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　５ 年　 ６ 月 ２０ 日現在

機関番号：１３９０３
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１１８６２

研究課題名（和文）異なる言語において入力音声の話者・感情を再現する深層学習に基づく多言語音声合成

研究課題名（英文）Multilingual speech synthesis based on deep learning to reproduce the speaker and emotion of input speech in different languages

研究代表者
橋本　佳（HASHIMOTO, Kei）

名古屋工業大学・工学（系）研究科（研究院）・准教授

研究者番号：１０６３５９０７

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、入力音声と異なる言語において入力音声の話者・感情を再現する多言語音声合成を実現するために、入力音声の言語・話者・感情それぞれに依存する音声の特徴を分離可能とする多言語音声合成モデルの学習に取り組んできた。特に、言語と話者の特徴を分離するため敵対的学習に基づく多言語音声合成や話者と感情を分離するモデル構造を提案した。また、補助特徴として顔画像を利用するモデルなどを提案した。提案法によって、入力音声と異なる言語において話者の特徴が再現された音声の生成が実現され、より自然なグローバルコミュニケーションの実現が期待される。

研究成果の学術的意義や社会的意義
本研究では、音声に含まれる話者・言語・感情といった3つの特徴に注目し、入力音声と異なる言語において入力音声の声質や感情を再現する多言語音声合成技術に確立に取り組んだ。本研究の成果は、音声翻訳システムに応用することで、自分の話すことができない言語においても、自分の声のまま、感情表現を含む自然なコミュニケーションを実現することが期待される。

研究成果の概要（英文）：To realize multilingual speech synthesis that reproduces the speaker and emotion of input speech in different languages, I have been working on deep neural network (DNN)-based multilingual speech synthesis that can separate speech features that depend on the language, speaker, and emotion of the input speech. I have proposed multilingual speech synthesis based on adversarial learning to separate language and speaker features, and a model structure to separate speaker and emotion. Additionally, I have proposed a speech synthesis model that uses face images as auxiliary features. The proposed method is expected to realize more natural global communication by generating speech that reproduces the characteristics of the speaker in different languages.

研究分野：音声情報処理

キーワード：音声合成

１．研究開始当初の背景

申請者はこれまでに統計モデルに基づく音声合成、近年では特に深層学習に基づく音声合成に取り組んできた。音声合成に適したモデル学習手法やモデル構造、様々な声質を再現する手法、音声波形モデルなどの研究に取り組み、合成音声の品質および声質の再現性を改善してきた。特定言語・特定話者・平静（読み上げ）という限定された条件においては非常に高品質な合成音声を生成することが可能となった。今後は、合成音声の品質だけでなく、感情音声や対話調といった発話スタイルなど、表現の多様性について人間と同程度、または人間以上の機能を備えた音声合成手法の研究開発が必要である。

申請者は、複数の言語・話者が混在する音声データから音声合成モデルを学習し、ある話者の個人性（声質）を異なる言語において再現する手法についても研究を進めている。これにより、自分が話すことができない言語においても自分の声の合成音声を生成することが可能となる。しかし、読み上げ音声に限定しており、感情や発話スタイルが異なる言語において再現されるには至っていない。自然なグローバル・コミュニケーションの実現のためには感情や発話スタイルを再現する必要がある。

２．研究の目的

本研究の目的は、入力音声と異なる言語において入力音声の話者・感情を再現する多言語音声合成技術を確立することである。この多言語音声合成技術を用いた音声翻訳システムを開発することで、自分の声のまま感情も伝えることができる自然な異言語間コミュニケーションの実現を目指す（図 1）。

