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　 ４ 月 ２０ 日現在

機関番号：１７１０２
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１１９６５

研究課題名（和文）発話動作を起点とした音声生成による代用発声技術の実現

研究課題名（英文）Construction of a substitute speech generation technique based on the input of articulatory motion

研究代表者

鏑木　時彦（Kaburagi, Tokihiko）

九州大学・芸術工学研究院・教授

研究者番号：３０３２５５６８

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では、喉頭摘出者が音声コミュニケーションを維持するための代用発声技術を創出することを目的として、口唇運動から音声波形を生成する合成モデルの検討、ならびにそのモデルを機械学習で実現するための音声コーパス作成を行った。合成モデルは低次元の音声特徴量を求めるエンコーダーと、メルスペクトログラムを推定するデコーダーから構成される。実験の結果、口腔の音響特性に加えて、アクセントやイントネーションを形成するピッチパタンを予測可能であり、十分に了解できる音声を合成できた。並行して、音声からリアルタイムMRIで測定した調音運動を復元するモデルを検討した。

研究成果の学術的意義や社会的意義
喉頭癌などの重度の疾患で喉頭を摘出した場合、その後の一生において日常のコミュニケーションに大きな支障をきたす。喉頭摘出者の代用発声法としては、電気式人工喉頭や食道の粘膜を声帯の代わりに振動させる食道発声などがあるが、それぞれ、抑揚のない機械的な発声になる、胃に空気を取り込むため高齢者では習得が難しいなどの問題がある。超高齢化した社会状況に鑑みても、喉頭疾患によるコミュニケーションの喪失に対処し得る情報技術の創出は不可欠であり、本研究で検討した新しい代用発声技術が意味を持つと考えられる。

研究成果の概要（英文）：In this project, a model for synthesizing speech from motion of the lips was constructed as a tool of substitute speech, that can help laryngectomees maintain voice communication, and a set of Japanese speech corpus was gathered for training the model. The model comprises an encoder, by which low-dimensional speech features are extracted from the motion input, and a decoder, by which mel-spectrogram is estimated as the output. As a result of experiments, the model is capable of estimating, not only the acoustic characteristic of the vocal tract, but also the pitch conture for expressing the accent and intonation. The synthesized speech was intelligible. In addition, a model was studied for estimating the motion of the vocal tract, which was measured using a real time MRI, from speech.

研究分野：音声情報処理

キーワード：音声合成　代用発声　調音運動　口唇動画　機械学習　ニューラルネットワーク

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

2014 年の喉頭ガンの罹患数は約 5,000 人であるが（国立がん研究センターのホームページによる）、このような重度の疾患で喉頭を摘出した場合、その後の一生において声を出すことができなくなり、日常のコミュニケーションに大きな支障をきたす。喉頭摘出者のための代用発声法としては、音源装置を使用する電気式人工喉頭、食道の粘膜を声帯の代わりに振動させる食道発声などがある。しかしながら、電気式人工喉頭は、抑揚のない機械的な発声になるだけでなく、使用者がもともと持っている個人的な声の質が失われてしまう。食道発声は、胃に空気を取り込んで吐き出すため、高齢者では習得が難しい。超高齢化した社会状況に鑑みても、喉頭疾患によるコミュニケーションの喪失に対処し得る情報技術の創出は不可欠であり、本研究が目指す新しい代用発声技術の着想に至った。

２．研究の目的

本研究は、ガンなどの重度の喉頭疾患による発声障害者がコミュニケーションを維持するための代用発声技術の確立を目的とする。喉頭疾患で喉頭を摘出した場合でも、口腔の調音器官（唇や舌など）は維持される。そこで、これらの調音器官の動作から音声を合成することにより、いわば「口パク」することで、音声による意図の伝達を可能とする。自然で明瞭な合成音を得るには、言語の音韻性を表す口腔の共鳴特性と、調音器官の運動とは直接的な関係性の低い音声の音源情報（声の高さ、大きさ、有声・無声の区別など）を、できるだけ正確に予測することが課題になる。本研究では、文節や文章全体にわたる口腔動作の時系列性に着目し、音韻系列に関する情報を間接的に得ることで、より品質の高い音声を合成することを検討する。具体的には、健常者の口唇の運動を音声とともに測定し、得られるデータベースと機械学習（人工知能）を基として、目的とする音声合成法を実現する。

