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　　元 年　　５ 月 ２３ 日現在

機関番号：３２６６０
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００２３４

研究課題名（和文）調音運動HMMとLSPデジタルフィルタを用いた音声合成

研究課題名（英文）Speech synthesis based on articulatory movement HMM and LSP digital filter

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・准教授

研究者番号：８０３２４４９０

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：人間の発声メカニズムに近い音声合成を実現するために，発話時の舌や唇の動作に基づいた合成方法を検討してきた．期間の前半では，口唇や舌の実際の動作ではなく，それらをパラメータ化したデータから音声を合成することを試みたが，より実際の人間の動作に近い音声合成を実現するために，期間の後半では口唇や舌の動作を収録するしてデータベースを作成することに取り組んだ．収録にはEMA (Electromagnetic Articulography)と呼ばれる特殊な機器を用い，現時点で男性アナウンサー1名分の収録を終えたところである．今後は引き続き収録を進めるとともに，音声合成システムの開発に取り組んでいきたい．

研究成果の学術的意義や社会的意義
近年，深層学習等の発展により音声合成のクオリティが格段に向上している．しかし一般的な音声合成では人間の発音に関する詳細な特徴を用いていないため，人間ならではの発音の失敗や声質の変化に対応することが難しい．本研究で取り組む調音運動ベースの音声合成は人間の発声の仕組みに近い方式をとるため，こうした人間ならではの声の変化に対応できる可能性がある．こうした合成のモデルを他者の発話の認識等に用いることで，言語情報だけではなく，その背後の発声方式の変化（風邪をひいたとか，口の中が痛いとか）を認識する補助情報として利用することも考えられる．

研究成果の概要（英文）：We have investigated how to synthesize speeches from articulatory features that represent movement of lip and tongue when humans utter. During the first half of the period, we have constructed a speech synthesizer from the features that parameterize the actual movement of lip and tongue. After that, we have collected the data of lip/tongue movement using EMA (Electromagnetic Articulography). We recorded the movement from a male announcer last year, and now we are labeling IPA (International Phonetic Alphabet) on it.

研究分野：音声合成

キーワード：調音運動　音声合成　データベース構築

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

親が発話する音声を乳幼児が模倣するという現象から，乳幼児は音韻情報ではなく調音運動を知覚しているのではないかという見解が古くから示されている．これを行うための知覚システムとして，音声の知覚と生成が 1-system であるとの仮説が提唱されている．1-system であるか 2-system であるかはまだ決着がついていないが，近年の研究では 1-system を支持する結果が示されつつある．

研究代表者らは 1-system の音声認識・合成法を実現するため，人間が行っている方法に近いであろう合成方式の開発を行ってきた．この合成方式では，発声の初期動作である調音運動（口や舌を動かす運動）を HMM でモデル化し，調音運動と声道形状（話者性）の対応関係を多層ニューラルネットワーク(Multilayer Neural network: MLN)でモデル化している．声道形状は線スペクトル対(Line Spectral Pairs: LSP)で表し，声帯振動（音高）を表わす励起信号と畳み込み演算することで合成音を生成している．一般的な音声合成では励起信号としてパルス列と白色雑音が用いられることが多いが，より高品質な合成音を得るため，研究代表者らは線形予測残差信号（以下，残差信号と呼ぶ）を励起信号として用いている．

本合成方式が HMM を用いた一般的な合成法と最も異なる点は，通常は HMM 内にモデル化される話者性（声道形状等）と調音運動を分離することにより，HMM を非常に少ないデータでモデル化することが可能である点である．しかし MLN の構築にはある程度の量のデータが必要である．また音質に関して，現時点では十分に高品質であるとは言い難い．そこで本申請では，本合成方式での高品質な合成音の生成を目指すと共に，少量データでの声道形状のモデル化を実現するために話者変換の技術を検討する．さらに，これらをベースにした合成による認識法についても検討する．

２．研究の目的

本研究では，調音から音響特性までを一つのモデルで扱う一般的な HMM(Hidden Markov Model)音声合成と異なり，調音運動を HMM で，声道形状を多層ニューラルネットワークで個別にモデル化した音声合成法の開発に取り組む．本手法は人間の発声に近い過程を経て音声を合成するという特徴を持っており，古くからの仮説の一つである認識と合成の 1-system 論を構成論的に検証するために本手法を用いることを想定している．この目標の実現には高品質な合成音の生成が重要になる．そこで本申請ではこれまで研究代表者らが検討してきた合成法において，話者の声道形状を表すパラメータと声帯振動を表す信号の平滑化，AutoEncoder を用いた少量データでの話者変換に取り組み，高品質な合成モデルを少量データで構築することを目指す．