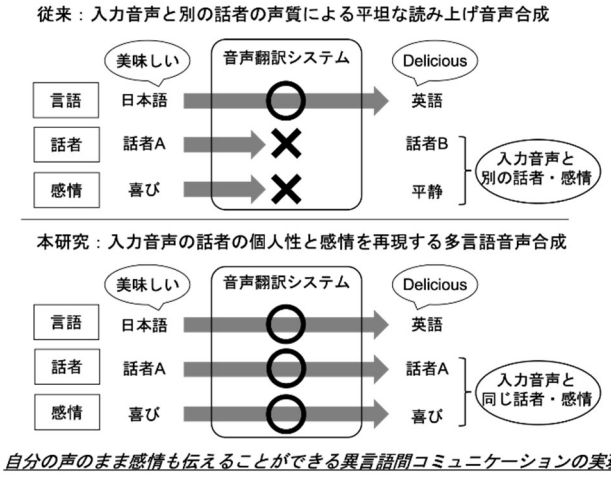


図 1　研究の目的

３．研究の方法

任意の言語・話者・感情の組み合わせの合成音声を生成可能とするためには、言語・話者・感情それぞれに依存する音声の特徴を分離し、さらにそれらを学習データにない組み合わせで音声を合成可能にする枠組みが必要である。そこで、本研究では、深層学習に基づく音声合成において、言語・話者・感情に依存する音声の特徴を分離しながらモデル化することが可能なモデル構造を明らかにする。また、敵対的学習を導入し、言語・話者・感情に依存しない中間表現を獲得することで、それぞれの要素に依存した特徴と依存しない特徴を効果的に表現することが可能なモデル学習法を明らかにする。さらに、異なる言語において入力音声の話者・感情を再現するための補助特徴量として利用者の顔画像や入力テキストを利用する手法を開発する。顔画像やテキストから話者・感情を表す情報を獲得し、異なる言語において話者・感情を再現する。

４．研究成果

(1) 言語依存層と話者コードによる複数言語・複数話者同時モデリング

従来の深層学習に基づく音声合成では、ある単一言語のテキストを入力として音声を出力する音声合成モデルを学習していた。そのため、単一の音声合成モデルは単一の言語のみに対応しており、複数の言語の音声を単一の音声合成モデルから生成することはできなかった。そこで、深層学習に基づく音声合成において言語依存層を持つモデル構造を用いることで複数言語を同時に学習することを可能にした。通常、単一話者は単一言語のみの音声データを収録しているため、本研究では言語依存層と話者コードによる複数言語・複数話者同時モデリングを提案し、様々な言語・話者の音声データを同時にモデル学習に利用可能とした（図 2）。言語依存層では言語ごとにネットワークを切り替えることで、複数言語の学習を可能としている。また、言語非依存層では全言語で共有する構造となっており、隠れ層に話者コードを入力することで、複数話者の学習を可能としている。

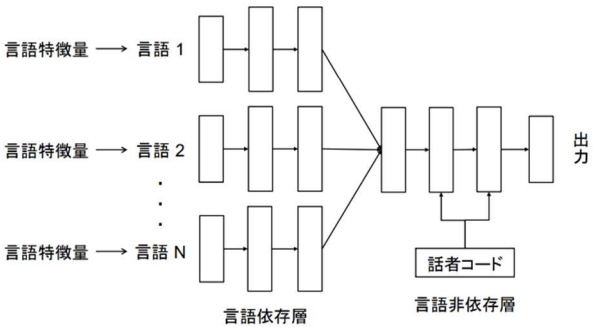


図 2　言語依存層と話者コードによる複数言語複数話者モデリングの概要図

目標話者の言語と異なる言語で音声を生

成する場合には、目標話者の話者コードと異なる言語の言語特徴量を学習済みモデルへと入力する。モデルは指定された話者コードと言語特徴量に合わせて、目標話者の声質のまま異なる言語の音声を生成することが可能となる。

(2) 敵対的学習に基づく多言語音声合成モデルの学習

入力音声の話者の特徴を、より再現するために深層学習に基づく複数言語・複数話者同時モデリングにおいて敵対的学習の枠組みを導入した、敵対的学習に基づく多言語音声合成モデルの学習を提案した[1]。テキストから音声を予測する多言語音声合成モデルを生成器 (Generator) とし、人間が発声した音声かモデルが生成した音声かを識別する識別器 (Discriminator) を導入する。ここでは、テキストから音声を予測する生成器として言語依存層と話者コードを持つモデル構造を用いることで、様々な言語・話者のデータを同時に取り扱う (図 3)。敵対的学習では、入力された音声が人間の発声した音声かモデルが生成した音声かを識別器が識別できなくなるように生成器の学習を行うことで、より人間の発声に近い音声をモデルから生成可能とする。さらに、より話者の特徴を学習するために、話者情報を識別器の入力に加える手法や、話者識別を補助的に行う識別器の導入する手法についても検討した。実験結果から、話者情報を識別機に入力する、または、話者識別を補助的に行う識別器を導入することで、異なる言語においてより話者の特徴を再現することが示された。

