wav2vec 2.0 と疑似ラベリングを活用した脳性麻痺者の音声認識 *

☆松坂勇樹 (神戸大), 高島遼一 (神戸大/JST さきがけ), 滝口哲也 (神戸大)

1 はじめに

構音障害 (Dysarthria) とは，発話したい内容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正しく発話できない症状のことを指す．構音障害の原因となる原因や疾患としては，脳性麻痺 (Cerebral Palsy; CP) や脊髄性筋萎縮症，口唇口蓋裂などが存在するが，本研究ではアテトーゼ型脳性麻痺に起因する構音障害を対象としている．アテトーゼ型脳性麻痺者は手足などの身体の部位において，不随意運動が生じてしまう．不随意運動は発話時にも影響を及ぼすため，それが構音障害の原因となっている．

脳性麻痺者の多くは，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で ASR モデルの学習をする必要がある．しかし，脳性麻痺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けるデータ不足の課題に対するアプローチはこれまで多く研究されてきた．代表的な手法の一つに，大量の健常者音声で事前学習を行い，後に少量のラベル付き構音障害者音声でファインチューニングをする方法がある [1]．また，収録した構音障害者音声に対してデータ拡張を施すことで，音声認識の学習に使用可能なデータを増やす方法がある [2]．上記のアプローチにより，構音障害者に対する音声認識精度が向上することを確認できたが，依然として収録音声は少量のままである．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集するための研究が必要であると考えられる．

本研究では，より多くの脳性麻痺者の音声を利用するために，脳性麻痺者のラベル無し音声を音声認識の学習に活用することを検討する．ラベル無し音声の活用法として，これまで我々の先行研究 [3] で提案した wav2vec 2.0 の自己教師あり学習による方法に加え，今回新たに疑似ラベリングの方法を追加で用いて実験を行う．また，これまでの研究と同様に音素認識による評価実験を行うだけでなく，文字認識による評価実験も実施する．

2 ラベル無し音声の活用法

ラベル無し音声を音声認識の学習に活用する代表的な手法としては，主に 2 つ挙げられる．一つ目は，ラベル無しデータを用いて特徴表現を学習可能な自己教師あり学習 (Self-supervised Learning; SSL) である．ラベル無し音声で自己教師あり学習を実施した後，ASR 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．二つ目は，ラベル無し音声に対して音声認識を行うことで，誤りも含めた疑似的なラベル (疑似ラベル) を生成し，ラベル無し音声に対応する正解ラベルの代わりとして用いることで，ASR モデルの学習データに活用する疑似ラベリング [4] の手法である．

我々の先行研究 [3] では，自己教師あり学習による活用法により，脳性麻痺者のラベル無し音声を活用した．自己教師あり学習のモデルとして wav2vec 2.0 [5] を使用しており，音声波形から音声の潜在表現を抽出する CNN 特徴量エンコーダと，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．文献 [5] では，大量のラベル無し音声で wav2vec 2.0 の自己教師あり学習を行うことで，その後の ASR 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．これはラベル付き音声の収集が困難な脳性麻痺者の音声認識において有効であると考えられ，我々の先行研究 [3] における実験でもその有効性を確認した．ただし，ラベル無し音声の活用法として，疑似ラベリングは用いていない．

文献 [6] では，自己教師あり学習と疑似ラベリングを併用することで，認識性能が向上することが報告されており，健常者の音声認識の枠組みで有効であることが確認されている．また，脳性麻痺者の音声認識においても，自己教師あり学習と疑似ラベリングを併用してラベル無し音声を活用した先行研究 [7] があ

* Speech recognition for persons with cerebral palsy utilizing wav2vec 2.0 and pseudo-labeling. by Yuki Matsuzaka (Kobe University), Ryoichi Takashima (Kobe University/JST PRESTO), Tetsuya Takiguchi (Kobe University)

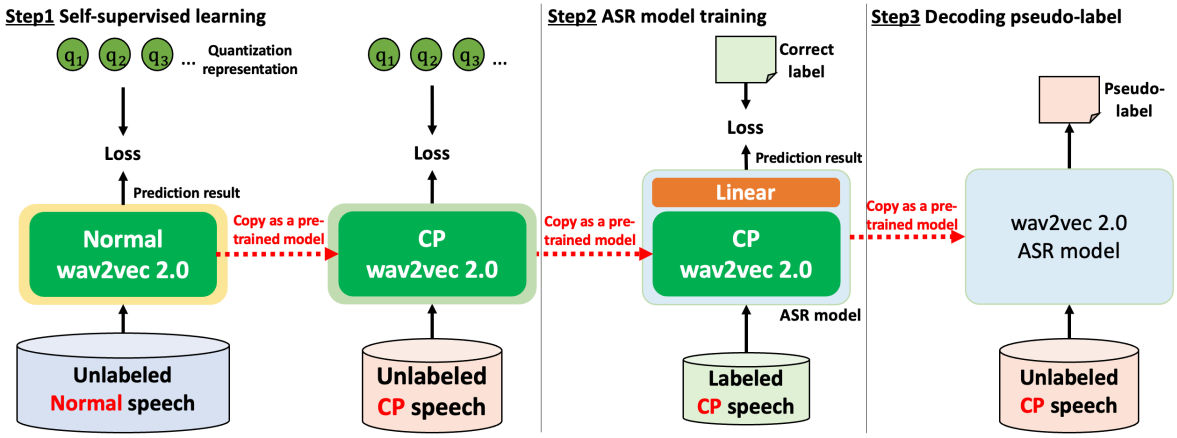

Fig. 1　Procedure for self-supervised learning and generation of pseudo-labels.

