wav2vec 2.0 によるラベル無し音声を用いた脳性麻痺患者の音声認識*

☆松坂勇樹, 高島遼一, 滝口哲也 (神戸大)

1 はじめに

構音障害とは，言葉を理解しているにもかかわらず，発声器官や神経などの異常により，言葉を正しく発話できない状態である．構音障害となる原因や病気はいくつか存在し，脊髄性筋萎縮症，口唇口蓋裂などが例として挙げられる．本研究では脳性麻痺 (Cerebral Palsy; CP) による構音障害を対象とし，中でもアテトーゼ型脳性麻痺に着目する．アテトーゼ型脳性麻痺患者は，意図した動作の際に筋肉の不随意運動を伴う．不随意運動は発話時にも起こり，これが構音障害の直接的な原因となっている．

脳性麻痺患者の多くは，筋肉の不随意運動や筋緊張のため，手足を自由に動かすことが困難であり，日常生活に支障をきたしている．このような背景から，音声認識 (automatic speech recognition; ASR) を用いたアシスタントシステムが，ハンズフリー入力可能な脳性麻痺患者の生活支援技術として期待されている．しかし，構音障害者の発話は健常者の発話特徴と大きく異なるため，健常者の音声で学習された従来の ASR システムでは構音障害者の発話を正確に認識することは困難である．したがって，構音障害者本人の音声を用いて ASR モデルの学習をする必要がある．しかし，脳性麻痺患者にとって音声の収録には身体への負担が大きいため，ASR モデル学習用の音声データを十分量用意することが困難であるという問題がある．

構音障害者の音声認識において，学習データ不足の改善手法はいくつか提案されてきた．代表的な手法として，事前に大量の健常者音声により ASR モデルを学習し，その後少量のラベル付き構音障害者音声を用いてモデルをファインチューニングする方法がある [1]．また，収録音声に対してデータ拡張を行うことで，モデル学習用のデータを増量するアプローチも研究されている [2, 3]．これらの手法は，構音障害者音声が少量に限られている条件下で音声認識性能を向上させるものであり，実際にその効果は確認されている．しかしながら健常者音声と比較して構音障害者音声は圧倒的に少ないことを鑑みると，健常者と同等にまで音声認識性能を向上させるためには，やはり構音障害者の音声をより多く収録するための研究が必要であると考えられる．

本研究では，使用可能な収録音声として，収録時に負担が大きいラベル付き音声だけではなく，より負担が小さいラベル無し音声を活用したアプローチに着目する．ラベル無し音声の活用法として wav2vec 2.0 を用いた自己教師あり学習を検討し，アテトーゼ型脳性麻痺患者の音声認識における認識精度の向上を検討する．

2 脳性麻痺患者のラベル無し音声の収録

脳性麻痺患者の音声を収録するためのアプローチとして，大きく分けて 2 つ挙げられ，それぞれの利点と欠点を示す．

一つ目のアプローチは，台本を使用した読み上げ発話を収録する方法である．この方法の利点は，ラベルを後付けする必要がないことである．つまり，収録した音声はラベル付き音声として利用できる．しかし脳性麻痺患者に台本を読ませることは身体への負担が大きいため，音声を大量に収録することが困難であることが，欠点として挙げられる．

二つ目のアプローチとして，自由発話音声を収録する方法がある．この方法の利点は，日常生活での発話は読み上げ発話の収録と比較して身体への負担が小さいと考えられるため，より多くの音声データを収集できることである．しかしこの方法では，あらかじめ発話する内容が決まっていないため，ラベルを後付けする必要がある．脳性麻痺患者の音声は人間でも聞き取りづらいため，したがって人手でのラベル付与が困難であることが欠点である．

従来研究では台本読み上げ発話を収録するアプローチがほとんどであったが，本研究では，音声データをより多く収集できるという利点に着目し，自由発話音声を利用することを検討する．ただし，ラベルの付与が困難という欠点があるため，音声に対してラベルの付与は行わない．すなわち，脳性麻痺患者のラベル無し自由発話音声を音声認識モデルの学習に使用することが，本研究の目的である．

3 wav2vec 2.0 による自己教師あり学習

ラベル無し音声を活用する方法はいくつか存在する．代表的な手法の一つに，ラベル無し音声に対して音声認識を行うことで擬似的なラベルを付与し，ASR モデルの学習データとして利用する疑似ラベリングの方法 [4] がある．また，近年研究されている手法と

*Speech recognition for cerebral palsy patients using unlabeled speech on wav2vec 2.0. by Yuki Matsuzaka, Ryoichi Takashima, Tetsuya Takiguchi (Kobe University)

