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　　９　日現在

機関番号：　１４５０１
研究種目：　基盤研究(B)（一般）
研究期間：　2013～2016
課題番号：　２５２８２０５３

研究課題名（和文）脳性麻痺障がい者の意図認識によるユニバーサルコミュニケーション支援機器の開発

研究課題名（英文）Development of communication support system based on intent recognition for cerebral palsy

研究代表者

滝口　哲也（Takiguchi, Tetsuya）

神戸大学・都市安全研究センター・准教授

研究者番号：４０３９７８１５

交付決定額（研究期間全体）：（直接経費）　　13,700,000円

研究成果の概要（和文）：脳性麻痺障がい者の発話スタイルは健常者と異なり，その発話内容を理解するのが困難な場合がある．本研究では，障がい者の自立した社会生活支援に資するコミュニケーション支援機器の開発を目指し，聞き取りの難しい障がい者発話に対して自動音声認識システムの研究開発，聞き取りの難しい障がい者発話を聞き取り容易な音声に変換する音声変換システムの研究開発，更に画像情報を用いたマルチモーダル発話認識システムの研究開発を行い，機械学習法を用いた新しい音声特徴量抽出法などを提案し，従来手法と比較して提案手法の有効性を示した．

研究成果の概要（英文）：An utterance style of a person with an articulation disorder, such as cerebral palsy, is different from that of physically unimpaired persons, and his/her utterance is often unstable or unclear, which makes it difficult for them to communicate. To develop a communication-support system for them, we propose a new automatic speech recognition (ASR) system using a new acoustic feature extraction technique, a voice conversion (VC) method for articulation disorders that converts unclear utterances to clear utterances, and a multi-modal utterance recognition system using a novel feature integration technique based on a machine-learning approach. Experimental results demonstrated that our ASR, VC, and multimodal recognition methods could improve the speech recognition accuracy, the listening speech quality, and the multimodal (speech and image) recognition accuracy in comparison with conventional approaches, respectively.

研究分野：　メディア情報処理

キーワード：　ヒューマン・インターフェース

１．研究開始当初の背景

近年，情報技術の福祉分野への応用が進んでいる．例えば，画像認識技術の応用による手話認識，文章読み上げシステム，ウェアラブル音声合成システムなど，その応用範囲は幅広い．本研究では，構音障がい者に焦点をあて，彼ら/彼女らの不安定な発話音声を，コンピュータを用いて自動認識すること，及び聞き取りやすく変換することを目指す．

構音障がいの一例として，脳性麻痺による場合がある．脳性麻痺とは，筋肉の動きをつかさどる脳の部分が受けた損傷が原因で筋肉の制御ができなくなり，けいれんや麻痺，その他の神経障がいが起こる症状のことである．

アテトーゼ型脳性麻痺者においては，筋肉が不随意に動き正常に制御できない症状が現れる．特に意図的な動作を行う場合や，緊張状態にある時に見られ，この運動障がいの一つとして，正しく構音できない場合がある．身体が不自由であるため，コミュニケーション支援システムとして，ハンズフリーな音声を中心にしたシステムが求められている．

２．研究の目的

構音障がい者の自立した社会生活支援に資するコミュニケーション支援機器の開発を目指し，主に下記３つのサブテーマ毎に研究を遂行していく．

(1)発話認識：音声認識技術を用いることにより，構音障がい者の発話内容を聞き取ることが困難な場合でもコミュニケーションが円滑になることが期待できる．しかし，構音障がい者において，筋肉の緊張から生じる発話スタイルの変動成分は，健常者と比較して非常に大きくなり，発話認識精度を劣化させる要因となっている．本研究では，構音障がい者の発話認識精度を向上させるため，新たな音響特徴量抽出手法を提案する．

(2)音声変換：構音障がい者の不安定な発話音声を，聞き取りやすい音声に変換することを目指す．これまでの声質変換技術はスペクトル変換に着目したものが多く，発話長，Durationを変換したものは少なかった．構音障がい者の発話の場合，健常者と比較してDurationが長くなる上，アテトーゼ現象により発話リズムが崩れる例が指摘されている．特に文章単位の長い発話においては，Duration が聞き取りやすさに与える影響が大きいと考えられる．声質変換において，ピッチ特徴量に対しては線形変換を用いる事が多いが，Durationは前後の音素関係により多様に変化するため，線形変換を用いることは困難である．音声合成においては，Duration は隠れマルコフモデルによりモデル化されることが多いが，入力音声の音素ラベルが与えられていない声質変換での使用は困難である．そこで本研究では，障がい者発話のDurationを健常者発話のDurationへと変換することで，聞き取りやすさの向上を目指す．

