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

平成　２９　年　　５　月　３１　日現在

機関番号：　１２６０１
研究種目：　挑戦的萌芽研究
研究期間：　2015～2016
課題番号：　１５Ｋ１２０５９

研究課題名（和文）EMA計測と統計的メディア変換技術の統合による音声・調音変換技術の確立

研究課題名（英文）Acoustic-to-articulatory conversion based on integration of EMA-based measurement and statistical media conversion techniques

研究代表者

峯松　信明（Minematsu, Nobuaki）

東京大学・大学院工学系研究科（工学部）・教授

研究者番号：９０２７３３３３

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では，調音測定技術，音響測定技術，及び，統計的メディア変換技術を統合することで，音声信号のみから調音器官運動の様子を高精度に推定する手法を検討した。また，研究の遂行に必要なコーパス構築も行った。その結果より詳細には，１）EMA を用いて一人の日中バイリンガル話者を対象に，日中各言語の音声・調音パラレルコーパスを構築した。２）音声からの調音運動推定問題に対して，話者正規化技術を応用して精度向上を実現した。３）音声の構造的表象を用い，当該話者が正確に発音できない音素に対しても，想定される調音運動を推定する技術を構築した。成果はトップカンファレンスにて発表し，雑誌論文にまとめることができた。

研究成果の概要（英文）：In this study, by integrating three techniques, articulatory measurement, acoustic measurement, and statistical media conversion, novel techniques were built, which can convert speech signals into the articulatory movement pattern that is assumed to be conducted to generate the speech signals. Further, a corpus was developed, which is required to realize the above technique. As a result of this study, 1) by using EMA, a parallel corpus between acoustic measurements and articulatory measurements was built from a single Japanese-Chinese bilingual speaker. 2) the performance of acoustic-to-articulatory mapping was improved by introducing speaker normalization techniques. 3) by using the structural representation of speech, a novel technique was built to predict the articulatory movement of a phoneme, which is impossible for the speaker to generate correctly. These results were presented at top conferences of speech communication and a journal paper was published.

研究分野：　音声コミュニケーション

キーワード：　音声・調音推定　EMA　パラレルコーパス　話者正規化　音声の構造的表象　外国語学習

１．研究開始当初の背景

音声の技術は基本的に音響の技術である。空気振動として（波形として）観測される音声信号がスペクトル解析され，音声認識，音声合成など様々なアプリケーション，サービスが実現されている。一方，音声は，調音器官（音声器官）の運動によって生成されるものであり，音声信号から調音運動の推定が可能であれば，例えば，発話障害の訓練，外国語学習の訓練，更には，よりリアルなヒューマノイドの実現（テキスト音声合成モジュールを使って喋るのではなく，ロボットの舌が適切に動いて喋る）へと繋がる。このような調音運動の測定・予測を基盤としたアプリケーション，サービスを考える場合，調音運動の直接的な測定や，また，音声信号からの調音運動の予測は，重要な基盤技術である。

従来の研究では，特定話者の声から，その人の調音運動を推定するタスクを検討することが多かった。この場合，別の人の声を入力すると，声の話者間差異のために，推定結果は求めるべき調音運動を大きくずれることとなる。本研究は，声の話者間差異に頑健な推定方法を検討することから開始され，その後，発声訓練をタスクとした場合に必要となる技術について検討した。

２．研究の目的

上記したように，本研究の第一の目的は，複数話者間の声の音響差異に頑健に動作する，音響・調音マッピング技術の構築である。また，発声訓練などの応用アプリケーションを考えた場合，当該話者が十分に正しく生成できない音素について，その話者が発声できるようになった場合の調音運動が推定できれば，応用範囲が格段と広がる。これらの考慮し第二の目的として，当該話者が発声に困難を抱える音素について，その話者が正しく発声できた場合の調音運動を推定する技術の構築を目指した。更には，第三の目的として上記研究成果をより拡充するために必要となる，単独話者の二言語（バイリンガル）音声・調音コーパスの構築を，世界に先駆けて検討することとした。

３．研究の方法

(1) 話者正規化技術の導入による，音声・調音マッピングの高精度化

音声・調音マッピングは技術的には，統計的メディア変換（マッピング）技術を応用することで構築できる。そこで話者の正規化を実現する場合，最も簡素な実装は，任意の話者の音声を，一旦，音声・調音マッピングが既に実装されている特定話者Ａの声に変換し（話者変換），変換の結果得られた話者 A の声を，音声・調音マッピングモジュールへ入力することである（**縦続モデル**）。

