Assistive Technology Using Machine Learning Based on Multi-Domain Data for Articulation Disorders

Takashima, Yuki

(Degree)
博士（工学）

(Date of Degree)
2020-03-25

(Date of Publication)
2021-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第7781号

(URL)
https://hdl.handle.net/20.500.14094/D1007781

（別紙様式３）

論文内容の要旨

氏　名　高島　悠樹

専　攻　情報科学

論文題目（外国語の場合は，その和訳を併記すること。）

Assistive Technology Using Machine Learning Based on Multi-Domain Data for Articulation Disorders

（構音障害者のための複数ドメインのデータに基づく機械学習を用いた支援技術）

指導教員　滝口　哲也　教授

本研究では，構音障害者の音声を対象としたコミュニケーション支援技術として，主に少量データの問題点に着目した手法を提案する．本提案の貢献として，以下の３点が挙げられる．

1. どんな発話内容の音声も声質変換の学習データとして用いることができる（第四章）
2. 追加のデータセットから所望の特徴を効率的に学習できる（第五章）
3. 雑音環境下の音声認識性能を向上させられる（第六章）

構音障害とは，知的な障害はないが，音を作る器官やその動きに問題があり正しく構音できない障害のことである．本稿では，アテトーゼ型脳性麻痺による構音障害者と重度難聴による構音障害者を対象としている．アテトーゼ型脳性麻痺による構音障害者は，意図した動作時に筋肉の不随意運動が生じるため，身体を自由に動かすことができない．発話時においても，調音器官を適切に動かすことができないため，安定した発話が困難になる．そのため，彼らの発話スタイルは健常者とは大きく異なり，聞き取りにくい音声となる．また，発話による身体への負担が大きく，大量の音声を収録することは難しい．重度難聴による構音障害は，聴覚フィードバックに異常をきたすために生じる構音障害である．脳性麻痺による構音障害者と同様に，彼らの発話スタイルも健常者とは大きく異なる．また，彼らは周囲の雑音を聞くことができないため，雑音環境下で自分の声量を適切に調整しにくい．本研究では，これらのタイプの構音障害者音声の発話内容を理解するための提案を行う．

上述の目的を達成するために，２つのアプローチが考えられる．彼らの音声を聞き取りやすく変換するアプローチとテキストとして認識するアプローチである．前者は声質変換と呼ばれる技術により実現可能である．声質変換とは，音声に含まれる音韻情報を維持しながら，話者性や感情性など特定の非言語情報のみを変換する技術である．また，健常者音声を障害者音声らしく変換することで，障害者モデルのためのデータ拡張としての応用も考えられる．後者のアプローチは音声認識であり，不規則に変動する障害者音声から正しく音韻情報を推定する必要がある．従来研究の多くはモデルの学習に大量の学習データを必要するため，障害者へ適用することは難しいという問題点があった．本研究では，障害者音声の限られたデータ量の音声から効率的に障害者らしい特徴を学習する手法を提案する．

第一章では，序論として，障害者支援技術に関する研究背景及び本研究の位置付け，本論文の構成について述べる．

第二章では，音声の特徴量やその抽出法など，本研究で用いる手法を理解する上で必要

となる音声信号処理の技術について述べる．

第三章では，従来の声質変換手法と音声認識手法について述べ，アルゴリズムについて概説する．

第四章では，パラレルデータフリー声質変換を提案している．パラレルデータとは，入力話者と出力話者が同一内容を発話した音声に対して，動的計画法により時間アライメントを取り発話長を揃えたデータである．従来の NMF (Non-negative matrix factorization) に基づく声質変換は，音韻項を共有しながらモデルの学習を行うため，パラレルデータを必要とする．パラレルデータを使用する問題点として，データに対して不自然な伸縮をすること，また，発話内容を制限するため使用できるデータセットが限られるという点がある．構音障害者は意図した内容を発話をすることが難しく，パラレルデータはシステムを利用する上で大きな制約となる．提案手法では，入力スペクトルの周波数ビン数・フレーム長に依存しない項を持つ NTD (Non-negative Tucker decomposition)を用いて入力スペクトルを分解する．この項を共有化することで、音韻項を話者ごとに表現することができ，パラレルデータを必要としないモデルの学習方法を提案する．客観評価指標に基づく評価実験では，代表的な声質変換手法である GMM (Gaussian mixture model)や NMF と比較を行い，提案手法の有効性を示した．また，話者性に対する主観評価実験により，提案手法が性能を維持したまま，従来の声質変換法を非パラレル拡張できることを示した．