舌は調音器官の中で最も重要であるが、口腔内にある舌の運動を、外部から直接的に目視することはできない。この問題に対して、従来は、マーカーである小型コイルの位置を外部磁界により測定する磁気センサ、下顎にあてたプローブで舌表面を画像化する超音波スキャナ、口腔全体の断層像を得る磁気共鳴画像法（MRI）などを用いて舌の運動を観測し、発話動作から音声を合成する試みがなされてきた。しかしながら、日常的なコミュニケーションにおいて、これらのセンシング装置を使用することは現実的ではない。対照的に、口唇の運動は外部から容易に観測できるが、その反面、主要な調音器官である舌に比べて、口唇がもたらす調音情報はきわめて限定的になる。本研究のチャレンジは、そのような限定的な調音情報から、音源特性も含めて音声を合成するところにある。

口腔の調音運動から音声を合成する研究においては、当初、調音運動と関係の深い口腔の共鳴特性のみを推定する研究が行われた。その後、機械学習の進展により、調音運動から音源のパラメータをも予測することで、音声波形を総合的に生成できる可能性が開けたため、代用発声法への応用を計画するに至った。調音運動データとしては、当初はコンパクトで扱いやすい磁気センサデータが多く使用され、比較的単純な構造を持つ多層パーセプトロンやリカレントニューラルネットワークによって合成音の生成が試みられた。近年では、ディープニューラルネットワークが大規模になるとともに、音声合成・音声認識や自然言語処理の分野で時系列データを扱う手法が発展したため、本研究では先端的なエンコーダー・デコーダーモデルを応用し、口唇の運動という最小限の調音情報から了解可能な音声を合成する。

最先端の機械学習技術の応用によって、これまでの代表的な代用発声法である電気式人工喉頭、食道発声、シャント発声の問題点を解決し、訓練や手術などの大きな負担がなく、ユーザーが自分自身の声質でコミュニケーションを維持できるようにすることが成果目標である。さらに、発声障害や構音障害に対処するための技術としては、電気式人工喉頭においてピッチの抑揚を制御するための改良、テキスト音声合成を利用した支援技術のほか、障害者の発した音声や電気式人工喉頭による音声を、より聞き取りやすくなるように変換・改善する方法が検討されている。この声質変換法は、周囲の騒音や部屋の残響などに影響されやすく、実環境での利用には限界がある。一方、本研究で検討する音声合成法は、口腔の動作入力に基づくものであり、このような音響環境の問題を回避できる。

３．研究の方法

通常の音声発話では、肺からの呼気をエネルギー源として、声帯の振動などによって音源波が作られる。声の高さや大きさは、基本的にこの音源波によって定められる。さらに、音源波が口腔内を伝わる際に、それぞれの言語音に特有の共鳴特性が付与され、言葉の意味を伴った音声が生成される。従って、ガンなどの喉頭疾患によって喉頭を摘出した場合には、音源波そのものを生成することが不可能になり、音声によるコミュニケーションに支障をきたす。

本研究では、喉頭疾患においても機能が保たれる口腔の調音器官、特に口唇に着目し、通常の音声発話と同様に口唇を動かすことで音声波形を生成する、代用発声技術を検討する。上記から明らかなように、音声の合成では音源波の生成と口腔の共鳴作用とを正確に模擬する必要があ

る。口腔の共鳴特性はその形状と関係があり、さらに口腔の形状は調音器官の運動によって決まる。従って、口唇の運動から口腔の共鳴特性を予測することは、ある程度可能と考えられる。一方、音源波は声の高さ（ピッチ）を決定し、アクセントやイントネーションを形成するため、了解性の高い音声を合成する上で非常に重要であるが、喉頭内の声帯の挙動は、口唇の運動とは直接的な関係が存在しない。従って、この問題をいかに解決するかが、本研究においてきわめて重要になる。この問題に対して、本研究では、文節や文章全体にわたる口腔動作の時系列性に着目する。このような長期の口腔動作には、音声発話時の言語的内容に対応した音韻系列の情報が、何らかの形で反映されるはずである。さらに、音韻系列の情報が与えられれば、そこから単語のアクセントを表出するピッチの時間変化や有声・無声の区別など、各種の音源情報を推定できる可能性がある。そこで本研究では、健常者の口腔動作を音声とともに測定し、得られる音声データベースと機械学習（人工知能）とを基として、口腔動作から音声波形を生成する音声合成法を実現させる。