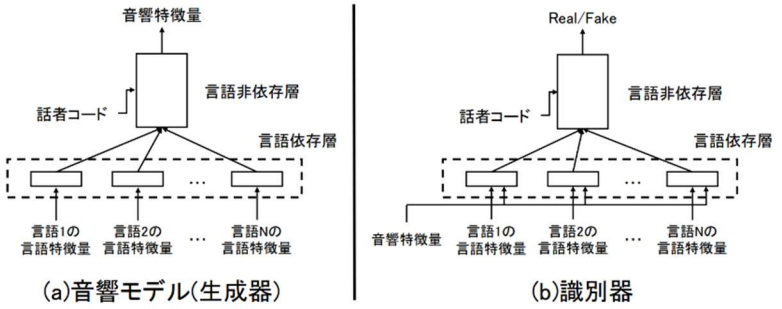


図 3　敵対的学習に基づく多言語音声合成モデルの概要

(3) 変分オートエンコーダに基づく多言語音声合成

入力音声の特徴を出力音声において再現するための枠組みとして、近年注目を集めている変分オートエンコーダ (Variational Autoencoder; VAE)の構造を導入した多言語音声合成を提案した。VAE はエンコーダとデコーダから構成され、音声合成においては、入力音声の声質を表す潜在変数をエンコーダによって獲得し、獲得した潜在変数とテキストをデコーダへと入力することで入力音声の特徴を再現した音声を生成する。提案法では、デコーダに言語依存層を導入することで多言語へと対応する。複数言語・複数話者のテキストと音声のペアデータを学習データとして用いて、VAE に基づく多言語音声合成モデルを学習する。これによって、言語依存の特徴は言語依存層においてモデル化され、話者依存の特徴は言語非依存なエンコーダにおいてモデル化されるため、音声に含まれる特徴のうち言語と話者に依存する特徴を分離することが可能となる。このようにして学習された VAE に基づく多言語音声合成モデルは、ある言語の音声をエンコーダへと入力し、エンコーダへ入力した音声と異なる言語のテキストをデコーダと入力することで、入力音声と異なる言語において入力音声の声質を再現する合成音声が生成される。さらに、敵対的学習を導入することで、より話者の特徴を再現する手法を検討した。実験結果から、言語依存層と話者コードを用いた手法よりも高い合成音声の自然性を示した。また、より少ない音声データから話者の声質を獲得することが可能となった。

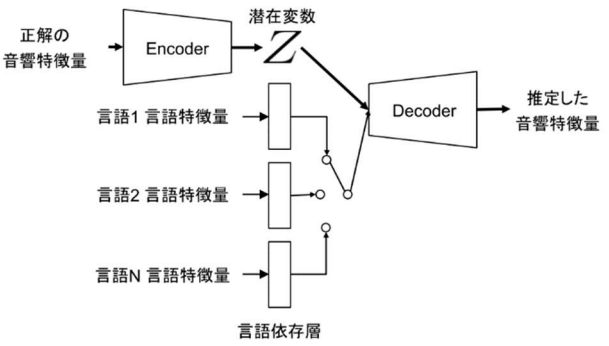


図 4　変分オートエンコーダに基づく多言語音声合成の概要

(4) 異なる話者で感情を再現するためのモデル構造の検討

従来の感情音声合成では、ある特定の話者の様々な感情音声を収録し、テキストと感情音声のペアデータを用いることで、感情音声合成モデルの学習が行われる。本研究では、ある特定の話者には感情音声のペアデータが存在するが、別の話者には感情音声のペアデータは存在せず、ナレーション調音声のペアデータのみが存在するという状況を想定する。ナレーション調音声のデータしか持たない話者の声質で感情音声を生成するために、感情と話者の特徴を分離して学習するためのモデル構造について検討を行った。提案法では、ナレーション調音声を用いた話者

の変換とナレーション調音声から感情音声への 2 段階の変換を実現するようなモデル構造を導入することで、感情音声を収録していない話者の声質のまま感情音声を生成した。

(5) 顔画像を補助特徴として利用した変分オートエンコーダに基づく複数話者音声合成

入力音声のみから音声の特徴を獲得するのではなく、顔画像から音声の特徴を獲得する枠組みを提案した。人間は顔から声の特徴を予測するため、顔画像と声質の間には何らかの相関があると考えられている。そこで、顔画像・テキスト・音声の関係を深層学習によってモデル化する手法を提案した [2]。提案法は、顔画像と音声をそれぞれ変分オートエンコーダ（VAE）によってモデル化する。テキストは音声をモデル化した VAE のデコーダの入力に用いる。顔画像をモデル化した VAE の潜在変数と、音声をモデル化した VAE の潜在変数の間で同一人物の顔画像と音声の場合は共通する潜在変数によって表すという制約を加えることで、顔画像と音声の関係をモデル化することが可能となった。提案法では、顔画像を入力することで、その顔画像から予測される声質の合成音声を生成する。また、提案法は顔画像と音声をそれぞれ VAE によってモデル化するため、顔画像・音声・テキストの 3 つ組のデータだけでなく、顔画像のみ、音声とテキストのペアデータもモデル学習に利用可能となった。主観評価実験では、被験者に顔画像を見せながら、実際にその人物の声質の合成音声と顔画像から予測した声質の合成音声を比較した。実験結果から、提案法は顔画像に対して違和感のない音声の合成を生成可能であることを示した。