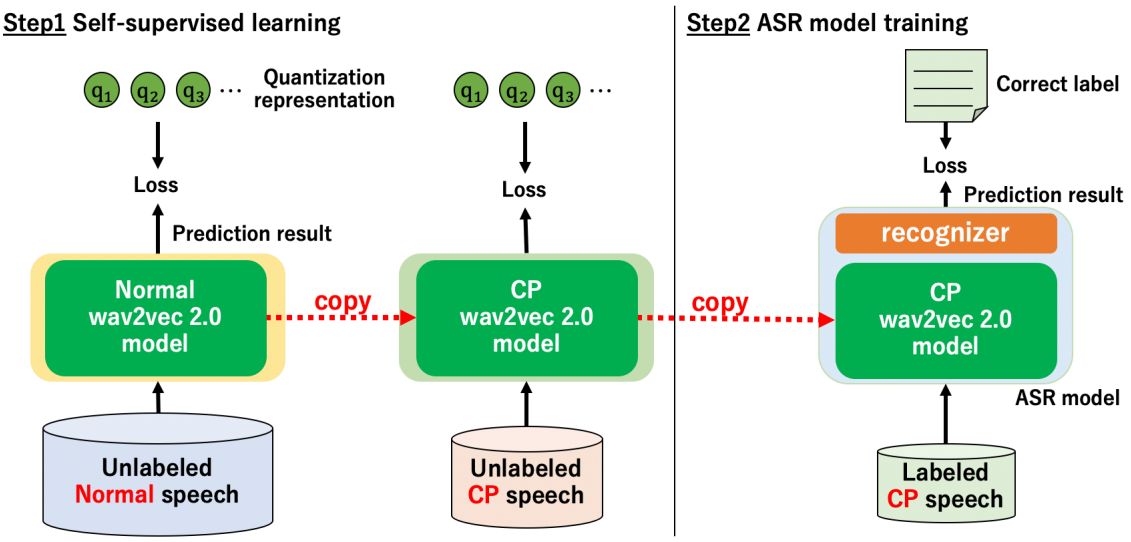

Fig. 1　Training procedure using wav2vec 2.0 for cerebral palsy (CP) speech recognition.

して，ラベル無し音声でも学習が可能な自己教師あり学習により特徴表現を学習し，ASR モデル学習時にはその学習済みモデルに対してラベル付き音声でファインチューニングを行う手法 [5] がある．本研究では後者の自己教師あり学習の枠組みにより，脳性麻痺患者のラベル無し音声を活用する．

脳性麻痺患者のラベル無し自由発話音声を用いて，自己教師あり学習を行った先行研究として文献 [6] がある．この研究では，自己教師あり学習の手法として Autoregressive Predictive Coding (APC) [7] を使用している．APC モデルは RNN と全結合層で構成されており，音声の現在までのフレームから未来のフレームを予測するタスクにより，音声の特徴表現を学習する．APC モデルによる自己教師あり学習で得られた特徴表現を音声認識に活用することで，脳性麻痺患者の音声認識における認識精度が向上したことが報告されている．

本研究でも同様にラベル無し音声を用いた自己教師あり学習を行うことで，脳性麻痺患者音声に対する認識精度の向上を目的とするが，自己教師あり学習のモデルとして wav2vec 2.0 [8] を使用する．wav2vec 2.0 は，音声波形から音声の潜在表現を抽出する CNN エンコーダと，一部が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．大量のラベル無しデータを用いて自己教師あり学習を行った後，wav2vec 2.0 は，その後のファインチューニングに用いるラベル付き音声が 10 分や 1 時間程度と少量であっても，高い認識率が得られることが報告されている．そのため，ラベル付き音声の収録が困難な脳性麻痺患者の音声認識において，有効であると期待される．

4　データの種類と学習手順

本研究で使用する音声データとして，主に 3 種類のデータを用意する．一つ目は脳性麻痺患者の台本読み上げによるラベル付きデータであり，少量データしか存在しない．二つ目は健常者音声であり，脳性麻痺患者と比べて大量に存在する．構音障害者を対象とした音声認識の研究では，これら 2 種類のデータが主に使用されている．少量の脳性麻痺患者のラベル付きデータでは，ASR モデルの学習が不十分となるため，事前学習用に健常者音声が使用されている．本研究ではこれらのデータに加えて，三つ目のデータとして脳性麻痺患者のラベル無し音声を使用する．このラベル無し音声のデータ量は，健常者音声と比較すると少量ではあるものの，ラベル付き音声よりは多く存在する．

Fig. 1 に，本研究で提案する wav2vec 2.0 を用いた音声認識の学習手順を示す．Step1 として，まず自己教師あり学習を行う．脳性麻痺患者音声の特徴表現を学習することが目的であるため，本来は脳性麻痺患者のラベル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．しかし現時点で我々が所有しているラベル無し音声は，ラベル付き音声より多いとはいえ，wav2vec 2.0 を学習できるほど多くはない．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モデルとして，脳性麻痺患者のラベル無し音声を利用して再度自己教師あり学習を行う．次に Step2 として，脳性麻痺患者のラベル付き音声を用いて，ASR モデルの教師あり学習を行う．このとき，ASR モデルの構造としては，自己教師あり学習時のモデルと同

Table 1　Speech dataset for experiment.