(3)マルチモーダル発話認識：人は発話内容を理解する際，種々の情報を統合的に利用している．音声が聞き取り難い場合，発話者の顔，特に唇の動きに注目して発話内容を理解しようとする．例えば，重度難聴者は耳で音を聞くことができないため，正確な発音をすることが難しく，発話スタイルが健常者と異なる．

彼ら/彼女らのコミュニケーション手段の一例として（手話以外に）口話を行う方々もおられ，訓練により意図した発話の唇の形状を作られている．しかし発話者が耳で音を聞くことが難しいために，聞き取りが難しい発話になる場合がある．そこで，彼ら/彼女らの音声を認識するために，唇画像を利用した音声認識システムの構築を行う．

３．研究の方法

３つのサブテーマ毎に研究方法について述べる．

(1)発話認識：研究代表者らは，これまでに畳み込みニューラルネットワークを用いた発話変動に頑健な音声特徴量抽出法を提案してきた．この手法は，ネットワークの学習に誤差逆伝播法を用いており，教師信号として隠れマルコフモデルによる強制アライメントの結果を用いている．しかし構音障がい者のスペクトルは変動が大きいため，精度の良いアライメントをとることが難しい．そのため，ネットワークの学習に用いる教師信号は誤りを含むことになり，より有効な特徴量抽出を阻害していると考えられる．さらなる構音障がい者の音声認識精度向上のために，より精度の高いアライメント情報を得る手法を研究する．

(2)音声変換：障がい者発話は音素の繋がりにより，Durationが様々に変化するため，単純な線形変換では対応が困難であると考えられる．そこで，本研究では従来の統計的声質変換モデルをベースに，パラレルデータ間のアライメント情報に基づいたフレームベースの Duration 特徴量を研究する．障がい者と健常者のパラレル発話間のマッチング距離を Duration 特徴量とすることで，入力話者に対する出力話者の Duration を，入力話者のフレームベースで記述することが出来る．入力話者のスペクトル特徴量と，求められた Duration 特徴量を統合しモデル学習することで，音素ラベルが与えられていない声質変換タスクにおいても Duration 変換が可能となる．図 1 に変換処理の概要を示す．また，Durationは前後の音素関係により変化すると考えられるため，入力特徴量として複数フレームを考慮した長距離特徴量を用いる．

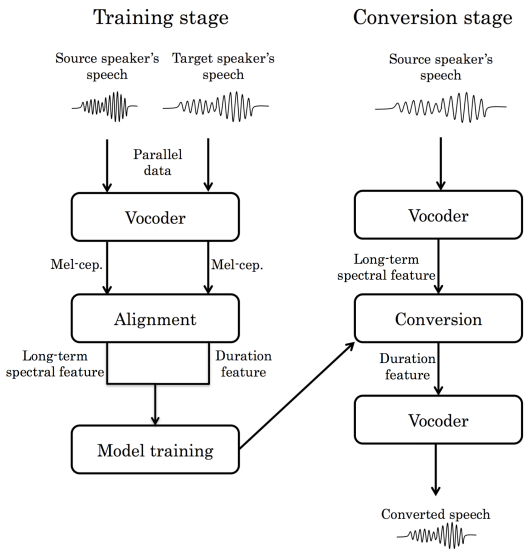

図 1. Duration 変換の概要

(3)マルチモーダル発話認識：音声と唇画像を統合した特徴量抽出のために，factored 3-way restricted Boltzmann machine (F3WRBM)を用いる．このモデルは restricted Boltzmann machine (RBM)を拡張したものであり，エネルギー関数に基づく確率モデルである．

F3RBM は，二つの観測変数と一つの潜在変数からなるモデルであり，音響特徴量と画像特徴量を観測変数とすることで，音声と唇画像を相補的に考慮できる特徴が潜在変数として現れると期待できる．

