた話者空間の構築

話者知覚の知見にもとづいて話者空間記述のための因子の検討を行い，これらの因子で張られる空間へ話者情報を展開することに成功した。その空間の中で発話スタイルの変換を行ったところ，この空間内の任意の位置での話者情報を持つ音声の合成が可能となった。具体的には，性別が異なる複数のプロのアナウンサーと素人の朗読音声を対象として，プロのアナウンサーの発話スタイルを真似た合成音声は，アナウンサー音声に特有の Clear Speech 特性を持つ音声となった。また，この変換法により，プロのアナウンサーの音声と同様に，雑音中の音声了解度が向上することを確認した。これは，将来の複数言語を入力とする話者情報の分離・変形・再合成への足掛かりとなる。

これらの成果は，次の雑誌および会議にて発表した。 2020 年 10 月に上海で開催された複数言語での音声変形に関する国際コンペティション Voice Conversion Challenge 2020 において，本グループが提案した新たな voice conversion 法を駆使した変形音声を出品し，第一位ではなかったものの好成績をあげた。また、この内容を拡張した論文を IEEE のオープンジャーナル（IEEE Access）に投稿し、すでに出版されている。さらに，音声研究に関する国際会議 InterSpeech2022 で発表した。

C) 話者の個人性に関連する特徴量の抽出方法

話者ごとに特徴的な形状を持つ声道形状の分岐管を考慮した声帯音源波形と声道フィルタの同時推定法（ARMA-LF モデル）を提案し，音声分析・合成に適用した。このモデルにより，分岐官により生成されるスペクトル上での零点の振舞いを精度よく推定できるようになった。こ

の成果は 2021 年にオンラインで開催された国際会議 APSIPA2021 において発表された。

また，ARMAX-LF モデルにおいて，声帯音源パラメータ推定に深層ニューラルネットワークを用いた方法を付加し，推定結果の高精度化および安定化を図った。この成果は，雑誌 Speech Communication に投稿中である。

推定した特徴量を話者変換のための特徴として使用することについては，その前段階として，Speaker Anonymization を目指した音声変形のために，声道形状に関係するパラメータを変形制御する方法の有効性を検討した。この結果は，国際会議 InterSpeech2022 で発表した。

D) 音声の自然性と了解度の保証

音声の属性変換を行えたとしても，変換後の合成音声の品質が悪く，音声了解度が高くないのであれば，変換の意味は薄れる。このため，合成音声の品質・了解度の保証は重要課題である。プロのアナウンサーの発話スタイルを真似た合成音声手法を検討し，雑音中での音声品質向上，特に音声了解度を向上させるメカニズムの解明を試みた。この結果，提案した変換法により，プロのアナウンサーの音声と同様に，雑音中の音声了解度が向上することを確認した。また，音声了解度向上とこれに関連する音響特徴の候補を発見した。この成果は，音響に関する国際会議 ICA2022，および，日本音響学会での招待講演で報告した。音響特徴の候補については，国際会議 APSIPA2023 で発表した。

E) 残された課題

本研究では，4 年間の研究期間中に，非言語情報の一つである話者属性の自由な変換操作を目指して，具体的な課題として，A) VC の Source 言語と Target 言語が異なる場合の話者情報表現，B) 誰でも話者となりえるシステムとするための多話者対多話者属性変換，C) 未学習話者の使用を想定した場合の話者特徴の記述法，D) 変換後の合成音声の品質・了解度の保証，を設定した。課題 A)〜D)それぞれに対応した手法の提案は行ってきた。しかし，多言語への展開および実際に多言語の環境における性能評価には若干の遅れが出ている。そこで今後の研究方向性として，次の計画を立てる。

話者性を多く含む特徴として，声道形状と声帯音源波形が知られている。これらは，性別，年齢，声質等に深くかかわっている。また，言語が異なっても音声生成系である声道形状と声帯音源波形の特徴は保持される。このため，声道形状と声帯音源波形は，言語によらない話者情報表現として適切な特徴と考えられる。そこで，音声波形から声道形状と声帯音源波形をより高精度で推定する手法を用いて推定した特徴量およびその関連量を話者変換のための特徴として使用すること，を検討する。

具体的には，(1) 声道形状と声帯音源波形の高精度推定に対しては，声帯音源波形を音声波形から同時に推定できる手法（ARMAX-LF モデル）の更なる高精度化を実施する。また，(2) 推定した特徴量を用いた話者変換に対しては，声道形状と声帯音源波形およびその関連量を用いて話者個人性の操作（ターゲットとして言語によらない他話者への変換）を行う。そして，(3) 多言語を入力とする属性変換および変換後の合成音声の品質・了解度の保証へとつなげ，これらの研究から得られた成果を取りまとめ学会発表を行う。

