wav2vec 2.0 によるラベル無し複数患者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 (Dysarthria) とは，言葉を理解しているが，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，脳性麻痺 (Cerebral Palsy; CP)，口唇口蓋裂，舌癌治療による舌摘出などが例として挙げられる．本研究では脳性麻痺による構音障害を評価対象とし，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，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) を用いたハンズフリー入力デバイスが彼らのための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]．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の向上が確認されている．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考慮すると，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に限らず，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．

近年では，自己教師あり学習 (SSL) がラベル無しのデータを用いて特徴表現を学習できる手法として注目されている．音声認識に有効な SSL モデルは多く提案されており，音声のフレーム特徴量を予測する APC モデル [3] のような生成タスクの SSL モデル，音声特徴量の一部をマスキングし，マスクされた特徴量を識別するタスクの SSL モデル [4] などがある．

以前，我々は wav2vec 2.0 [4] の SSL モデルを用いて脳性麻痺患者の音声認識を行い，評価話者本人のラベル無し音声を用いた自己教師あり学習の有効性を確認した．本研究では，さらに評価話者本人のラベル無し音声を使用する場合に限らず，評価話者とは別の患者も含めた複数患者の音声による自己教師あり学習の有効性を検証する．

2 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 [5] がある．また，近年研究されている手法として，ラベル無し音声でも学習が可能な自己教師あり学習 (SSL) 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を初期値として，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，ラベル無し患者音声を活用する．

自己教師あり学習のモデルとして wav2vec 2.0 を使用する．これは，近年の音声分野における SSL の研究でよく使用されているモデルである．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．我々は以前の研究で，脳性麻痺患者の音声認識においても wav2vec 2.0 が有効であることを確認した [6].

3 複数患者音声を用いた自己教師あり学習

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

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

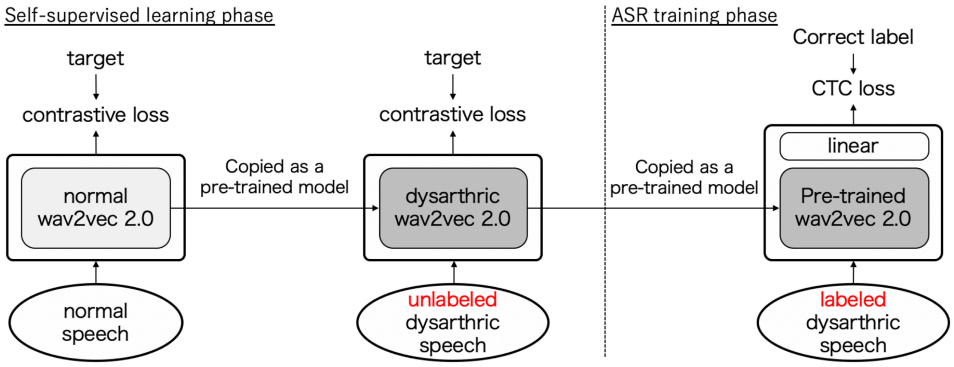

Fig. 1　Proposed training procedure using unlabeled dysarthric speech based on wav2vec 2.0.

読み上げによるラベル付き患者音声であり，少量データしか存在しない．二つ目は健常者音声であり，脳性麻痺患者の音声と比べて大量に存在する．三つ目は脳性麻痺患者によるラベル無し患者音声である．このラベル無し患者音声のデータ量は，健常者音声と比較すると少量ではあるが，ラベル付き患者音声よりは多く存在する．

我々の先行研究 [6] で提案した，wav2vec 2.0 を用いた音声認識の学習手順を Fig. 1 に示す．自己教師あり学習のフェーズでは，患者音声の特徴表現を学習することが目的であるため，本来はラベル無し患者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，ラベル無し患者音声は収録しやすいとはいえ，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，ラベル無し患者音声を用いて自己教師あり学習フェーズにおけるファインチューニングを行う．

自己教師あり学習のファインチューニングにおいて，我々の先行研究 [6] では，評価話者によって発話されたラベル無し患者音声のみを用いていた．これは，ラベル無し音声を収録している評価話者のみ可能となる学習方法である．この方法に対して，本研究では評価話者以外のラベル無し患者音声を利用する以下 2 種類の方法を検証する．一つ目は，評価話者と同じ疾患 (つまり脳性麻痺) を持つ複数の患者によるラベル無し患者音声を用いる方法である．この方法の狙いは，複数患者のラベル無し音声を利用することで，自己教師あり学習時における患者音声のデータ量を増やすこと，また，ラベル無し音声を収録できていない話者であっても，複数患者の音声で代用することで，患者音声による自己教師あり学習を可能にすることである．二つ目は，一つ目のデータセットに加えて，評価話者と異なる疾患を持つ複数の患者によるラベル無し患者音声も用いる方法である．本実験では，脳性麻痺の患者に加えて，口唇口蓋裂 (CLP) 患者と舌摘出者の音声を使用する．