本法では、口唇の運動を撮像した動画を入力、メルスペクトログラムを出力とするディープニューラルネットワークを構成し、さらにメルスペクトログラムを音声に変換するディープニューラルネットワークによって波形生成を行う。後者はテキストからの音声合成などでも使用される既存のモデルを用いるため、前者のモデル構築が研究の主眼となる。入力と出力はいずれも時系列データであるため、本研究では自然言語処理の分野で飛躍的な発展を遂げたトランスフォーマーで使用されている、エンコーダー・デコーダー型のネットワーク構造を用いる。エンコーダーは、文章全体の時系列性、すなわち口唇運動の時間的連続性や前後の関係に考慮しつつ、音声合成において有効となる、冗長性を排除した本質的な特徴量を抽出する。デコーダーは、この特徴量からメルスペクトログラムを推定するものであり、多くの場合、再帰的構造を有する。すなわち、過去のメルスペクトログラムの推定値を参照しつつ、現時点での推定を効率的に行う。この処理を文章全体について適用することで、音声合成に必要なメルスペクトログラムを得ることができる。ディープニューラルネットワークに含まれる多数のパラメータを学習するには、口唇の運動と音声を同時に収録した大規模パラレルコーパスを用いる。最終的に、音声合成システムを種々の方法で評価することによりその有効性を検証する。

４．研究成果

(1) 機械学習のためのパラレルコーパスの構築

口唇動画から音声のメルスペクトログラムを推定するニューラルネットワークを機械学習によって構成するため、健常者の顔を正面から撮像した動画データならびに同時録音による音声データを収集し、日本語のパラレルコーパスを構築した。発話者は女性のプロ・ナレーター１名である。発話文の選定においては、日本語の各音素を可能な限り前後の音素分脈を考慮する必要がある。これは、音素分脈によって調音器官の運動が影響を受ける調音結合現象が生じるためである。本研究では、既往研究で開発された短文章の発話リストとともに、音素分脈を考慮して独自に作成した短文章を用い、合計 3887 文章、時間長として約 4.8 時間におよぶパラレルコーパスを収録して音声合成に用いた。

一方、口唇動画では重要な調音器官である舌の情報が欠落している。従って、舌を含む調音器官全体の音声発話時の運動を測定し、基礎的データとして蓄積することは重要であると考えられる。そこで本研究では、口唇動画による情報を補うため、MRI（磁気共鳴画像）を用いて頭部の正中断面における断層イメージをリアルタイムで収録した。発話者は男性 1 名、女性 2 名の合計 3 名、発話内容は音素バランスの考慮された ATR503 文である。これらの文章は上記のパラレルコーパスにも含まれるため、相互参照が可能である。収録は ATR Promotions 社の脳活動イメージングセンタで 3 テスラの MRI 装置を有償で使用して行った。このリアルタイム MRI のフレームレートは毎秒約 27 フレームであり、発話時の母音・子音の調音運動を十分な時間分解能で高精度に撮像することができた。

(2) ベースライン合成システムの構築

次に、口唇動画から音声を合成するベースラインシステムの構築を行った。本システムは、自然言語処理の分野で使用されるトランスフォーマを参考にしている。システムは動画情報を圧縮する残差ネットワーク、エンコーダおよびデコーダからなる。アテンション機構を含むエンコーダにおいて口唇動画から特徴マップを推定し、デコーダでは再帰構造を取り入れてメルスペクトログラムを推定する。より詳細には、デコーダーでは Gated linear unit (GLU) block を積層することで因果的畳み込みをおこなう。本システムは再帰構造を含むため、前時刻に予測したメルスペクトログラムを全結合層で構成されるプリネットを通してデコーダーに入力するとともに、デコーダー出力は非因果的畳込み層からなるポストネットによって、メルスペクトログラムへの誤差の累積を低減させた。提案したネットワークを学習するための損失は、ポストネットに入力する前のデコーダー出力と正解となるメルスペクトログラムとの平均自乗誤差と、ポストネットの出力と正解となるメルスペクトログラムの平均自乗誤差の和である。学習データには前述のパラレルコーパスを用いた。学習を効率化させるため、スケジュールドサンプリングを適用した。客観評価指標による本システムの評価では、PESQ が 1.25、STOI が 0.63、ESTOI が 0.54 などの結果を得た。合成音は、発話内容を聞き取れる程度の明瞭性を確認した。