５．主な発表論文等

〔雑誌論文〕　計10件（うち査読付論文　10件／うち国際共著　4件／うちオープンアクセス　1件）

１．著者名 Dung Kim Tran, Masato Akagi, and Masashi Unoki	４．巻 -
２．論文標題 Increasing Speech Intelligibility by Mimicking Professional Announcers’ Voices and Its Physical Correlates	５．発行年 2023年
３．雑誌名 Proc APSIPA2023	６．最初と最後の頁 1162, 1167
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kai Li, Xugang Lu, Masato Akagi, Jianwu Dang, Sheng Li, Masashi Unoki	４．巻 -
２．論文標題 Relationship Between Speakers’ Physiological Structure and Acoustic Speech Signals: Data-Driven Study Based on Frequency-Wise Attentional Neural Network	５．発行年 2022年
３．雑誌名 Proc. EUSIPCO2022	６．最初と最後の頁 379, 383
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Tuan Vu Ho, Maori Kobayashi, Masato Akagi	４．巻 -
２．論文標題 Speak Like a Professional: Increasing Speech Intelligibility by Mimicking Professional Announcer Voice with Voice Conversion	５．発行年 2022年
３．雑誌名 Proc. Interspeech2022	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kai Li, Sheng Li, Xugang Lu, Masato Akagi, Meng Liu, Lin Zhang, Chang Zeng, Longbiao Wang, Jianwu Dang, Masashi Unoki	４．巻 -
２．論文標題 Data Augmentation Using McAdams-Coefficient-Based Speaker Anonymization for Fake Audio Detection	５．発行年 2022年
３．雑誌名 Proc. Interspeech2022	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Masato Akagi	4．巻 -
2．論文標題 Increasing speech intelligibility in noise based on concepts of modulation spectrum and voice conversion to professional announcer voice	5．発行年 2022年
3．雑誌名 Proc. of the 24th International Congress on Acoustics	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Dung Kim Tran, Masato Akagi, and Masashi Unoki	4．巻 -
2．論文標題 Deep Hashing for Speaker Identification and Retrieval Based on Auditory Sparse Representation	5．発行年 2022年
3．雑誌名 Proc. APSIPA2022	6．最初と最後の頁 938, 944
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Li Yongwei、Tao Jianhua、Erickson Donna、Liu Bin、Akagi Masato	4．巻 29
2．論文標題 F0-Noise-Robust Glottal Source and Vocal Tract Analysis Based on ARX-LF Model	5．発行年 2021年
3．雑誌名 IEEE/ACM Transactions on Audio, Speech, and Language Processing	6．最初と最後の頁 3375～3383
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TASLP.2021.3120585	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Kai Li, Masashi Unoki, Yongwei Li, Jianwu Dang, Masato Akagi	4．巻 -
2．論文標題 Study on Simultaneous Estimation of Glottal Source and Vocal Tract Parameters by ARMAX-LF Model for Speech Analysis/Synthesis	5．発行年 2021年
3．雑誌名 Proceeding of APSIPA2021	6．最初と最後の頁 36～43
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Ho Tuan Vu、Akagi Masato	４．巻 9
２．論文標題 Cross-Lingual Voice Conversion With Controllable Speaker Individuality Using Variational Autoencoder and Star Generative Adversarial Network	５．発行年 2021年
３．雑誌名 IEEE Access	６．最初と最後の頁 47503～47515
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2021.3063519	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Ho Tuan Vu、Akagi Masato	４．巻 -
２．論文標題 Non-parallel Voice Conversion based on Hierarchical Latent Embedding Vector Quantized Variational Autoencoder	５．発行年 2020年
３．雑誌名 Proceeding of Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020	６．最初と最後の頁 140～144
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計7件（うち招待講演　1件／うち国際学会　0件）

１．発表者名 赤木正人
２．発表標題 確実に情報を伝える音声避難誘導システムの構築に向けて
３．学会等名 日本音響学会音声研究会
４．発表年 2023年

１．発表者名 Kimdung Tran, Masato Akagi and Masashi Unoki
２．発表標題 Increasing Speech Intelligibility for Evacuation Guidance by Mimicking Professional Announcers’ Voice: Discussion on Speech Intelligibility and Its Physical Correlates
３．学会等名 電子情報通信学会音声研究会
４．発表年 2023年

１．発表者名
赤木正人

２．発表標題
音声変形による雑音残響環境での音声了解度向上

３．学会等名
日本音響学会2023年度春季研究発表会（招待講演）

４．発表年
2023年

１．発表者名
Kai Li, Masato Akagi, Masashi Unoki

２．発表標題
Estimation of Glottal Source Parameters of the LF Model Using Feed-forward Neural Network

３．学会等名
日本音響学会令和4年春季大会

４．発表年
2022年

１．発表者名
Tuan Vu Ho and Masato Akagi

２．発表標題
Cross-lingual voice conversion with Multi-codebook Hierarchical Vector-Quantized Variational Autoencoder

３．学会等名
ASJ '2020 Fall Meeting

４．発表年
2020年

１．発表者名
Tuan Vu Ho and Masato Akagi

２．発表標題
Improving spectral detail and F0 modelling for VAE-based cross-lingual voice conversion with adversarial training

３．学会等名
ASJ '2021 Spring Meeting

４．発表年
2021年

１．発表者名 Kai Li, Yongwei Li, Jianwu Dang, Masashi Unoki, and Masato Akagi
２．発表標題 Estimation of Glottal Source Waveforms and Vocal Tract Shapes Based on ARMAX-LF Model
３．学会等名 ASJ '2021 Spring Meeting
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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
研究分担者	鵜木　祐史 (Unoki Masashi) (00343187)	北陸先端科学技術大学院大学・先端科学技術研究科・教授 (13302)	

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