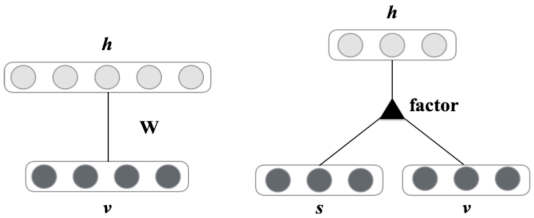

図 2. RBM (左側) と factored 3-way RBM

図 2 に示されるように，restricted Boltzmann Machine (RBM)は，可視素子 v と隠れ素子 h からなる無向グラフィカルモデルである．RBM のパラメータは，学習データを用いて対数尤度最大化に基づいて推定される．本研究では RBM を拡張し，音響特徴量を表す s，画像特徴量を表す v，潜在特徴量を表す h の 3 変数間の関係性を 3-way RBM を用いて表現する．ここで，3 つの確率変数間に，ファクターと呼ばれる概念を設け，音響特徴量-ファクター，画像特徴量-ファクター，ファクター-潜在特徴量の間の結合行列を求める事により，パラメータ数を削減し，効率的にパラメータ推定を行う．

4．研究成果

上記において述べた 3 つのサブテーマ毎に研究成果を示す．

(1)発話認識：図 3 に，健常者と構音障がい者が，/i k i o i/と発話した際のスペクトルを示す．

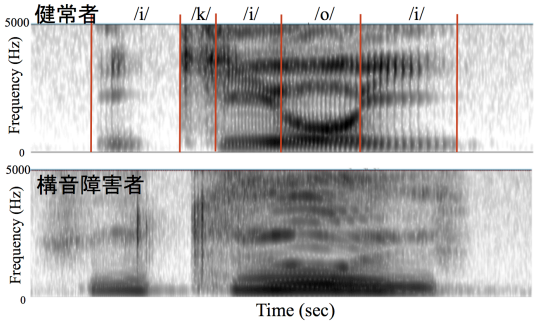

図 3. 健常者と構音障がい者の音声スペクトル

色が濃い箇所は，強い周波数成分を持つことを表している．健常者発話では，特に後半の/i o i/において，白黒の違い（周波数成分の減り張り）がはっきり目視でも確認できる．一方，構音障がい者発話においては，健常者発話のような明瞭な周波数成分の変化は見られず，音素境界が曖昧になっている．

本研究では，音素境界の曖昧性を考慮し，正規分布を用いた確率表現による音素のソフトラベリング法を提案する．ある発話において各音素区間の中心を平均とする正規分布を構成し，これを用いた生存確率により各時間における音素ラベルを与えることにより，音素境界のラベリングを，曖昧性を考慮して行うことが可能になる．

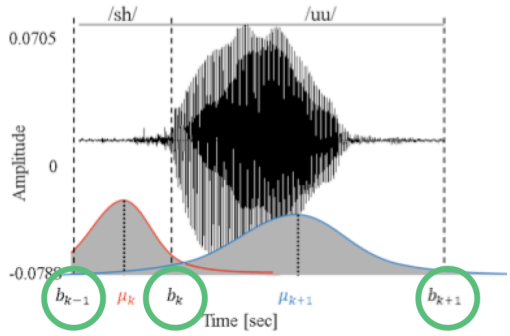

図 4. 正規分布を用いたソフトラベリング

図 5 に音声認識実験の結果を示す．評価には 216 単語認識タスクを用いた．“forced”は，従来の強制アライメントによるハードラベルを用いて畳み込みニューラルネットワークを学習した結果を表し，“manual”は手動でアライメントを求めた結果を表し，“gaussian”は正規分布を用いたソフトラベルによる結果（提案手法）を表している．

音声認識実験結果より，提案手法により約 92%の音声認識精度が得られ，従来手法よりも音声認識精度が改善された．

また，構音障がい者の発話の場合，ある音素が発話されていない事もあり，ソフトラベリングだけでは十分に対応しきれない事もある．間違えている音素ラベル情報で畳み込みニューラルネットワークを学習すると，誤