(3) 合成モデル構築における敵対的生成ネットワーク(GAN)の応用

上記のベースラインシステムは自己回帰的、再帰的な予測を基にしており、長時間の音声を生成する際には予測誤差が累積しやすく、また、スペクトル、ピッチなどの音響特徴量が時間的に平坦化されやすいといった問題がある。特に、自己回帰モデルではスケジュールドサンプリングを用いた学習時と推論時のスペクトル生成手順が異なるため、この学習法が推論時の誤差に影響を与える可能性がある。そこで、この課題を解決するため、GAN によるモデル学習を検討した。識別器を導入して予測したメルスペクトログラムを入力し、スケジュールドサンプリングで生成したものか、free running で生成したものかを識別した。客観評価実験の結果、GAN の導入については、ピッチ曲線の変化幅などに改善が見られたが、評価指標の値そのものについては改善されず、さらなる精度向上のための検討が必要となった。

(4) 学習データが少量である条件下での合成モデル構築に関する検討

最終年度において、口唇動画から音声波形を生成する合成モデルに関して、ベースラインシステムを高度化するための検討を行った。本法は機械学習に基づくため、口唇動画と音声のパラレルデータセットが大量に必要になる。現時点で得られているデータセットは、本研究課題において収集した約 4000 個の短文章からなり、日本語のデータセットとしては比較的大規模と言えるものの、海外において英語音声を対象に収集されたデータセットと比較するとまだ十分とは言えない。モデル学習用のデータセットの品質と規模は合成音の品質に直結するため、データ量が少ない条件下での有効な合成モデルあるいは学習法を確立することは重要である。

本研究の合成モデルは自然言語処理で用いられるトランスフォーマーを基としており、エンコーダーとデコーダーから構成される。テキストと音声からなるパラレルデータは既存の大規模なオープンリソースが利用できることをもとに、これらのエンコーダー、デコーダーを別手法で構成し、転移学習によって口唇動画音声合成を実現する検討を行った。エンコーダーについてはテキスト音声合成、デコーダーについては音声スペクトルを自己復元するネットワークを学習し、転移学習した。さらに、中間表現である特徴量をベクトル量子化によって離散化することを試みた。転移学習とベクトル量子化の効果を種々のデータ量に対して比較し、それぞれの有効性を明らかにした。

(5) リアルタイム MRI を用いた音声からの調音運動の推定

口唇動画からの音声合成と並行して、音声から調音運動を推定する逆推定法の検討を行った。先に(1)で述べたリアルタイム MRI は、頭部の正中断面における調音器官の時系列的な形状変化に関する詳細な情報を取得できる。音声からこの調音運動情報を推定することで、発話トレーニングなどへの応用が期待される。そこで、このタスクを実行するニューラルネットワークを教師あり学習によって実現した。ネットワーク構造は全結合層と時系列処理に適する LSTM（長期短期メモリ）層の積層である。本モデルを高解像度の画像データの復元に適用した際には、画像データ中の高周波成分に対する推定性能が不十分になることが考えられる。そこで本研究では、リアルタイム MRI における解像度向上が起因となる推定性能の低下に対処するため、上記のベースラインモデルに加えて、画像情報の特徴抽出によって情報を有効に圧縮する畳み込みオートエンコーダと、推定後の画像の高解像度化を行う SRCNN を組み合わせたモデルを検討し、それらの有効性を検証した。