第五章では，日本人の脳性麻痺による構音障害者を対象とした，複数データベースを用いた音声認識を提案する．このタイプの構音障害者は，発話による身体への負担が大きく，モデルの学習のために十分なデータ量を確保することが難しい．障害者音声に含まれる要素として，言語的な特徴と障害者らしい特徴が考えられる．それぞれ，言語依存特徴，言語非依存特徴と仮定し，前者は日本人健常者音声から，後者は構音障害を持つ英語話者音声から転移させる手法を提案した．本研究では，近年盛んに研究が行われている深層学習に基づく音声認識モデル Listen, attend and spell (LAS)モデルを採用する．LAS モデルは特徴抽出器と音素推定器から構成される．複数言語を１つモデルで表現するために，言語ごとの音素推定器を持つモデルを提案した．音素認識実験により，ランダム初期化や従来の転移学習法と比較して，提案手法が構音障害を持つ英語発話から効果的に障害者性を学習できることを示した．

第六章では，重度難聴者を対象とした雑音環境下マルチモーダル音声認識を提案する．実用シーンでは背景雑音は音声認識性能を劣化させる要因の１つとなる．健常者であれば雑音の大きさや種類に応じて声量を調整できるが，重度難聴者は周囲の音を聞くことが難

しいため，自分の声量を適切に調整することができない．そのため，このような環境では音声認識性能は著しく劣化する．しかし，聴覚障害者の中には唇の動きを読むことで会話を行う，口話ができる方もおり，雑音環境下でも唇の形を適切に作ることができる．そこで，本研究では，唇画像を用いて音声認識精度を補う手法を提案する．音声・唇画像からの特徴抽出に畳み込みニューラルネットワークを使用することで，音声の時間的な微小変動や画像の切り出しによるズレに頑健な特徴抽出を行う．また，ネットワークの中間層で次元数を小さく絞ることで情報を集約させる効果が得られるため，この層をボトルネック特徴量として認識に使用する．単語認識実験では，より雑音が大きい環境で提案手法が有効であることが示した．

最後に第七章にて，全体を通してのまとめと，今後の課題について述べる．本論文で提案した手法は，障害者分野だけでなく，十分な学習データ量を用意することが難しい領域においても応用可能であり，今後さらなる発展が望まれる．

（別紙１）　　論文審査の結果の要旨

氏名	高島　悠樹		
論文題目	Assistive Technology Using Machine Learning Based on Multi-Domain Data for Articulation Disorders （構音障害者のための複数ドメインのデータに基づく機械学習を用いた支援技術）		
審査委員	区　分	職　名	氏　　名
	主　査	教授	滝口　哲也
	副　査	教授	玉置　久
	副　査	教授	佐野　英樹
	副　査	准教授	高島　遼一
	副　査		印

要　　旨

本研究では、構音障害者の音声を対象としたコミュニケーション支援技術として、少量学習データにおける問題を解決する手法を提案している。構音障害とは、調音器官やその動きに問題があり、発音が正しく出来ない状態のことを言う。音声はヒトにとって、重要なコミュニケーション手段の１つであるが、構音障害者は周囲の人たちに意図した内容を伝えることが難しい。ここでは、アテトーゼ型脳性麻痺による構音障害者と重度難聴による構音障害者を対象としている。