ったモデルが学習されてしまい認識精度が劣化する恐れがある．本研究では，一部の学習データを使わない dropout を適用し，曖昧な学習を行う事で誤りを含む音素ラベリングに頑健な学習を行った．図 5 に示す音声認識実験結果より，dropout を適用する事により音声認識精度が改善されているのが分かる．

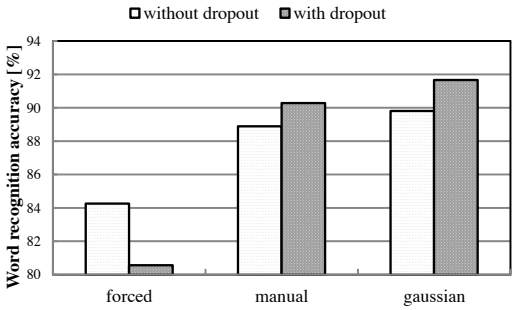

図 5. 構音障がい者の 216 単語音声認識率

今回の評価実験では，話者特定の音響モデルを用いており，かつ比較的認識精度の高い発話者の結果を示している．今後更に多くの発話者に対して評価実験を進め，不特定話者モデルを用いた発話認識の実現を目指していく．

(2)音声変換：本実験では，特徴量変換として，Gaussian Mixture Model (GMM)に基づく最尤特徴量変換を用いる．入力話者に対応する出力話者の Duration 特徴量を用いて GMM を学習することで，特徴量空間は教師無しで音響クラスに分離され，その音響クラスごとに Duration 変換が行われる．50 文を学習データ，他の 50 文を評価データとした．本実験では，Duration に着目するため，スペクトル包絡，基本周波数，非周期成分は入力話者のものを用いた．客観評価指標として，入力話者あるいは変換音声と出力話者のケプストラムによる DP (Dynamic Programming) 距離を用いた．

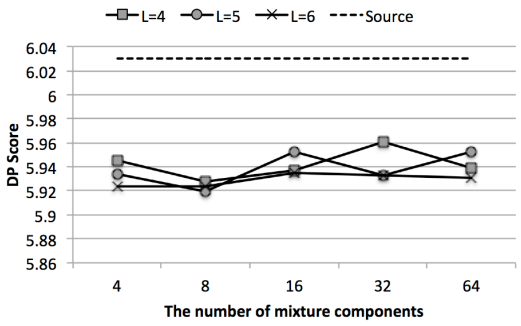

図 6. Duration 変換後の DP 距離

図 6 に長距離特徴量の窓幅による DP 距離の変換を示す．Score は変換前の障がい者音声と目標となる健常者音声間の DP 距離を示し，L は窓幅を表す．図より，Duration 変換により，無変換の場合（点線：Source との距離）と比較して DP 距離が削減できており，提案手法の有効性が確認できる．L=5 の場合が最も適切に変換できることがわかる．

今後は，提案した Duration 変換を行なった上で，スペクトル変換やピッチ変換を行い，障がい者音声を，話者性を維持しながら聞き取り易い音声へと変換することを目指す．更に話者数を増やして提案手法の有効性を確認していく予定である．

(3)マルチモーダル発話認識：図 7 にマルチモーダル音声認識結果を示す．学習データには 2,620 単語，評価データには 216 単語を用いた．“MFCC+Δ+DCT”は，音響特徴量に従来のメルケプストラム（MFCC）特徴量とその一次差分（Δ）及び画像特徴量に離散コサイン変換（DCT）を用いた際の認識結果を示している．“RBM”は，音声のメル周波数スペクトル 39 次元と DCT30 次元を連結して RBM を学習し，潜在特徴量を新たな特徴量として用いた際の認識結果を示している．F3WRBM だけでは十分な音声認識精度が得られていないが，従来の音響特徴量メルケプストラム（MFCC）を連結して認識を行うと，一番良い認識制度が得られているのが分かる．

今回は，パラメータ学習時に制約を入れていないため，3 つある結合行列のうち，ファクターと潜在変数間の結合行列に学習が偏る傾向が見られた．今後は，それぞれの結合行列がバランス良く学習できる正規化を導入し，精度の改善を試みていく．