３．研究の方法

本研究では高品質な合成音の生成のために(i)および(ii)の課題に，少量データでの合成のために(iii)の課題に取り組む．

(i)　RNN，LSTM を用いた LSP の平滑化による音質向上

調音特徴（舌の位置や口形を表す特徴量）は HMM の一つの状態（一つの音素の 1/3～1/5 の長さ）内で尤度最大化基準に基づいて選択される．この調音特徴を MLN に入力する際に前後時刻(t=±3)の特徴量を同時に入力しているが，HMM の状態変化に伴う調音特徴の変化を MLN 内で平滑化するには不十分であった．結果として，MLN の出力である声道形状パラメータ（LSP）の急激な変化が目立ち，音質の低下に繋がっていた．そこで時系列情報を内部で扱える RNN (Recurrent Neural Network)，もしくは LSTM (Long Short-Term Memory)を MLN の代わりに用いることによって LSP の平滑化を強化する．平滑化に伴う音質の劣化に対しては，鈍ったパラメータを先鋭化する分散補償の導入等を検討する．

(ii)　残差信号の平滑化，リアルタイム選択アルゴリズムによる音質向上

声帯振動を表わす残差信号について，これまで合成音の生成時には HMM の一つの状態に対して一つの残差信号を対応させていた．このため，HMM の状態変化による残差信号の急激な変化によって音質が低下する場合があった．研究代表者らはこれまで残差の平滑化に関する予備的検討を行ってきたが，本研究では MLN から出力される LSP に最もマッチする残差信号を選択するアルゴリズムの導入により，音質の向上を図る．また，残差信号の平滑化により音質を向上させる方法も引き続き検討を進める．

(iii)　AutoEncoder を用いた LSP 変換による話者変換

本合成方式では，HMM は少量のデータで構築できるものの，MLN の構築にはある程度のデータを要する．そこで研究代表者らは MLN の再学習による話者適応によって話者性を変換する方法を検討してきた．しかし，この方式では音質の点で改善の余地があった．本申請では，AutoEncoder を用いて LSP のパラメータを圧縮し，このパラメータを直接変換する話者依存のニューラルネットワークを学習することにより，少量データでの声質変換を実現する．声道形状を表す LSP パラメータを直接変換するため，話者性の変換を効率的に実現できると考えている．

４．研究成果

研究開始の当初は研究背景・目的等で示した通り，調音特徴をベースに HMM を利用した音声合成システムを開発することを予定していた．しかし，近年の深層学習の発展により HMM と比較して高性能な深層学習を用いた手法が提案され，人間の声と比較して遜色のない合成音が生成される状況となったため，当初の目標を変更せざるを得なくなった．

当初の提案では中間表現として調音特徴を利用して合成音を生成することを想定していた．調音特徴とは，例えば「破裂音」や「口唇音」のように，人間が発声するときの発音方法について音声学的な音素の分類に基づいて一つの音素を 19 次元のベクトルで表したものである．しかしこれらは声道形状の物理的な特徴を表したものではないため，実際の音声の音響的特徴との乖離が大きい．また，一つの音素には様々な発声方法があり得るため，音素単位の合成では発声方法の差異を表現することが難しいことが分かった．

そこで研究代表者らは 2017 年度から実際の調音運動（口形等の物理的な動き）と，実際に発音された音を表す「単音」に基づく音声合成器を構築するための予備的検討を進めることにした．研究代表者らはまず調音運動の計測を行うため，EMA(Electromagnetic Articulography)という機器を用いることにした．EMA とは，発話者周辺に磁場を発生させ舌上のコイルに発生する誘導電流からコイルの位置を推定する機器で，位置推定の精度が良くサンプリング周波数も高いという特性を持っている．これは一定程度のサンプリング周波数を必要とする音声合成には最適であると考えた．研究代表者は九州大学の鏑木教授，若宮助教の協力の下，2017 年度から EMA による調音運動の収録を開始した．2018 年度末の時点で，収録対象文章（ATR503 文）を決定し，ナレーター1 名分の収録を終えたところである．

EMA による調音運動の収録

調音運動の収録と同時に，発音の正確なラベリングを行うために,研究代表者らは「単音」による発音ラベリングを開始した．「音素(phoneme)」が言語ごとに規定されているのに対し，「単音(phone)」は国際音声記号(IPA: International Phonetic Alphabet)により決められた言語非依存の発声を表すものである．例えば日本語の音素の一つである子音の/h/は，実際には数種類の単音[h]，[ç]，[ɸ]等で発音されており，調音の方法が異なる．こうした単音による正確なラベリングを行うため，2017 年度から音声学の専門家である中央大学の牧野教授を研究分担者に加え，単音ラベリングを進めている．2018 年度末の時点で，EMA データを収録したナレーター1 名分の音声に対する IPA 精密ラベリングを行っている段階である．

５．主な発表論文等

〔雑誌論文〕（計 1 件）

1. Seng Kheang, Kouichi Katsurada, Yurie Iribe and Tsuneo Nitta: "Using Reversed Sequences and Grapheme Generation Rules to Extend the Feasibility of a Phoneme Transition Network-based Grapheme-to-Phoneme Conversion", IEICE Transaction on Information and System, Vol.E99-D, No.4, pp.1182-1192 (2016-4).