ASR モデルの学習フェーズでは，評価患者によって発話された少量のラベル付き患者音声を用いて，特定患者 ASR モデルの教師あり学習を行う．ASR モデルとして，原著論文 [4] に倣って wav2vec 2.0 とその後続の線形層で構成される CTC [7] のモデルを使用し，wav2vec 2.0 は自己教師あり学習によって学習されたパラメータを初期値とする．本論文では比較実験として，自己教師あり学習に用いた音声データセットごとの認識性能で比較を行う．

4　評価実験

4.1　データ設定

Table 1 に本実験で使用する音声データを示す．脳性麻痺 (CP) の日本人患者 4 名を評価話者としており，CP-SPK1 ではラベル付き患者音声として，ATR 日本語音声データベース [8] 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おり，そのうち 329 発話を学習データ，50 発話を検証データ，50 発話を評価データに使用した．ラベル無し患者音声として，講演音声および新聞の読み上げ音声を計約 3 時間収録した[1]．CP-SPK2 ではラベル付き患者音声として同様の ATR503 文のうち，501 文 (約 80 分) の読み上げ発話を収録しており，そのうち 402 発話を学習データ，50 発話を検証データ，49 発話を評価データに使用した．CP-SPK3 でも同様に 181 文の収録音声 (学習データ 100 発話，検証データ 31 発話，評価データ 50 発話)，CP-SPK4 では 177 文の収録音声 (学習データ 100 発話，検証データ 27 発話，評価データ 50 発話) をラベル付き患者音声として用意した．ただし，CP-SPK2，CP-SPK3，CP-SPK4 の 3 話者はラベル無し患者音声を用意しておらず，これが CP-SPK1 と大きく異なる点である．

[1] 実際には読み上げ音声 (ラベル付き音声) として収録しているが，本研究ではラベル無し音声として使用する．

Table 1　Dysarthric speech used in the experiment.

Speaker	Label	Content
CP-SPK1	✓	ATR503 (429 utterances)
		unlabeled speech (about 3 hours)
CP-SPK2	✓	ATR503 (501 utterances)
CP-SPK3	✓	ATR503 (181 utterances)
CP-SPK4	✓	ATR503 (177 utterances)
CP-ALL		CP-SPK1 + CP-SPK2 + CP-SPK3 + CP-SPK4 (about 6 hours)
DYS-ALL		CP-ALL + CLP & TR speech (about 15.5 hours)

そのため，複数患者によるラベル無し患者音声 (e.g., CP-ALL) を自己教師あり学習の際に代用する．

複数患者をまとめた音声のサブセットとして，脳性麻痺の患者 4 名の音声をまとめた CP-ALL(約 6 時間) を用意した．CP-ALL では，患者 4 名の音声のうち，評価発話を除いた音声を全て含めている．また，複数の疾患における音声をまとめた DYS-ALL(約 15.5 時間) も用意しており，CP-ALL に加えて，口唇口蓋裂 (CLP) の日本人患者 2 名の音声 (約 5 時間)，舌摘出者 (TR) の日本人患者 5 名の音声 (約 4.5 時間) が含まれる．健常者音声としては，日本語話し言葉コーパス (CSJ) [9] を使用し，約 660 時間の音声が含まれている．これは wav2vec 2.0 の自己教師あり学習のフェーズにおける事前学習に使用する．

4.2　モデル設定

自己教師あり学習における wav2vev 2.0 のモデルは，原著論文 [4] 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．

ASR モデルへのファインチューニングは espnet [10] を用いて行った．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．

4.3　実験結果

4.3.1　CP-SPK1 の実験結果

評価話者本人のラベル無し患者音声，及びラベル付き患者音声の両方を持つ評価患者として，CP-SPK1 に関する比較結果を Table 2 に示す．評価指標として

Table 2　PERs[%] for CP-SPK1 on self-supervised learning of wav2vec 2.0 using unlabeled speech.

SSL [normal]	SSL [patient]	PER[%]
labeled training data 329 utterances		
	CP-SPK1	51.3
	CP-ALL	47.7
	DYS-ALL	38.1
CSJ		23.5
CSJ	CP-SPK1	**21.3**
CSJ	CP-ALL	22.1
CSJ	DYS-ALL	23.4
labeled training data 50 utterances		
CSJ		30.7
CSJ	CP-SPK1	**27.0**
CSJ	CP-ALL	27.4
CSJ	DYS-ALL	29.2

は，音素誤り率 (Phone Error Rate; PER) を用いており，自己教師あり学習に使用したデータセットによる認識性能を比較している．また，ASR モデルの学習時に使用する評価話者のラベル付き学習データが 329 発話，50 発話の場合それぞれにおいて同様の比較を行っている[2]．