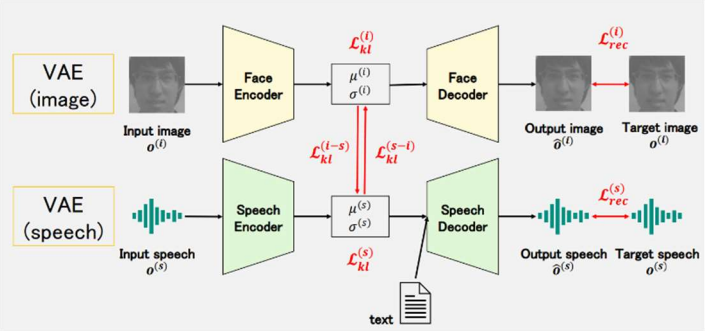


図 5　顔画像を補助特徴として利用した変分オートエンコーダに基づく複数話者音声合成の概要

＜引用文献＞

[1] 大谷眞史, 佐藤優介, 高木信二, 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一, “音声合成における敵対的生成ネットワークを用いた複数言語・複数話者モデリング,” 日本音響学会 2020 年秋季研究発表会講演論文集, pp. 695-696, 2020.

[2] 平光啓祐, 橋本佳, 南角吉彦, 徳田恵一, “深層学習に基づく音声合成における顔画像情報を用いたクロスモーダル話者適応,” 日本音響学会 2022 年春季研究発表会講演論文集, pp. 905-906, 2022.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

| | |
|---|---|
| １．著者名<br>Yukiya Hono, Shinji Takaki, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda | ４．巻<br>9 |
| ２．論文標題<br>PeriodNet: A Non-Autoregressive Raw Waveform Generative Model With a Structure Separating Periodic and Aperiodic Components | ５．発行年<br>2021年 |
| ３．雑誌名<br>IEEE Access | ６．最初と最後の頁<br>137599-137612 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1109/ACCESS.2021.3118033 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

〔学会発表〕　計24件（うち招待講演　0件／うち国際学会　3件）

| |
|---|
| １．発表者名<br>Yukiya Hono, Shinji Takaki, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda |
| ２．発表標題<br>PeriodNet: A non-autoregressive waveform generation model with a structure separating periodic and aperiodic components |
| ３．学会等名<br>2021 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)（国際学会） |
| ４．発表年<br>2021年 |

| |
|---|
| １．発表者名<br>高木信二，牛田光一，橋本佳，南角吉彦，徳田恵一 |
| ２．発表標題<br>因子分析に基づくHSMMを利用した構造化アテンション音声合成 |
| ３．学会等名<br>日本音響学会2021年秋季研究発表会 |
| ４．発表年<br>2021年 |

| |
|---|
| １．発表者名<br>藤本崇人，橋本佳，南角吉彦，徳田恵一 |
| ２．発表標題<br>隠れセミマルコフモデルによる構造化アテンションを用いた自己回帰型VAEに基づくsequence-to-sequence音声合成 |
| ３．学会等名<br>日本音響学会2021年秋季研究発表会 |
| ４．発表年<br>2021年 |

１．発表者名
平光啓祐，橋本佳，南角吉彦，徳田恵一

２．発表標題
深層学習に基づく音声合成における顔画像情報を用いたクロスモーダル話者適応

３．学会等名
日本音響学会2022年春季研究発表会

４．発表年
2022年

１．発表者名
佐々木一匡，吉村建慶，高木信二，橋本佳，南角吉彦，徳田恵一

２．発表標題
声質・声の高さ・話速を変更可能なニューラルボコーダ構成法の検討

３．学会等名
日本音響学会2022年春季研究発表会

４．発表年
2022年

１．発表者名
藤本崇人，橋本佳，南角吉彦，徳田恵一

２．発表標題
HSMM構造化アテンションに基づく音声合成のためのメモリ削減手法

３．学会等名
日本音響学会2022年春季研究発表会

４．発表年
2022年

１．発表者名
法野行哉，高木信二，橋本佳，中村和寛，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
非周期性指標を考慮したニューラルボコーダの学習

