公募シンポジウム
公募シンポジウム7
歯科医療情報共有化と異分野融合によるデータ駆動型時代の歯科医療
2023年11月24日(金) 13:30 ～ 16:00 C会場 (EX1-B)

[3-C-3-04] 発話困難者の音声コミュニケーションを支援する音声認識技術

*滝口 哲也[1]、北条 直樹[1]、高島 遼一[1]、杉山 千尋[2]、田中 信和[2]、野原 幹司[2]、野崎 一徳[2]（1. 神戸大学大学院システム情報学研究科、2. 大阪大学歯学部附属病院）
*Tetsuya Takiguchi[1], Naoki Hojo[1], Ryoichi Takashima[1], Chihiro Sugiyama[2], Nobukazu Tanaka[2], Kanji Nohara[2], Kazunori Nozaki[2]（1. Graduate School of System Informatics, Kobe University, 2. Osaka University Dental Hospital）
キーワード：Automatic Speech Recognition, Organic Articulation Disorder, Deep Neural Networks

近年，スマートフォンやスマートスピーカーなど音声を用いた端末入力サービスが発表されている．これらのサービスでは，音声で機器を操作して天気予報や店情報を検索し，また家電操作などが可能であるが，明瞭な発話ができる人を対象としており，発話困難者には対応していない．発話困難者の発話音声の特徴は，その原因（舌切除，脳性麻痺など）によりさまざまであり，聞き取りが困難な発話も存在する．そのような方々の発話音声コミュニケーションを支援するための音声研究への期待は大きい．そこで本発表では，発話困難者として器質性構音障害者を対象とした音声認識技術について紹介する．

発話困難者の発話スタイル特性は一人一人多様なため，カスタマイズされた特定話者専用の音響モデル構築が必要となる．またそのモデル構築には大量の（特定話者）モデル学習用音声データが必要である．通常は当事者による原稿（テキスト）読み上げ音声を収録して，モデル学習用データ（音声＋テキスト）として使用するが，当事者への負担などを考慮すると，原稿読み上げデータのみで十分な量の学習用データを集めるのは難しい場合がある．そのような学習データ量が少ない課題に対する解決策として，日常生活における自由発話音声を学習データとして活用することが考えられる．自由発話音声を音声認識の学習に使用するためには，音声に対応するテキスト書き起こしを手動で行う必要があるが，聞き取りが難しい発話困難者の音声を書き起こすことは難しい．そこで本発表では，器質性構音障害者の書き起こしテキストの無い音声データを活用した音声認識モデルの学習について報告する．

発話困難者の音声コミュニケーションを支援する音声認識技術

滝口 哲也[*1]、北条 直樹[*1]、高島 遼一[*1]、杉山 千尋[*2]、田中 信和[*2]、野原 幹司[*2]、野崎 一徳[*2]
*1 神戸大学大学院システム情報学研究科、
*2 大阪大学歯学部附属病院

Automatic Speech Recognition for Supporting Speech Communication of Persons with Speech Disorders

Tetsuya Takiguchi[*1], Naoki Hojo[*1], Ryoichi Takashima[*1],
Chihiro Sugiyama[*2], Nobukazu Tanaka[*2], Kanji Nohara[*2], Kazunori Nozaki[*2]
*1 Graduate school of system informatics, Kobe University,
*2 Osaka University Dental Hospital

Conventional approaches based on deep neural networks to automatic speech recognition usually require a large amount of speech data. But it is very difficult for persons with articulation disorders, in particular, to utter a large amount of speech data, and their utterances are often unstable. In this paper, a pre-trained model, wav2vec 2.0, using a large amount of speech data of physically unimpaired persons is applied to speech recognition for persons with organic articulation disorders, where the model is based on a framework for self-supervised learning of speech representations on unlabeled data for speech processing. Our experiment results show the effectiveness of the pre-trained wav2vec 2.0 using a large amount of speech data of physically unimpaired persons and also fine-tuning using a small amount of speech data of a person with an organic articulation disorder.

Keywords: automatic speech recognition, organic articulation disorder, deep neural networks

1. はじめに

近年、スマートフォンやスマートスピーカーなど音声を用いた端末入力サービスが発表されている。これらのサービスでは、音声で機器を操作して天気予報や店情報を検索し、また家電操作などが可能であるが、多くのサービスでは明瞭な発話ができる人を対象としており、発話困難者の使用が困難な場合がある。発話困難者の発話音声の特徴は、その原因（舌切除，脳性麻痺など）によりさまざまであり、聞き取りが困難な発話も存在する。そのような方々の発話音声コミュニケーションを支援するための音声研究への期待は大きい。これまで我々は、脳性麻痺者、重度難聴者、脊髄性筋萎縮症者らの音声認識や音声変換に関するコミュニケーション支援技術の研究 [1) 2)] を進めてきている。本稿では、発話困難者として器質性構音障害者を対象とした音声認識について紹介する。