脳性麻痺は、筋肉をつかさどる脳の部分に受けた損傷が原因で筋肉の制御ができなくなり、痙攣や麻痺、そのほかの神経障害が起こる症状である。そのなかでも、脳性麻痺患者の約 20%に発生するアテトーゼ型は、筋肉が不随に動き正常に制御できない症状が現れる。この症状は特に意図的な動作を行う場合や、緊張状態にあるときに見られ、その運動障害の一つとして、正しく構音できない場合がある。アテトーゼ症状には知能障害を合併していないケースや比較的知能障害の程度が軽いケースも多いのが特徴であることから、アテトーゼ型脳性麻痺による構音障害者を対象とした発話支援システムが求められている。また重度難聴者は、難聴により聴覚フィードバックに異常をきたすために構音障害を持つことがある。自分が発声した言葉を自分の耳で聞いて確認することができないため、発話スタイルが健常者とは大きく異なり、周囲の人たちに意図した内容を伝えることが難しい。コミュニケーション支援技術の実現により彼らの社会進出を促進し、生活の質を向上させることが期待される。

構音障害者の音声コミュニケーション支援方法として、音声として伝えるアプローチと、文字として伝えるアプローチの二つが考えられる。これらはそれぞれ、声質変換と音声認識と呼ばれる技術により実現される。声質変換とは、音声中の言語情報を維持しつつ、特定の非言語情報を変換する技術である。音声認識とは、音声中の発話内容を文字として書き起こす技術である。これらの技術は機械学習に基づいたものである。一般に、機械学習モデルは大量の学習データを必要とし、特に近年盛んに研究されている深層学習モデルの性能は学習データ量に大きく依存する。しかし、構音障害者は発話による身体への負担が大きいことが多く、使用できるデータ量が限られている。従って、限られた量のデータから効率的に障害者の特徴を学習する必要がある。本研究では、声質変換に関して１つ、音声認識に関して２つの提案を行う。

第一章では、序論として、障害者支援技術に関する研究背景及び本研究の位置付け、本論文の構成について述べている。また、２つのアプローチと、その関連研究について概説している。

第二章では、音声の特徴量やその抽出法など、本研究で用いる手法を理解する上で必要となる音声信号処理の技術について解説している。

第三章では、従来の声質変換手法と音声認識手法について解説し、その利点と欠点をまとめている。

第四章では、パラレルデータフリー声質変換を提案している。本研究では、障害者の明瞭性向上へ応用されている、非負値行列因子分解（NMF: non-negative matrix factorization）に基づく声質変換に着目している。NMF 声質変換では、入力スペクトルを辞書とその線形結合係数であるアクティビティに分解する。アクティビティが音韻情報を表現すると仮定し、これを話者間で共有しながら辞書学習を行うため、NMF 声質変換は学習データとして、パラレルデータを必要とする。パラレルデータとは、入力話者と出力話者のデータに対してフレームレベルで同期を取ったものであり、同一発話内容の音声を必要とする。そのため、使用できるコーパスが限られるという問題点がある。特に、構音障害者は発話困難な音があり、また学習データ量も限られているため、パラレルデータを用意することが難しい。そのため、障害者音声への応用が難しくなる。本研究では、非負値タッカー分解（non-negative Tucker decomposition）により、話者ごとに可変長の音韻情報項へ分解する。これにより、学習データとして任意の発話内容の音声を使用できる。また、提案手法は、辞書に対して行列分解を施したと解釈することができ、次元圧縮の効果

氏名　高島　悠樹

により、辞書のパラメータ数を小さくすることができる。客観評価・主観評価に基づく声質変換評価実験では、代表的な GMM （Gaussian mixture model）や従来の NMF、パラレルデータを必要としない ARBM（adaptive restricted Boltzmann machine）声質変換と比較を行っており、提案手法の有効性を示している。