(6) まとめと今後の展望

本研究では、喉頭摘出者が音声コミュニケーションを維持するための代用発声技術を創出することを目的として、口唇運動から音声波形を生成する合成モデルの検討、ならびにそのモデルを機械学習で実現するための音声コーパス作成を行った。並行して、新規性の高い観測技術であるリアルタイム MRI による調音運動の観測と、音声からの調音運動推定モデルを検討した。これらの研究は、音声を発話する際の調音運動と、その結果として生じる音声波形の間の相互変換が可能であることを示している。特に、口唇運動からアクセントやイントネーションを形成するピッチパタンを予測できることは興味深い結果であると言える。今後の課題としては、日本語のより大規模なパラレルコーパスを構築すること、任意の話者においてより少量のコーパスから良質な合成音を生成すること、感情表現などのパラ言語情報に対応した合成音を生成することなどが挙げられる。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Kato Hikari, Lee Yogaku, Wakamiya Kohei, Nakagawa Takashi, Kaburagi Tokihiko	４．巻 -
２．論文標題 Vocal Fold Vibration of the Whistle Register Observed by High-Speed Digital Imaging	５．発行年 2023年
３．雑誌名 Journal of Voice	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） 10.1016/j.jvoice.2023.08.026	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 Kaburagi Tokihiko, Kuroki Chiho, Hidaka Shunsuke, Ishikawa Satoshi	４．巻 44
２．論文標題 Numerical method for analyzing steady-state oscillation in trumpets	５．発行年 2023年
３．雑誌名 Acoustical Science and Technology	６．最初と最後の頁 269～280
掲載論文のDOI（デジタルオブジェクト識別子） 10.1250/ast.44.269	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 Shunsuke Hidaka, Yogaku Lee, Moe Nakanishi, Kohei Wakamiya, Takashi Nakagawa, Tokihiko Kaburagi	４．巻 -
２．論文標題 Automatic GRBAS Scoring of Pathological Voices using Deep Learning and a Small Set of Labeled Voice Data	５．発行年 2022年
３．雑誌名 Journal of Voice	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） 10.1016/j.jvoice.2022.10.020	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 鏑木時彦	４．巻 77
２．論文標題 磁気共鳴画像(MRI)を用いた管楽器吹奏時の声道計測	５．発行年 2021年
３．雑誌名 日本音響学会誌	６．最初と最後の頁 572-579
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 無
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10件（うち招待講演　0件／うち国際学会　1件）

1．発表者名 藤田直明，南汰翼，鏑木時彦
2．発表標題 転移学習を用いた少量データからの口唇動画音声合成
3．学会等名 日本音響学会春季研究発表会
4．発表年 2024年

1．発表者名 南汰翼，藤田直明，鏑木時彦
2．発表標題 自己回帰及び非自己回帰モデルによる口唇動画を用いた音声合成
3．学会等名 日本音響学会秋季研究発表会
4．発表年 2023年

1．発表者名 加藤日花里，李庸學，鏑木時彦，若宮幸平
2．発表標題 高速度ディジタル撮像を用いたボーカルフライ声区における声帯振動の分析
3．学会等名 日本音響学会秋季研究発表会
4．発表年 2023年

1．発表者名 鏑木時彦，加藤日花里，李庸學
2．発表標題 発声における仮声帯振動の影響に関する数値流体解析
3．学会等名 日本音響学会秋季研究発表会
4．発表年 2023年

１．発表者名
Shunsuke Hidaka, Kohei Wakamiya, Tokihiko Kaburagi

２．発表標題
An investigation of the effectiveness of phase for audio classification

３．学会等名
IEEE ICASSP 2022（国際学会）

４．発表年
2022年

１．発表者名
南汰翼，藤田直明，鏑木時彦

２．発表標題
自己回帰及び非自己回帰モデルによる口唇動画を用いた音声合成

３．学会等名
日本音響学会九州支部　学生のための研究発表会

４．発表年
2022年

１．発表者名
藤田直明，南汰翼，鏑木時彦

２．発表標題
TransformerとGANを用いた口唇動画音声合成

３．学会等名
日本音響学会春季研究発表会

４．発表年
2023年

１．発表者名
江崎蓮，鏑木時彦

２．発表標題
系列変換モデルを用いた口唇動画・音声変換システムに関する研究

３．学会等名
日本音響学会九州支部 学生のための研究発表会

４．発表年
2021年

１．発表者名
江崎蓮，鏑木時彦

２．発表標題
系列変換モデルを用いた口唇動画からの複数話者音声合成

３．学会等名
日本音響学会春季研究発表会

４．発表年
2022年

１．発表者名
日髙駿介，若宮幸平，鏑木時彦

２．発表標題
音分類課題において有効な位相情報の表現に関する検討

３．学会等名
日本音響学会秋季研究発表会

４．発表年
2021年

〔図書〕　計1件

１．著者名	４．発行年
滝口哲也（編著）鏑木時彦他（著）	2021年
２．出版社	５．総ページ数
コロナ社	309
３．書名	
音響学講座　音声（上）	

〔産業財産権〕

〔その他〕

九州大学研究者情報　鏑木時彦
https://hyoka.ofc.kyushu-u.ac.jp/search/details/K002357/index.html

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