この場合，二つの変換を経ることから，予測誤差が累積される可能性が高い。そこで，二つの変換モデルを一つに統合する，**統合モデル**も検討した。入力話者→話者 A，話者 A の声→話者Ａの調音運動，という二つのモデルがある訳だが，前者の出力，後者の入力がともに同一の確率分布に従うと考えた場合，三種の観測量（入力話者の声，話者 A の声，話者Ａの調音運動)全体が従う確率分布をGMMでモデル化し，最終的に，入力話者の声から，話者Ａの調音運動を一つのモデルで推定する。この場合，三種の観測量が常に揃った環境とはならないため，欠損データを予測しつつ，変換を行うことになる。

(2) 発声困難な音素に対する調音運動の推定技術の構築

一般に音声・調音変換は，当該話者が発声した音声を入力することから，その話者の発話能力を前提とした応用しかできない。その話者にとって発声が難しい音素は，当然，調音運動推定はできない。しかし，外国語学習応用など，話者が発声できない音素に対して，当該話者が（訓練の結果）できるようになると想定される調音運動を提示することが求められる。ここでは，当該音素を発声できる話者の音声に対して，これを，音声の構造的表象を用いて表現し（音声の構造的表象は，性別，年齢など，その話者固有の属性に強く依存する声の成分を捨象した表象として提案されている），その表象を参照しつつ，当該話者の当該音素に対する調音運動を推定することを試みた。発音ができる話者の構造的表象から得られる（音響空間での）制約を調音空間に変換し，当該話者の調音空間において，この制約を満たしつつ探索を行い，所望の調音運動を得る方法を検討した。

(3) 日中バイリンガル話者を用いた音声・調音パラレルコーパスの構築

上記二種の研究は，イギリス英語のコーパスを使って行われたが，この結果を日本語や中国語の外国語応用する場合は，日中の音声・調音パラレルコーパスが必要となる。更には，同一話者（つまり，バイリンガル話者）の日本語音声・調音コーパス，中国語音声・調音コーパスが準備されていると開発の効率化に繋がる。音声工学の分野では，音声・調音パラレルコーパスは数種類のものが公開されているが，同一（バイリンガル）話者による二言語の音声・調音パラレルコーパスは存在していない。本研究では，調音音声学に関する研究基盤を充実させることも狙い，日中バイリンガル話者の，両言語に対する音声・調音パラレルコーパスを構築した。

４．研究成果

(1) 話者正規化技術の導入による，音声・調音マッピングの高精度化

実験の結果，正規化を行わない場合よりも正規化を行った方が精度が高く，また，二つのマッピングモデルの比較では，統合モデルが有意に高い変換精度を示すことが得られ，

統合モデルの優位性が証明された。

図１は調音観測位置毎に計測した推測エラーの大きさであり，図２は音素毎に計測した推定エラーの大きさである。いずれもbaseline より，縦続（concatenation）／統合（unification）モデルの方が精度が高く，また縦続モデルよりも統合モデルの方が精度が高いことが分かる。

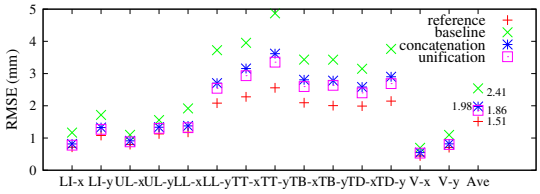

図１：観測位置別に計測した推定誤り

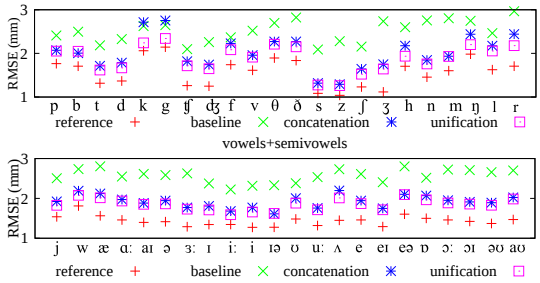

図２：各音素別に計測した推定誤り

(2) 発声困難な音素に対する調音運動の推定技術の構築

他話者から得られた構造的表象を制約として用いて，当該話者の調音位置を探索する訳だが，調音音声学的な事前知識を有効的に導入することで，母音を対象にした実験の結果，良好な精度で推定できることが示された。図３は上図が男性話者，下図が女声話者を対象として行った実験である。いずれも，適切な制約を使って探索を行うことで，予測精度が向上していることが分かる。