発話困難者の発話スタイル特性は一人一人多様なため、カスタマイズされた特定話者専用の音響モデル構築が必要となる。またそのモデル構築には大量の（特定話者）モデル学習用音声データが必要である。通常は当事者による原稿（テキスト）読み上げ音声を収録して、モデル学習用データ（音声＋テキスト）として使用するが、当事者への負担などを考慮すると、原稿読み上げデータのみで十分な量の学習用データを集めるのは難しい場合がある。そのような学習データ量が少ない課題に対する解決策として、日常生活における発話困難者の自由発話音声を学習データとして活用することが考えられる。しかし自由発話音声を音声認識の学習に使用するためには、音声に対応するテキスト書き起こしを手動で行う必要があるが、聞き取りが難しい発話困難者の音声を書き起こすことは難しい。そこで近年研究されているテキスト不要で音声特徴表現を学習できる自己教師あり学習に注目する。これまでに我々は wav2vec 2.0 [3)] の自己教師あり学習モデルを用いて、脳性麻痺者の音声認識において有効性を確認 [4)] している。本稿では、大量に準備することが可能な健常者音声データを wav2vec 2.0 の事前学習に活用すること、また日常生活における発話困難者の自由発話音声を学習データとして活用することを検討し、器質性構音障害者の音声認識実験にて有効性を報告する。

2. 自己教師あり学習：発話テキスト無しによる音声特徴表現の学習

本稿では、近年の音声分野において注目されている wav2vec 2.0 を自己教師あり学習のモデルとして使用する。wav2vec 2.0 は、音声波形から音声の潜在表現を抽出する CNN（Convolutional Neural Network）エンコーダと、一部マスクされた潜在表現からコンテキスト表現を学習する Transformer エンコーダで構成されている。我々の先行研究 [4)] で提案した wav2vec 2.0 を用いた音声認識の学習手順を図 1 に示す。自己教師あり学習のフェーズでは、発話困難者の音声特徴表現を学習することが目的であるため、本来は発話困難者のテキスト無し音声のみを用いて自己教師あり学習を行うことが望ましい。しかし発話困難者のテキスト無しの音声データの収録が（原稿読み上げ音声データより）容易になるとはいえ、健常者音声のように wav2vec 2.0 を学習できるほど大量に準備することは現実的には困難である。そのため、事前に大規模な健常者音声を用いて wav2vec 2.0 の自己教師あり学習を行う。次に、健常者音声で学習した wav2vec 2.0 の学習済みモデルを初期値として、発話困難者のテキスト無し音声を用いて自己教師あり学習フェーズにおける fine-tuning を行う。

音声認識モデルの学習フェーズでは、評価者により発話された少量のテキスト付き音声を用いて、特定話者の音声認識モデルの教師あり学習を行う。音声認識モデルとして、

wav2vec 2.0 とその後続の線形層で構成される CTC (Connectionist Temporal Classification)[5] のモデルを使用し、wav2vec 2.0 は自己教師あり学習により学習されたパラメータを初期値として、少量の発話テキスト付き音声を用いてモデル(Neural Networks)の学習を行う。