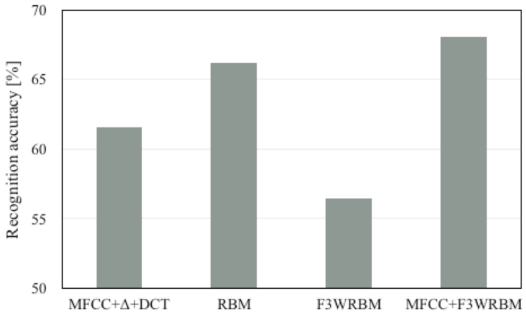

図 7. 音声と唇特徴を用いたマルチモーダル単語音声認識率

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 10 件）

① Yuki Takashima, Toru Nakashika, Tetsuya Takiguchi, Yasuo Ariki, Phone Labeling Based on the Probabilistic Representation for Dysarthric Speech Recognition, American Journal of Signal Processing, 査読有, Vol. 6, No. 1, 2016, pp. 19-23, DOI:10.5923/j.ajsp.20160601.03

② Ryo Aihara, Tetsuya Takiguchi, Yasuo

Ariki, Individuality-preserving Voice Conversion for Articulation Disorders Using Phoneme-categorized Exemplars, ACM Transactions on Accessible Computing, 査読有, Vol. 6, Issue 4, Article No. 13, 2015, pp. 1-17

③ Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki, Exemplar-Based Voice Conversion in Noisy Environments, IEICE, 査読有, Vol. E96-A, No. 10, 2013, pp. 1946-1953

〔学会発表〕(計 38 件)

① 上田 怜奈, 滝口 哲也, 有木 康雄, 話速補正に基づく話者性を維持した構音障害者のための音声合成システム, 日本音響学会 2016 年秋季研究発表会講演論文集, 2016 年 9 月 15 日, 富山大学(富山県)

② 高島 悠樹, 中鹿 亘, 滝口 哲也, 有木 康雄, Factored 3-Way Restricted Boltzmann Machine を用いたマルチモーダル音声認識の検討, 日本音響学会 2016 年秋季研究発表会講演論文集, 2016 年 9 月 15 日, 富山大学(富山県)

③ Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki, Dysarthric Speech Modification Using Parallel Utterance Based on Non-negative Temporal Decomposition, 7th Workshop on Speech and Language Processing for Assistive Technologies, 2016 年 9 月 13 日, サンフランシスコ (アメリカ)

④ Yuki Takashima, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki, Nobuyuki Mitani, Kiyohiro Omori, and Kaoru Nakazono, Audio-Visual Speech Recognition Using Bimodal-Trained Bottleneck Features for a Person with Severe Hearing Loss, Interspeech, 2016 年 9 月 10 日, サンフランシスコ (アメリカ)

⑤ 羅 兆杰, 滝口 哲也, 有木 康雄, Emotional Speech Conversion Using Deep Neural Networks, 日本音響学会 2016 年春季研究発表会講演論文集, 2016 年 3 月 10 日, 桐蔭横浜大学 (神奈川県)

⑥ Yuki Takashima, Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, Feature Extraction Using Pre-Trained Convolutive Bottleneck Nets for Dysarthric Speech Recognition, EUSIPCO, 2015 年 9 月 1 日, ニース (フランス)

〔その他〕
ホームページ
http://www.me.cs.scitec.kobe-u.ac.jp/~takigu/index.html

６．研究組織

(1)研究代表者
滝口　哲也 (TAKIGUCHI, Tetsuya)
神戸大学・都市安全研究センター・准教授
研究者番号：４０３９７８１５

(2)研究分担者
有木　康雄 (ARIKI, Yasuo)
神戸大学・都市安全研究センター・名誉教授
研究者番号：１０１３５５１９

(3)研究分担者
高田　哲 (TAKADA, Satoshi)
神戸大学・大学院保健学研究科・教授
研究者番号：１０２１６６５８

(4)研究分担者
中川　誠司 (NAKAGAWA, Seiji)
千葉大学・フロンティア医工学センター・教授
研究者番号：７０３５７６１４

(5)研究分担者
中井　靖 (NAKAI, Yasushi)
宮崎大学・教育学部・准教授
研究者番号：８０４６２０５０

(6)研究分担者
榎並　直子 (ENAMI, Naoko)
神戸大学・大学院システム情報学研究科・助教
研究者番号：８０６２８９２５