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：３２６６０
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１２０２４

研究課題名（和文）調音運動データベースの構築とデータベース間の正規化および調音運動ベースの音声合成

研究課題名（英文）Construction of articulatory movement database, normalization of databases, and speech synthesis based on the database

研究代表者
桂田　浩一（Katsurada, Kouichi）

東京理科大学・理工学部情報科学科・教授

研究者番号：８０３２４４９０

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）：(1)EMAデータからの音声合成，(2)rtMRIデータからの音声合成，(3)調音運動データの収録，についてそれぞれ研究を実施した．(1)については，LSTMとD-vectorによる話者識別器を用いた多人数話者用の音声合成器を構築し，特に話者closeの合成については十分な合成音が生成できることを確認した．(2)はtransposed convolutionによる時系列データの補間を用いた合成器を構築し，ストライドサイズを大きくして補間強度を強くしたときに合成音の品質が向上することを確認した．(3)については7名分の調音運動データの収録を終えており，そのうち1名についてIPAの付与が完了した．

研究成果の学術的意義や社会的意義
本研究によって，舌や唇の動きを表す調音運動から音声が良好に生成できることが確認できた．収録方法の異なる2種類の調音運動データ（EMA，rtMRI）の双方で生成できることを確認しており，当該分野の研究進展に微力ながら貢献できたと考えている．調音運動のデータは一般的に収録が困難ではあるが，本研究で日本語用の調音運動データを収録することによって，音声学や音声情報処理の研究分野において調音運動データを利用することが可能になった．これにより，音声学および音声情報処理の発展に多少なりとも寄与できたと考えている．

研究成果の概要（英文）：We developed (1) a speech synthesis system from EMA data, (2) a speech synthesis system from rtMRI data, and built (3) an articulatory movement database using EMA. The speech synthesis system from EMA data is constructed for multiple speakers using LSTM and D-vector, and we confirmed it can generate sufficient synthesized sounds, especially for speaker-close synthesis. For speech synthesis from rtMRI data, we used transposed convolution which interpolates time series data, and the results showed the quality improved when the stride size is increased. As for articulatory database, we have completed the recording of articulatory movement data for seven persons, and IPA assignment has been completed for one of them.

研究分野：音声情報処理

キーワード：EMA　調音運動　音声合成　rtMRI

１．研究開始当初の背景

音声合成の分野では，WaveNet の登場によって声質が劇的に向上し，Tacotron2 によって音質が人間の声と差異の無いレベルにまで至っている．これらの合成方式はテキストから音声を直接合成する方式を採っていることから，end-to-end の音声合成方式と呼ばれる．これに対して，人間が行う発音過程を模した合成方法も検討されている．その一つが調音運動ベースの音声合成法である．この音声合成法では，EMA (Electromagnetic Articulography)をはじめとする機器で調音運動を収録し，その動作から音声を合成する．この方式の利点は人間ならではの調音運動の相違や変化をモデル化可能な点にある．例えば，ある話者が調音の異なる他言語を話した場合の音声や，高齢になり調音運動が正確に行えなくなった人物の音声，風邪をひいた時の鼻声等を正確にシミュレート可能になる．これにより，内部がブラックボックスである end-to-end の合成方式では難しい，人間の発話で実際に生じ得る微妙な変化や違いを合成音において表現可能になる．


EMA による調音運動の収録

しかし，調音運動の収録でしばしば利用される EMA は特殊な機器であるため，現状で日本語調音運動データベースは公開されていない．また，音声合成で調音運動を利用するには発音ラベルが付与されている必要があるが，ある音素は実際には様々な調音方法で発音される（例えば日本語の子音/h/は単音[h]，[ç]，[ɸ]等で発音される）ため，調音運動に基づく詳細な単音ラベリングがされている事が望ましい．このような背景を踏まえて，本研究では正確な単音ラベルを付与した日本語調音運動データベースを構築することに加え，調音運動ベースの音声合成の各研究も並行して行うことにより，高品質な音声合成器を開発することを目標とする．