まず，ラベル付き学習データを 329 発話使用した場合において，健常者音声を使用しない場合は認識性能が悪く，複数患者を用いた場合においても最大 15.5 時間程度であり，wav2vec 2.0 の学習には不十分であることがわかる．しかし，大量の健常者音声である CSJ データセットを事前学習として使用することで，大幅な改善が確認できる．次に，CSJ による健常者事前学習を初期モデルとして，評価話者のラベル無し音声のみを使用した場合 (CP-SPK1)，同じ脳性麻痺の複数患者の音声セット (CP-ALL)，異なる疾患も含めた複数患者の音声セット (DYS-ALL) で学習した場合，全ての場合において音声認識性能が向上していることが確認でき，ラベル無し患者音声による自己教師あり学習の有効性がわかる．しかし、評価話者のラベル無し音声のみを使用した場合 (CP-SPK1) の方が性能が良く，複数患者による学習のメリットがないことがわかった．これは，患者音声の自己教師あり学習において，評価話者本人のラベル無し音声が占める割合が小さくなったことや，発話特徴の異なる音声を加えたことが悪影響を及ぼしたことが原因と考えられる．

ラベル付き学習データを 50 発話使用した場合[3]も同様に，ラベル無し患者音声による認識性能の向上が確認でき，ラベル付き学習データ 329 発話よりも向上率が高かった．これは，少量のラベル付き患者音

[2] ここで示しているラベル付き学習データは，検証データを含めていない．検証データは別に使用している (以降の表も同様)．
[3] ラベル付き学習データを 329 発話から 50 発話に減らした場合，ラベル無し患者音声も 279 発話分減らしている．

Table 3　PERs[%] for CP-SPK2, CP-SPK3 and CP-SPK4 on SSL of wav2vec 2.0 using unlabeled speech.

SSL [normal]	SSL [patient]	SPK2 PER	SPK3 PER	SPK4 PER
labeled training data 402 utterances				
CSJ		33.2	—	—
CSJ	CP-ALL	**29.2**	—	—
CSJ	DYS-ALL	33.0	—	—
labeled training data 100 utterances				
CSJ		42.0	24.3	44.7
CSJ	CP-ALL	**39.3**	**22.8**	**39.4**
CSJ	DYS-ALL	42.9	25.5	40.1
labeled training data 50 utterances				
CSJ		47.4	31.8	55.8
CSJ	CP-ALL	**44.9**	**26.2**	**45.6**
CSJ	DYS-ALL	51.8	29.0	45.8

声による不十分な学習を，ラベル無し患者音声が補ったことによるものと考えられる．しかし，ラベル付き学習データ 329 発話の場合と同様に複数患者による学習に関しては有効性が見られなかった．

4.3.2　CP-SPK2,3,4 の実験結果

評価話者本人のラベル無し患者音声が無く，ラベル付き患者音声のみを持つ話者として，CP-SPK2，CP-SPK3，CP-SPK4 の 3 人の評価話者に関する比較結果を Table 3 に示す．Table 2 と同様に，評価指標として音素誤り率 (PER)，自己教師あり学習に使用したデータセットによる認識性能を比較している．また，CP-SPK2 では，ASR モデル学習時に使用するラベル付き学習データが 402 発話，100 発話，50 発話の場合それぞれにおいて同様の比較を行っている．CP-SPK3，CP-SPK4 では，100 発話，50 発話の場合で比較を行っている．

表より，全ての話者，全てのラベル付き学習データ数において，ラベル無し患者音声の学習により，認識性能が向上していることが確認できる．これより，本人のラベル無し音声が収録できていない話者であっても，複数患者によるラベル無し患者音声を用いることで認識性能が改善することがわかる．また，複数患者の音声サブセットとして，異なる疾患を含めた DYS-ALL よりも，同じ疾患の患者を集めた CP-ALL の方が性能が良い傾向がある．このことから，本人のラベル無し音声の代用としては，同じ疾患である脳性麻痺患者のみでまとめる方が良いと考えられる．

5　おわりに

本論文では，脳性麻痺患者の音声認識において，複数患者によるラベル無し音声を用いた wav2vec 2.0 の自己教師あり学習を行った．実験の結果，評価話者本人のラベル無し音声を持つ患者 (CP-SPK1) に対しては，複数患者音声の学習は有効でなかったが，話者本人のラベル無し音声が未収録の患者に対しては，複数患者音声で学習することで性能が向上することがわかった．今後は，患者数やデータ量を増やした場合の実験や，より実用的な文字認識へ拡張することを検討する．

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] Y. Matsuzaka *et al.*, “Data augmentation for dysarthric speech recognition based on text-to-speech synthesis,” in *LifeTech*, pp. 399-400, 2022.

[3] Y.-A. Chung, W.-N. Hsu, H. Tang, J. Glass, “An unsupervised autoregressive model for speech representation learning,” in *Interspeech*, pp. 146-150, 2019.

[4] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *NeurIPS*, vol. 33, pp.12449-12460 , 2020.

[5] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, *et al.*, “Iterative pseudo-labeling for speech recognition,” in *Interspeech*, pp. 1006-1010, 2020.

[6] 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患者の音声認識,” 日本音響学会秋季研究発表会講演論文集, pp. 1317-1320, 2022.

[7] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *ICML*, pp.369-376, 2006.

[8] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, Vol. 9, No. 4, pp. 357-363, 1990.

[9] 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.

[10] S. Watanabe, T. Hori, S. Karita, T. Hayashi, *et al.*, “ESPnet: End-to-end speech processing toolkit,” in *Interspeech*, pp. 2207-2211, 2018.