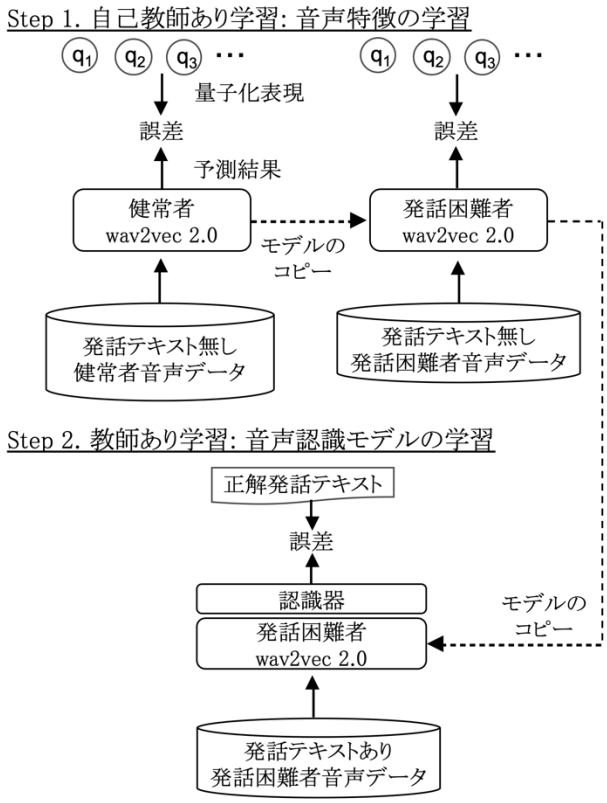

図 1　発話テキスト無し音声を活用した音声特徴の学習

3. 器質性構音障害者の音声認識実験

3.1 実験条件

本稿では、8 名(口唇口蓋裂 2 名、舌癌 6 名)の器質性構音障害者の音声データを収録し評価を行った。読み上げテキストとして、ATR 日本語音声データベースに含まれる音素バランス 503 文を利用した。各話者が 503 文を発話しており、そのうち 50 文を評価データ、別の 50 文を学習時の検証データ、残りを学習データとした。図 1 に示される発話テキスト無しの健常者音声データベースとしては、日本語話し言葉コーパス[6] を使用し、約 660 時間の音声が含まれている。

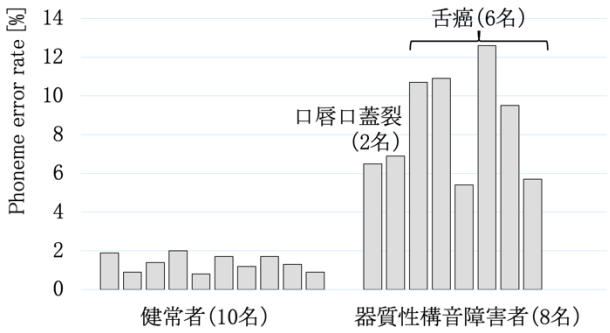

図 2　健常者 wav2vec を用いた音声認識誤り率

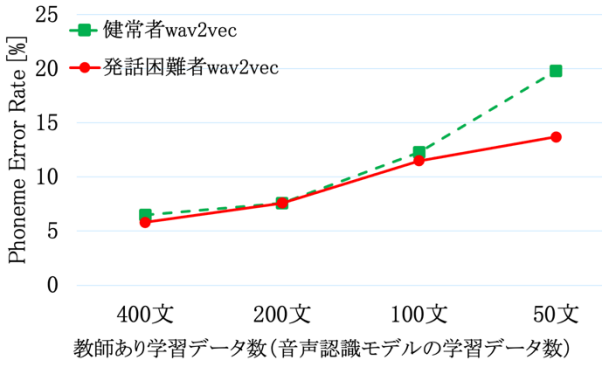

図 3　発話困難者 wav2vec を用いた音声認識誤り率

3.2 音声認識実験結果

図 2 に健常者 wav2vec 2.0 を用いた音声認識実験の結果を示す。ここで使用しているモデルでは、図 1 の Step 1 の発話テキスト無し発話困難者音声データでの学習をせずに、Step 2 の音声認識モデルの学習を行っている。図 2 に示す認識実験結果より、健常者の平均音素誤り率は 1.38%であるが、器質性構音障害者の平均音素誤り率は 8.5%であり、誤り率は健常者発話に比べると高く、また発話者ごとに精度のばらつきが大きくなっているのが分かる。

次に口唇口蓋裂 1 名の発話テキスト無し音声データ 4,495 文を用いて発話困難者 wav2vec 2.0 を学習したときの音声認識結果を図 3 に示す。音声認識モデルの教師あり学習データ数を変化させて実験を行ってみたところ、約 400 文(395 文)では、健常者 wav2vec と比べてあまり大きな変化が見られないが、100 文、50 文と学習データ数が少ない場合、発話困難者の wav2vec の効果が大きいのが分かる。

4. おわりに

本稿では、器質性構音障害者の音声認識において、発話テキストなし音声データを活用した wav2vec 2.0 の自己教師あり学習を適用し、その効果を検証してみた。今後は、発話テキストなし(少量)音声データ数に応じた wav2vec モデルの最適な構造、および学習方法について検討する。また、器質性構音障害者の発話ばらつき、誤り傾向などについても解析を進めていく予定である。

参考文献

1) 滝口哲也. 構音障害者のための話者性を維持した音声変換. 日本音響学会誌 2018 ; 74 : 144-147.

2) Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki. Two-step acoustic model adaptation for dysarthric speech recognition. IEEE ICASSP 2020 : 6104-6108.

3) A. Baevski, Y. Zhou, A. Mohamed, M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. NeurIPS 2020 : 12449-12460.

4) 松坂勇樹, 高島遼一, 滝口哲也. wav2vec 2.0 によるラベル無し音声を用いた脳性麻痺者の音声認識. 日本音響学会秋季研究発表会講演論文集 2022 : 1317-1320.

5) A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber. Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks. ICML 2006 : 369-376.

6) K. Maekawa. Corpus of spontaneous Japanese: Its design and evaluation. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition 2003 : 7-12.