第五章からは第四章とは別のアプローチとして、音声認識に関する提案を行っている。構音障害者の発話スタイルは健常者とは大きく異なるため、健常者の音声を用いて学習された音声認識システムを用いて彼らの音声を認識することは極めて難しい。そのため、障害者固有の音声認識システムが求められる。脳性麻痺による構音障害者は筋肉の不随意運動により、発話による身体への負担が大きい。そのため、大量の音声を収録することが難しく、モデル学習に使用できるデータ量は限られている。構音障害者と比べて健常者の音声は大量収集が比較的容易であるため、これを用いたデータ拡張が考えられる。しかし、例えば日本語健常者音声を用いた場合では、一般的な日本語音声の特徴を学習する上では役立つが、日本語障害者音声そのものは増えていないため、障害者音声らしい特徴の学習は行えない。一方で、日本語以外の障害者音声データベースはいくつか公開されており、利用可能である。そこで、本研究では、対象話者と異なる言語の障害者音声を用いて、障害者らしい特徴を学習する手法を提案している。提案手法の枠組みでは、音声の特徴を言語依存な特徴と言語非依存な特徴に分けて捉え、障害者特有の特徴は言語非依存特徴にのみ含まれると仮定している。提案手法はこの仮定に従い、言語依存な特徴に対しては、同一言語の健常者の音声を用いることで学習を補助し、言語非依存な障害者特徴に対しては、他言語の障害者音声を用いることで学習を補助する。具体的には、これらの補助データを用いてモデルのパラメータを事前学習し、その後、ターゲットとなる障害者の少量音声を用いてモデルを再学習（チューニング）する。ただし、従来の音声認識モデルでは異なる言語を同一のモデルで扱うことはできないため、提案のアプローチを使うことは不可能である。そこで、言語固有の音素推定器を持つ新しいモデル構造を提案し、それにより複数言語の障害者音声を利用可能にしている。音素認識による評価実験では、追加のデータベースを用いない場合と従来の健常者音声のみを用いる手法と比較を行っており、提案手法の有効性を示している。

第六章では、聴覚障害者のためのマルチモーダル音声認識を提案している。実環境下において、背景雑音は音声信号に悪影響を及ぼすが、聴覚障害者の場合には特に顕著に現れる。なぜなら、彼らは周囲の音を聞くことができないため、自分の声量を適切に調節することが難しいからである。本研究では、コミュニケーション手段として、唇の動きを読む「口話」を用いる重度難聴者を対象としている。彼らの中には、発話の際に自然な唇の形を作ることができる方もおり、本研究ではこの唇の動きを音声認識性能を補うために利用する。モデルの学習データ量が限られている場合には、入力特徴量そのものの識別能力を高めることが有効であることが知られており、本研究では唇画像モダリティを使用して、より識別的な特徴量を得る。提案するマルチモーダル音声認識では、音素識別のための畳み込みニューラルネットワーク（CNN: convolutional neural network）からボトルネック特徴量を計算する。CNN のシフト不変性により、音声の局所的な時間変動や唇画像の切り出しズレなどに頑健な特徴抽出を行う。ボトルネック特徴量は、隣接層と比べて隠れ素子数の小さな層から抽出される特徴量であり、音素の識別に有効な情報を集約させることができる。単語認識による評価実験では、従来の音響特徴量とマルチモーダル特徴量と比較を行っており、提案手法の有効性を示している。

最後に第七章にて、全体を通してのまとめと、今後の課題について述べている。

提案手法は、障害者支援技術としてだけでなく、十分な学習データ量を用意することが難しい領域においても応用可能である。各提案の貢献はそれぞれ「行列分解によるパラメータ数の小さいコンパクトなモデル構造の実現」、「共通の特徴を有する他のデータベースの利用可能化」、「複数モダリティ利用による特徴量の識別能力の向上」として解釈可能である。実用シーンでは、対象となる環境のデータ量は少ないことが多いため、本提案が応用できると考えられる。

以上のように、本研究は、構音障害者の音声を対象としたコミュニケーション支援技術について、少量学習データによる声質変換と音声認識の新しいアルゴリズムを研究したものであり、障害者の声質変換と音声認識について重要な知見を得たものとして価値ある集積である。提案された論文はシステム情報学研究科学位論文評価基準を満たしており、学位申請者の高島悠樹は、博士（工学）の学位を得る資格があると認める。