３．学会等名
日本音響学会2022年春季研究発表会

４．発表年
2022年

１．発表者名
藤本崇人，橋本佳，南角吉彦，徳田恵一

２．発表標題
学習時と合成時の一貫性を考慮したVAEに基づく自己回帰型sequence-to-sequence音声合成

３．学会等名
日本音響学会2021年春季研究発表会

４．発表年
2021年

１．発表者名
角谷健太，吉村建慶，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
隠れセミマルコフモデルに基づく構造化アテンションを用いたSequence-to-Sequence音声合成

３．学会等名
日本音響学会2021年春季研究発表会

４．発表年
2021年

１．発表者名
法野行哉，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
周期・非周期成分の分離に基づくニューラルボコーダによる音声波形のモデル化の検討

３．学会等名
日本音響学会2021年春季研究発表会

４．発表年
2021年

１．発表者名
岩田康平，高木信二，橋本佳，南角吉彦，徳田恵一

２．発表標題
勾配ブースティング決定木を用いた音声合成手法の検討

３．学会等名
日本音響学会2021年春季研究発表会

４．発表年
2021年

１．発表者名
平光啓祐，橋本佳，徳田恵一，南角吉彦

２．発表標題
深層学習に基づく音声合成における顔画像を用いた話者適応

３．学会等名
第18回情報学ワークショップ

４．発表年
2020年

１．発表者名
久野宏彰，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
音声合成における特徴的な発話スタイルの転移学習

３．学会等名
第18回情報学ワークショップ

４．発表年
2020年

１．発表者名
大谷眞史，佐藤優介，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
音声合成における敵対的生成ネットワークを用いた複数言語・複数話者モデリングの検討

３．学会等名
第18回情報学ワークショップ

４．発表年
2020年

１．発表者名
岩田康平，高木信二，橋本佳，南角吉彦，徳田恵一

２．発表標題
勾配ブースティング決定木を用いた高速な音声合成手法の検討

３．学会等名
第18回情報学ワークショップ

４．発表年
2020年

１．発表者名
Yukiya Hono, Kazuna Tsuboi, Kei Sawada, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

２．発表標題
Hierarchical Multi-Grained Generative Model for Expressive Speech Synthesis

３．学会等名
Interspeech 2020（国際学会）

４．発表年
2020年

１．発表者名
藤本崇人，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
感情音声合成のためのDirichlet VAE

３．学会等名
日本音響学会2020年秋季研究発表会

４．発表年
2020年

１．発表者名
法野行哉，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
DNNに基づく音声ボコーダにおける周期・非周期成分のモデル化の検討

３．学会等名
日本音響学会2020年秋季研究発表会

４．発表年
2020年

１．発表者名
大谷眞史，佐藤優介，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
音声合成における敵対的生成ネットワークを用いた複数言語・複数話者モデリング

３．学会等名
日本音響学会2020年秋季研究発表会

４．発表年
2020年

１．発表者名
Takenori Yoshimura, Shinji Takaki, Kazuhiro Nakamura, Keiichiro Oura, Yukiya Hono, Kei Hashimoto, Yoshihiko Nankaku, and Keiichi Tokuda

２．発表標題
Embedding a differentiable mel-cepstral synthesis filter to a neural speech synthesis system

３．学会等名
ICASSP 2023（国際学会）

４．発表年
2023年

１．発表者名
藤本崇人，橋本佳，南角吉彦，徳田恵一

２．発表標題
半教師あり学習を用いた階層化生成モデルに基づく日本語 end-to-end 音声合成

３．学会等名
日本音響学会2022年秋季研究発表会

４．発表年
2022年

１．発表者名
吉村建慶，高木信二，中村和寛，大浦圭一郎，法野行哉，橋本佳，南角吉彦，徳田恵一

２．発表標題
微分可能なメルケプストラム合成フィルタを組み込んだend-to-end 音声合成システムの検討

３．学会等名
日本音響学会2022年秋季研究発表会

４．発表年
2022年

１．発表者名
石田龍成，藤本崇人，橋本佳，南角吉彦，徳田恵一

２．発表標題
隠れセミマルコフモデルに基づく構造化アテンションを用いた音声合成におけるパラメータ共有構造の検討

３．学会等名
日本音響学会2022年秋季研究発表会

４．発表年
2022年

| |
|---|
| １．発表者名<br>Takato Fujimoto, Kei Hashimoto, Yoshihiko Nankaku, and Keiichi Tokuda |
| ２．発表標題<br>Autoregressive variational autoencoder with a hidden semi-Markov model-based structured attention for speech synthesis |
| ３．学会等名<br>ICASSP 2022 |
| ４．発表年<br>2022年 |

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（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 |
|---|---|