Speaker	label	content
CP	✓	ATR503 (429utts)
		Lecture & newspaper
Normal		CSJ

じ構造を含めており，脳性麻痺患者のラベル無し音声による学習済みモデルを初期モデルとしている．また文献 [8] に倣って，自己教師あり学習時のモデル構造に加えて，後続の層に認識器として線形層と CTC [9] を加える．

5　評価実験

5.1　実験条件

Table 1 に本実験で使用する音声データを示す．脳性麻痺患者の対象話者として，アテトーゼ型脳性麻痺患者の男性 1 名の音声を使用する．ラベル付き脳性麻痺患者音声として，ATR 日本語データベース [10] に含まれる音素バランス文 503 文のうち，429 文 (約 50 分) の読み上げ発話を収録している．そのうち 329 文を学習データ，50 文を開発データ，50 文を評価データに使用した．ラベル無し脳性麻痺患者音声として，患者による講演音声および新聞の読み上げ音声[1]を計約 3 時間収録した．ラベル無し音声を用いて自己教師あり学習を行う際には，ラベル付き音声における学習データと開発データも加えて使用している．この理由としては，脳性麻痺患者のラベル無し音声は現状では大量に用意できていないため，ラベル付き音声を加えることでデータ量を増やそうとしたこと，また，事前にラベル付きデータを自己教師あり学習時に含めておくことで，後の教師あり学習で使用するデータセットと近いドメインのデータセットで自己教師あり学習を行うためである．健常者音声には，日本語話し言葉コーパス (CSJ) [11] を使用し，約 660 時間の音声が含まれている．また，LibriSpeech(約 960 時間) による学習済みモデル (Wav2Vec 2.0 Base, No finetuning[2]) も健常者モデルとして使用し，CSJ データセットで学習した場合と比較を行う．

自己教師あり学習における wav2vev 2.0 のモデルは，文献 [8] を参考にして，Base モデルと同じ構造にした．CNN エンコーダは 7 ブロックで構成されており，チャネルサイズは 512，カーネルサイズは各ブロックごとに [10,3,3,3,3,2,2]，ストライドは各ブロックごとに [5,2,2,2,2,2,2] としている．Transformer エンコーダは 12 ブロックで構成されており，モデル次元は 768，内部次元は 3,072 としている．

音声認識モデルには，wav2vec 2.0 の Base モデルの後続の層に，線形層 1 層と CTC を追加した．本実験では音素単位での認識を行うため，出力層は 39 種類の音素に加えて CTC の blank トークン，未知トークン ⟨unk⟩，始端/終端記号 ⟨sos/eos⟩ からなる計 42 種類のトークンで定義した．オプティマイザには AdaDelta を使用し，認識の際には検証損失が最小のエポックを採用した．

[1]新聞読み上げ音声は実際はラベルが存在することになるが，本実験ではラベル無し自由発話と見立てて使用している．

[2]https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec

Table 2　PER for Cerebral Palsy (CP) Patients Using wav2vec 2.0 self-supervised learning (SSL).

SSL [Normal]	SSL [CP]	PER[%]
	✓	82.2
✓(LibriSpeech)		29.9
✓(LibriSpeech)	✓	26.5
✓(CSJ)		23.5
✓(CSJ)	✓	21.3

5.2　実験結果

wav2vec 2.0 を用いた自己教師あり学習による脳性麻痺患者の音声認識の実験結果を Table 2 に示す．自己教師あり学習時における学習手順や学習データの構成による結果を比較しており，音素誤り率 (phone error rate; PER) で評価を行った．

まず，健常者音声を使用しない場合の結果では認識性能が悪く，大規模な健常者音声を使用することにより大幅な性能向上が確認できる．健常者音声を使用しない場合に性能が悪い点については，現状の脳性麻痺患者のラベル無し音声のデータ量では wav2vec 2.0 の大規模なモデル学習に対して不十分であることが原因として考えられる．また，健常者音声として LibriSpeech よりも CSJ データセットを使用した場合のほうが良い性能となっている．この理由としては，脳性麻痺患者の音声は日本語で発話されているため、日本語の大規模データによる事前学習が有効であったためと考えられる．そして，健常者音声による学習済みモデルを初期モデルとして，脳性麻痺患者のラベル無し音声による自己教師あり学習を行うことで，認識性能がさらに向上した．この理由としては，自己教師あり学習において，より脳性麻痺患者音声に適した特徴表現が学習できたためと考えられる．

5.3　ラベル付きデータ量による認識性能の変化

ASR モデルの学習に使用するラベル付き学習データの発話数を変更し，認識性能の比較を行った．Table 3 に，使用するラベル付き学習データの発話数と認識

Table 3　Comparison of PER with number of training labeled data for patients with cerebral palsy.