〔学会発表〕（計 12 件）

1. 小口 優人, 大村 英史, 桂田 浩一：“Active Appearance Models を用いた読唇”, 第 5 回サイレント音声認識ワークショップ, 講演番号 2 (2018-9).
2. 深井 健大郎, 大村 英史, 桂田 浩一, 平田 里佳, 入部 百合絵, 新田 恒雄：“発話時脳波を利用した音声言語情報の識別”, 第 5 回サイレント音声認識ワークショップ, 講演番号 13 (2018-9).
3. 森口 寛生, 大村 英史, 桂田 浩一：“変分オートエンコーダーを用いた多重音解析の性能評価”, 情報処理学会第 80 回全国大会, 3N-04 (2018-3).
4. 桂田 浩一：“Suffix Array を用いた高速 STD におけるキーワード分割の最適化に関する検討”, 日本音響学会 2017 年春季研究発表会講演論文集, 2-P-16 (2017-3).
5. 浅原 康平, 中根 丈司, 神崎 卓丸, 桂田 浩一, 杉本 俊二, 新田 恒雄, 堀川 順生：“日本語音節発話・想起時の脳波解析”, 日本音響学会 2017 年春季研究発表会講演論文集, 1-7-1 (2017-3).
6. 神崎 卓丸, 浅原 康平, 中根 丈司, 桂田 浩一, 杉本 俊二, 堀川 順生, 新田 恒雄：“発話時と想起時の脳波による日本語短音節認識の比較”, 日本音響学会 2017 年春季研究発表会講演論文集, 3-Q-47 (2017-3).
7. 渡辺 拓也, 桂田 浩一, 金澤 靖：“顔画像の対称 3D-AAM による顔方向非依存な発話認識”, 電子情報通信学会技術研究報告, PRMU2016-127, pp.135-140 (2017-1).
8. Kohei Asahara, Jozi Nakane, Takumaru Kanzaki, Shunji Sugimoto, Kouichi Katsurada, Tsuneo Nitta, and Junsei Horikawa: "EEG during Japanese syllable recall and speech tasks", Proc. of The 3rd Annual Meeting of the Society for Bioacoustics, p.24

(2016-12).
9. Takumaru Kanzaki, Shunji Sugimoto, Kouichi Katsurada, Junsei Horikawa, and Tsuneo Nitta: "Japanese monosyllable recognition from EEG", Proc. of The 3rd Annual Meeting of the Society for Bioacoustics, p.26 (2016-12).
10. Takuya Watanabe, Kouichi Katsurada, and Yasushi Kanazawa: "Lip Reading from Multi View Facial Images Using 3D-AAM", Proc. of ACCV2016 Workshops, W9-04 (2016-11).
11. 浅原 康平, 中根 丈司, 神崎 卓丸, 中澤 香太, 桂田 浩一, 杉本 俊二, 新田 恒雄, 堀川 順生：“日本語単音節発話時と想起時の脳波解析”, 日本音響学会 2016 年秋季研究発表会講演論文集, 2-P-27 (2016-9).
12. 神崎 卓丸, 浅原 康平, 中根 丈司, 中澤 香太, 桂田 浩一, 杉本 俊二, 堀川 順生, 新田 恒雄：“脳波からの日本語単音節認識方式の検討”, 日本音響学会 2016 年秋季研究発表会講演論文集, 2-P-29 (2016-9).

６．研究組織

(1)研究分担者
研究分担者氏名：新田恒雄
ローマ字氏名：Tsuneo Nitta
所属研究機関名：早稲田大学
部局名：グリーンコンピューティング・システム研究機構
職名：招聘研究員
研究者番号（8 桁）：70314101

研究分担者氏名：牧野武彦
ローマ字氏名：Takehiko Makino
所属研究機関名：中央大学
部局名：経済学部
職名：教授
研究者番号（8 桁）：00269482

研究分担者氏名：金澤靖
ローマ字氏名：Yasushi Kanazawa
所属研究機関名：豊橋技術科学大学
部局名：大学院工学研究科
職名：准教授
研究者番号（8 桁）：50214432

(2)研究協力者
研究協力者氏名：鏑木時彦
ローマ字氏名：Tokihiko Kaburagi

研究協力者氏名：若宮幸平
ローマ字氏名：Kohei Wakamiya