２．研究の目的

(1) 日本語調音運動の収録，単音ラベリングとデータベースの構築

現状で最も利用されている調音運動データベースはエジンバラ大学の Korin Richmond 講師が構築した mngu0 である．このデータベースは EMA の他にも MRI で計測したデータを含む多様なデータを含んでいるものの，一人の話者のデータのみからなっており，また音声に対するラベリングは音声認識器の出力結果を用いている．このため，例えば英語の音素/r/に対する単音[ɹ]や[ɻ]のように，同じ音素が別の調音方法で発音され得る場合にラベリングが不正確な可能性がある．また，現在公開されている調音運動データベースには mngu0 の英語をはじめ，フランス語，中国語などが存在するが，日本語のデータベースは公開されていない．そこで本研究では多人数（少なくとも 4 話者）が同一文を発話した日本語データを収録し，それらに音声学の専門家である中央大学の牧野教授が手作業で単音ラベリングを行うことにより，多人数の正確なラベリングを行ったデータベースの構築を目指す．日本語のデータベースを構築することにより，日本語ならではの音素の調音運動のモデル化など，日本語の音声研究分野での活用が望める．

(2) 調音運動ベースの音声合成に関する研究

調音運動を用いた音声合成，およびその周辺研究は概ね次のテーマに分類される．

① テキスト（音素や単音の系列）から調音運動系列を生成
② 調音運動系列から合成音声を生成
③ 音声から調音運動系列を生成（②の逆変換）

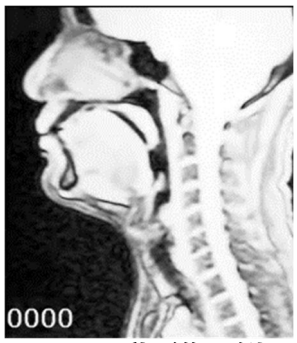
本研究ではまず②の研究に取り組む．②の研究では，深層学習を用いた研究が盛んであり，Bi-directional LSTM を用いたものが良好な成果を収めている．本研究では Wavenet や Transformer Model のような生成モデルを用いることで，より高品質な合成モデルの構築を目指す．その後，③および①に取り組む．③は②の逆変換に相当するため，②と同様に Wavenet や Transformer モデルを用いることを検討すると同時に，AutoEncoder のようなモデルで②と③の双方向変換を実現することも検討する．①については，従来法では HMM 等の古典的なモデルが用いられている事が多いが，本研究ではここまでに挙げた深層学習の各技術を用いてモデル化することを検討している．

３．研究の方法

研究代表者は 2019 年 4 月から 9 月末までエジンバラ大学に在外研究員として滞在した．エジンバラ大学は音声合成研究分野において著名な研究機関であり，音声合成器の性能を評価するための最も大きなコンペティションである Blizzard Challenge を主催する Simon King 教授の他，本研究と関係が深い mngu0 データベースを構築した Korin Richmond 講師も在席している．研究代表者は在外研究の期間に(2)②の調音運動系列からの合成音声の生成の研究に着手し，Richmond 講師を始めとするエジンバラ大学の教員と共に研究を実施した．研究代表者の帰国後，2020 年度からは大学院生の協力の下，(2)①のテキスト（音素や単音の系列）からの調音運動系列の生成について検討した．(2)③の「音響特徴量⇒調音運動」の生成については，2021 年度に大学院生と予備的な検討を行った．

また，EMA からの音声合成に加えて，2020 年度から rtMRI により収録された調音運動の動画像を入力とする音声合成法についても検討を始めた．rtMRI は体内の水素原子を計測することができる MRI 機器を体内組織の時系列動画像の収録に応用したもので，正中矢状面の声道断面を収録することにより，右図のように調音器官全体の断面図を取得できる．この rtMRI で収録した調音運動を入力とし，音響特徴を生成する深層学習器を構築し，その有用性を確認した．

(1)の調音運動データベースの構築については，研究協力者である九州大学の鏑木教授，および研究分担者である九州大学の若宮助教と共に調音運動の収録作業を実施した．収録の対象者はナレーター・アナウンサーである．収録したデータは研究分担者である中央大学の牧野教授が単音ラベリングを行った．