SSL [Normal]	SSL [CP]	Number of labeled data	PER[%]
✓(CSJ)		50	30.7
	✓		27.0
		200	26.0
	✓		23.6
		329	23.5
	✓		21.3

性能の結果を示す．比較する各発話数において，ラベル付きデータ 50 発話は約 6 分，200 発話は約 24 分，329 発話は約 50 分の発話時間である．また，自己教師あり学習では健常者音声として CSJ データセットを用いており，各発話数において脳性麻痺患者のラベルなし音声を用いた自己教師あり学習の有無による比較を行っている．

表の結果より，比較する全ての発話数において，脳性麻痺患者のラベルなし音声を利用した自己教師あり学習が有効であることが確認できた．また，ラベル付き音声が非常に少量の 50 発話であっても，脳性麻痺患者のラベル無し音声を利用することで 30%を下回る認識性能が得られた．ラベル付き音声が非常に少量でも認識性能が大きく劣化しないことは，脳性麻痺患者にとって負担が大きいラベル付き音声の収録回数が少なく済むという利点がある．

6　おわりに

本研究では，脳性麻痺患者の音声認識における学習データ不足の問題を解決するために，読み上げ発話音声より比較的多く収録可能なラベル無し自由発話音声を音声認識に活用することを検討した．また，ラベル無し音声の活用法として wav2vec 2.0 による自己教師あり学習を行い，脳性麻痺患者のラベル無し音声を使用することで認識性能が向上することを確認した．また，wav2vec 2.0 は非常に少量のラベル付きデータでも認識性能が大きく劣化しないことが確認されたことから，脳性麻痺患者の音声認識において今後も期待できるモデルであった．

今後の拡張としては，文献 [12, 13] を参考にして，wav2vec 2.0 の自己教師あり学習に加えて，疑似ラベリングを併用した学習方法を検討していく．さらに，今回は音素認識で実験を行ったが，wav2vec 2.0 を用いた文字認識を今後検討する．ただし，ラベルの種類が少ない音素とは異なり，文字単位の認識ではラベルの種類が増えるため，言語モデルの使用も検討する．

謝辞　本研究の一部は，JSPS 科研費 JP21H00906, JP22K12168 の支援を受けたものである．

参考文献

[1] R. Takashima *et al.*, “Two-step acoustic model adaptation for dysarthric speech recognition,” in *ICASSP*, pp. 6104-6108, 2020.
[2] K. Fujiwara *et al.*, “Data augmentation based on frequency warping for recognition of cleft palate speech,” in *APSIPA*, pp. 471-476, 2021.
[3] Y. Matsuzaka *et al.*, “Data augmentation for dysarthric speech recognition based on text-to-speech synthesis,” in *LifeTech*, pp. 399-400, 2022.
[4] J. Kahn *et al.*, “Self-training for end-to-end speech recognition,” in *ICASSP*, pp. 7084-7088, 2020.
[5] W. Wang *et al.*, “Unsupervised pre-training of bidirectional speech encoders via masked reconstruction,” in *ICASSP*, pp. 6889-6893, 2020.
[6] 澤佑哉 *et al.*, “自己教師あり学習によるラベル無し自由発話を用いた構音障害者音声認識,” 日本音響学会春季研究発表会講演論文集, pp. 1045-1048, 2021.
[7] Y.-A. Chung *et al.*, “An unsupervised autoregressive model for speech representation learning,” in *Interspeech*, pp. 146-150, 2019.
[8] A. Baevski *et al.*, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *arXiv preprint arXiv:2006.11477*, 2020.
[9] A. Graves *et al.*, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *ICML*, pp. 369-376, 2006.
[10] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, Vol. 9, No. 4, pp. 357-363, 1990.
[11] K. Maekawa, “Corpus of spontaneous Japanese: Its design and evaluation,” in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pp. 7-12, 2003.
[12] Q. Xu *et al.*, “Self-training and pre-training are complementary for speech recognition,” in *ICASSP*, pp. 3030-3034, 2021.
[13] 澤佑哉 *et al.*, “疑似ラベリングと特徴表現学習を併用した構音障害者音声認識,” 日本音響学会秋季研究発表会講演論文集, pp. 847-850, 2021.