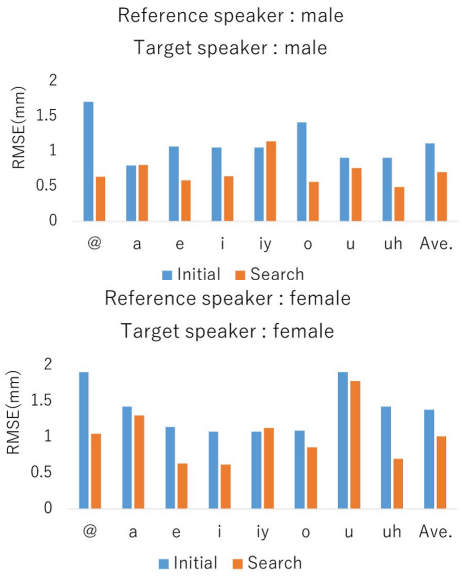

図３：母音別に計測した推定誤りの大きさ

(3) 日中バイリンガル話者を用いた音声・調音パラレルコーパスの構築

コーパスを構築するのみに留まらず，同一話者の日本語調音・中国語調音間に存在する，言語間差異について，前田モデルを前提とした統計分析を行った。その結果，中国語にしか存在しない調音運動が存在することが，純粋にデータドリブンな解析から示唆される興味深い結果を得た。これらの知見は，外国語学習に応用することが期待される。

上記の成果を，音響学会全国大会，電子情報通信学会音声研究会などの国内会議，また，INTERSPEECH や ISSP などの国際会議にて発表し，最終的に音響学会誌に雑誌論文を発表するに至った。

５．主な発表論文等

〔雑誌論文〕（計　１件）

① 内田秀継，齋藤大輔，峯松信明，音声の構造的表象を用いた未観測音素の調音運動の推定，音響学会誌，査読有，2017（採録済・掲載予定）

〔学会発表〕（計　６件）

① H. Uchida, T. Hashimoto, D. Saito, N. Minematsu and A. Suemitsu, “A recording of bilingual acoustic-articulatory data from a Japanese-Chinese bilingual speaker with a 3D-EMA system”, ISSP, 2017/10, 中国天津（中国），採択済

② H. Uchida, D. Saito, and N. Minematsu, “Acoustic-to-articulatory mapping based on mixture of probabilistic canonical correlation analysis,” INTERSPEECH, 2017/8, ストックホルム（スウェーデン），採択済

③ H. Uchida, D. Saito, N. Minematsu, “Prediction of the articulatory movements of unseen phonemes of a speaker using the speech structure of another speaker,” INTERSPEECH, 2016/9/9, サンフランシスコ（米国）

④ 内田秀継，斉藤大輔，峯松信明，“音声の構造的表象を用いた未観測調音運動の推定に関する検討,” 電子情報通信学会音声研究会，2016/1/14，サンピアンかわさき（神奈川）

⑤ H. Uchida, D. Saito, N. Minematsu, K. Hirose, “Statistical acoustic-to-articulatory mapping unified with speaker normalization based on voice conversion,” INTERSPEECH, 2015/9/7, ドレスデン（ドイツ）

⑥ 内田秀継，齋藤大輔，峯松信明，広瀬啓吉，統計的音声-調音マッピングにおける声質変換を利用した話者正規化法の検討，電子情報通信学会音声研究会，

2014/11/14, 九州大学筑紫キャンパス
（福岡）

６．研究組織

(1)研究代表者

峯松 信明 (MINEMATSU, Nobuaki)

東京大学・大学院工学系研究科・教授

研究者番号：９０２７３３３３

(2)研究分担者

齋藤 大輔 (SAITO, Daisuke)

東京大学・大学院工学系研究科・講師

研究者番号： ４０６１５１５０

(3)連携研究者

(4)研究協力者

内田 秀継 (UCHIDA, Hidetsugu)