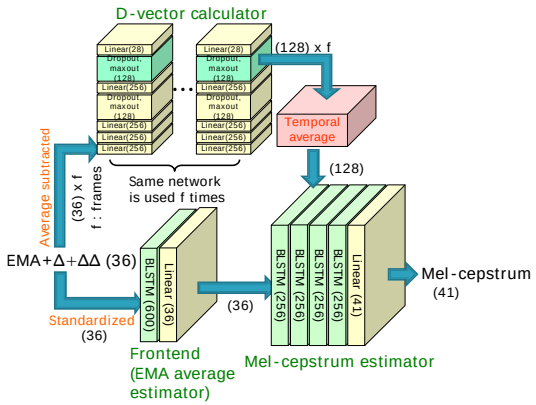


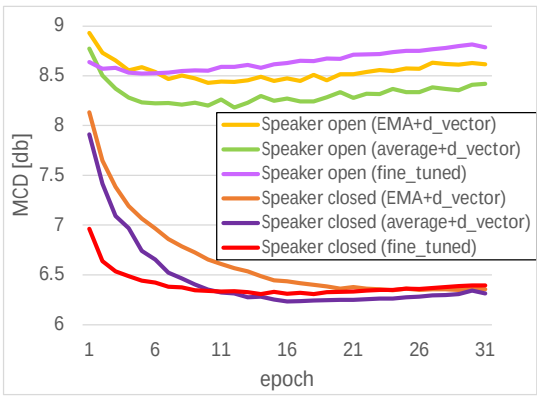
rtMRI 動画像の例

４．研究成果

(1) EMA データからの音声合成に関する成果

前処理のネットワークとして EMA データから平均トライフォンを出力するネットワークを導入し，実際の平均トライフォンを求める代わりに用いる方法を検討する．右図にシステムの構成を示す．システムは平均トライフォンを出力するフロントエンドネットワークとメルケプストラム推定器，d-vector 算出器から構成される．フロントエンドネットワークはノード数 600 の BLSTM と一つの線形結合層から構成される．フロントエンドネットワークの学習には平均トライフォンを用い，学習時のメルケプストラム推定器への入力にも平均トライフォンを用いた．これらのネットワークを個別に学習した後に，結合してファインチューニングを行った．右図に本ネットワークを用いた時のメルケプストラム推定の結果を示す．赤いラインが示す通り話者クローズテストでは従来手法の結果を若干上回ることができたが，紫のラインの通り話者オープンテストでは良好な結果を得ることができなかった．これはフロントエンドネットワークの性能が不十分であったことが原因であると考えられる．フロントエンドネットワークの性能が十分であれば，緑のラインと同様の結果を得られると考えられる．

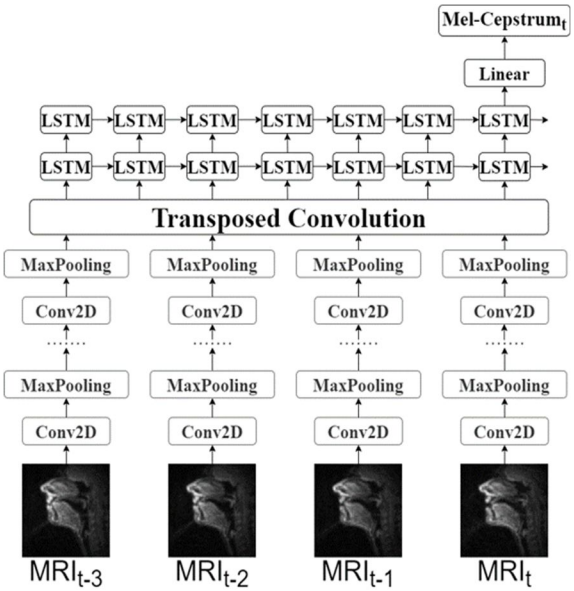




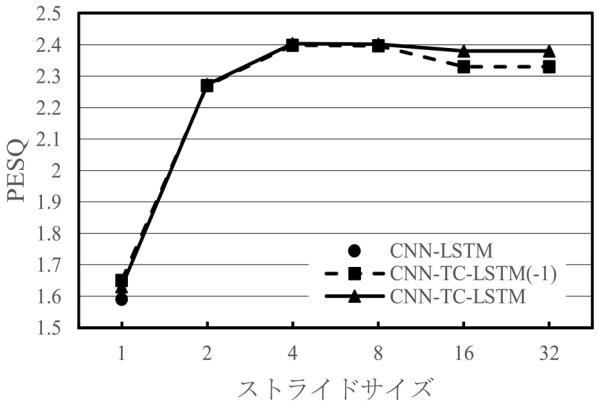
(2) rtMRI データからの音声合成に関する成果

提案するネットワークは基本的に Csapó の CNN-LSTM モデルの CNN と LSTM の間に転置畳み込みニューラルネットワークを挿入した形になっている．これに加えて，CNN の畳み込み層（フィルタサイズ 3×3，ストライド 1×1，フィルタ数 32）とマックスプーリング層（フィルタサイズ 2×2，ストライド 2×2）を 1 層追加し，4 層にした．これにより rtMRI 画像の比較的大域的な情報をより深く学習できるようにし，転置畳み込みニューラルネットワークに送られ

るデータの次元圧縮も実現している．また，このモデルでは LSTM の後の 2 層の全結合層をなくしており，ネットワークから直接メルケプストラムを出力する形にしている．これらの工夫によって精度の向上とネットワーク全体のパラメータ数の削減を実現している．このモデルを本研究では CNN-TC-LSTM と呼ぶことにする．右図に CNN-TC-LSTM のネットワーク構成を示す．



メルケプストラム歪みを右下図に示す．図に示す通り，ストライドサイズを大きくするとメルケプストラム歪みが若干大きくなる傾向にあるものの，提案手法である CNN-TC-LSTM の結果は CNN-LSTM と比較すると良好であることが分かる．ストライドサイズが 1 の時を比較と，CNN-TC-LSTM モデルは CNN-LSTM モデルと比較してメルケプストラム歪みを ATR503 文で 0.41dB 低減できていることが分かる．ストライドサイズが 1 の時の CNN-TC-LSTM と CNN-TC-LSTM(-1)の結果を比べると，CNN-TC-LSTM の方が若干メルケプストラム歪みを低減できているものの，CNN-LSTM と比較すると，共にメルケプストラム歪みを大幅に改善できていることが分かる．CNN-LSTM と CNN-TC-LSTM (-1)の違いは転置畳み込み層の有無である．転置畳み込み層ではフィルタリングとベクトルの加算の処理だけしか行われていないにもかかわらず，メルケプストラム歪みはこれら二つのネットワークで大きく異なる．フィルタリングに相当する処理はネットワーク内の他の層でも行われているため，この差はベクトルの加算によって生じたと考えられる．この加算の処理は近接した時系列データの関係をモデル化していると見做せる．CNN-LSTM モデルにおいても LSTM で時系列情報を扱っているが，転置畳み込みネットワークでは過去数フレーム分に限定した関係を捉えて現在の出力に反映している．我々はこの差がメルケプストラム推定の大きな差に繋がったと考えている．これを確かめるため，CNN-TC-LSTM(-1)から LSTM を取り除いたモデルで実験を行ったところ，メルケプストラム歪みは 6.78dB となり，CNN-LSTM よりも精度が良かった．この結果から，時間的に連続する数フレームの rtMRI 画像を明示的に処理することが精度の良いメルケプストラムの推定に重要であると分かった．



(3) 調音運動データベースの構築に関する成果

EMA 収録装置を所有する若宮助教および九州大学の鏑木教授の協力の下で 2018 年度より収録を継続しており，2022 年度末の時点で 7 名分の EMA データおよび音声の収録が完了している．また，そのうち 1 名の音声に対して IPA ラベリングが完了している．話者によって EMA 機器のコイルの剥離状況が異なるため，収録したデータを精査した上で，今後公開の可否を判断する予定である．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計17件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 Takeshi Koshizuka, Hidefumi Ohmura, Kouichi Katsurada
２．発表標題 Fine-tuning pre-trained voice conversion model for adding new target speakers with limited data
３．学会等名 InterSpeech2021（国際学会）
４．発表年 2021年

１．発表者名 Ryo Tanji, Hidehumi Ohmura, Kouichi Katsurada
２．発表標題 Using Transposed Convolution for Articulatory-to-Acoustic Conversion from Real-Time MRI Data
３．学会等名 InterSpeech2021（国際学会）
４．発表年 2021年

１．発表者名 丹治 涼, 澤田 隼, 大村 英史, 桂田 浩一
２．発表標題 転置畳み込みニューラルネットワークを用いたrtMRIデータからの調音-音響変換
３．学会等名 言語資源活用ワークショップ発表論文集, vol.6
４．発表年 2021年

１．発表者名 飯山 智晴, 澤田 隼, 大村 英史, 桂田 浩一
２．発表標題 IPA を介した音素－調音データ変換のためのIPA 継続長推定手法の検討
３．学会等名 日本音響学会2021年秋季研究発表会
４．発表年 2021年

１．発表者名
Kouichi Katsurada, Korin Richmond

２．発表標題
Speaker-Independent Mel-Cepstrum Estimation from Articulator Movements Using D-Vector Input

３．学会等名
InterSpeech2020（国際学会）

４．発表年
2020年

１．発表者名
Yuta Ogura, Hidefumi Ohmura, Yui Uehara, Satoshi Tojo, Kouichi Katsurada

２．発表標題
Expectation-based parsing for Jazz Chord sequences

３．学会等名
SMC2020（国際学会）

４．発表年
2020年

１．発表者名
池上 凌, 大村 英史, 桂田 浩一

２．発表標題
Cycle-Consistency を利用したマルチモーダル音声強調システムの各種ノイズに対する効果の検証

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
飯山 智晴, 大村 英史, 桂田 浩一

２．発表標題
BLSTM を用いた音素－調音変換

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
越塚 毅，大村 英史，桂田 浩一

２．発表標題
事前学習したvq-wav2vecの音声特徴表現を用いたボコーダフリーのAny-to-Many音声変換

３．学会等名
電子情報通信学会技術報告

４．発表年
2021年

１．発表者名
丹治 涼，大村 英史，桂田 浩一

２．発表標題
real-time MRI 動画像を用いた音声合成システムの作成

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
飯山 智晴，大村 英史，桂田 浩一

２．発表標題
IPA を介した音素－調音データ変換のための音素－IPA 変換手法の検討

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
若宮幸平，田口史朗，渡辺莉子，桂田浩一，牧野武彦，鏑木時彦

２．発表標題
大規模日本語調音・音声パラレルデータの収集

３．学会等名
電子情報通信学会技術報告vol．119，no．80，SP2019-2，pp．7-12

４．発表年
2019年

１．発表者名
池上 凌，大村 英史，桂田 浩一

２．発表標題
マルチモーダル音声強調に対するCycle-Consistencyの導入の検討

３．学会等名
日本音響学会2020年春季研究発表会，3-P-3（2020-3）

４．発表年
2020年

１．発表者名
柴宮 怜，大村 英史，桂田 浩一

２．発表標題
StarGAN-VCモデルにおける潜在表現への制約の有効性について

３．学会等名
日本音響学会2020年春季研究発表会，3-P-39（2020-3）

４．発表年
2020年

１．発表者名
深井 健大郎，大村 英史，桂田 浩一，平田 里佳，入部 百合絵，付 明川，田口 亮，新田 恒雄

２．発表標題
音声想起脳波からの言語表象抽出と音節認識

３．学会等名
電子情報通信学会技術研究報告，SP2019-28（2019-10）

４．発表年
2019年

１．発表者名
深井 健大郎，大村 英史，桂田 浩一，新田 恒雄

２．発表標題
音声想起時脳波中の音節識別について

３．学会等名
人工知能学会第33回全国大会，3P3-0S-20-04（2019-6）

４．発表年
2019年

１．発表者名 桑原 健太，大村 英史，桂田 浩一
２．発表標題 Universal Transformerを使用した対話破綻検出
３．学会等名 人工知能学会第33回全国大会，4J3-J-13-01（2019-6）
４．発表年 2019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	牧野　武彦 (Makino Takehiko) (00269482)	中央大学・経済学部・教授 (32641)	
研究分担者	若宮　幸平 (Wakamiya Kohei) (70294999)	九州大学・芸術工学研究院・助教 (17102)	

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