る．この研究では，自己教師あり学習のモデルとして Autoregressive Predictive Coding (APC) [8] を活用しており，自己教師あり学習と疑似ラベリングを併用することで認識性能が向上したことが報告されている．本研究でも同様に，wav2vec 2.0 の自己教師あり学習に加え，疑似ラベリングを併用することで，脳性麻痺者のラベル無し音声を活用することを検討する．

3　wav2vec 2.0 の自己教師あり学習と疑似ラベリングによるラベル無し音声の活用

3.1　自己教師あり学習と疑似ラベルの生成

Fig. 1 に，wav2vec 2.0 の自己教師あり学習を用いた ASR モデルの学習手順及び疑似ラベルの生成までの手順を示す．まず Step1 では，自己教師あり学習を行う．脳性麻痺者の音声の特徴表現を学習することが目的であるため，本来は脳性麻痺者のラベル無し音声のみで学習するべきだが，現状では wav2vec 2.0 の学習に十分な量を用意できないため，大量の健常者音声で事前学習を行なっておく．その後，事前学習モデルを初期モデルとして，脳性麻痺者のラベル無し音声を用いて自己教師あり学習におけるファインチューニングを行う．次に Step2 では，脳性麻痺者のラベル付き音声を用いて ASR モデルの学習を行う．このとき，ASR モデルの構造は wav2vec 2.0 のモデル構造を含んでいるため，その部分のパラメータの初期値を自己教師あり学習における学習済みモデルから取得し，その上でファインチューニングを行う．そして Step3 では，学習済みの ASR モデルを用いて，ラベル無し音声に対してデコード処理を実施する．これにより，ラベル無し音声に対応する疑似ラベルを生成する．

3.2　疑似ラベルを用いた ASR の学習

Fig. 2 に，疑似ラベルを用いた ASR モデルの学習手順を示す．Step4 では，ASR の学習を行うが，用いる学習データとして，脳性麻痺者のラベル付き音声に加え，ラベル無し音声及びそれに対応する疑似ラベルを用いる．またこのとき，ASR モデルにおける wav2vec 2.0 の部分を，Step1 で自己教師あり学習を行った学習済みモデルを初期値とする．次に Step5 では，Step4 で学習した ASR モデルを初期値として，ラベル付き音声のみでファインチューニングを行う．最後にラベル付き音声のみで学習している理由としては，ラベル無し音声に対応する疑似ラベルは誤りを含んでおり，最終的には誤りを含まないラベル付き音声のみで実施すべきと考えたためである．

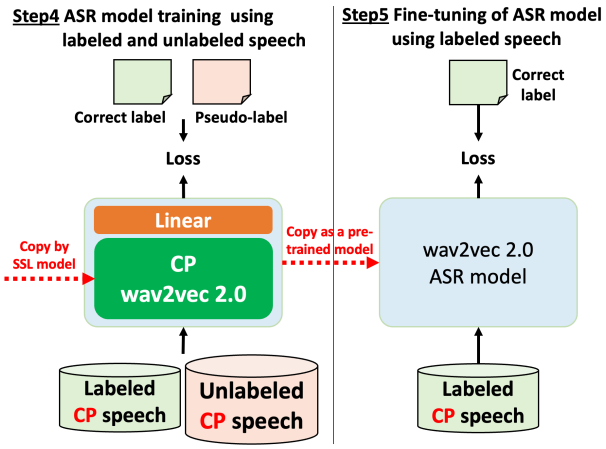

Fig. 2　Training procedure for ASR models with pseudo-labels.

4　評価実験

4.1　データ設定

本実験では，アテトーゼ型脳性麻痺者の男性 1 名を音声認識の対象話者とし，日本語音声を収録している．脳性麻痺者のラベル付き音声として，ATR 日本語音声データベース [9] に含まれる音素バランス文 503 文のうち，429 発話 (計約 50 分) 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．ラベル付き音声のデータ量は少量であるため，本研究では 8-fold の交差検証により評価を行う．脳性麻痺者のラベル無

し音声として，講演音声および新聞の読み上げ音声 (計約 3 時間) を収録している[1]．健常者音声としては，日本語話し言葉コーパス (CSJ) [10] を使用している (約 660 時間)．また，英語音声として Librispeech (約 960 時間) [11] 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．ASR モデルの学習の際には，話速変化のデータ拡張である Speed Perturbation [12] を実施しており，速度因数を 0.9, 1.0, 1.1 に設定した．

4.2 モデル設定

自己教師あり学習における wav2vec 2.0 のモデルは，文献 [5] を参考にして，Base モデルと同じ構造にした．CNN 特徴量エンコーダは 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⟩ からなる計 41 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 P2C テキストモデルによる文字認識への拡張

本研究における ASR モデルでは文字認識ではなく音素認識を実施する．その理由としては，脳性麻痺者のラベル付き音声は少量のため，文字認識のように数千以上の出力ノードを持つ ASR モデルの場合，学習できない文字が多く存在してしまうためである．しかし，実用的な音声認識を考慮すると，文字認識による評価も必要であると考えられる．よって本研究では，音素認識による評価だけでなく，文字認識でも評価を実施するために，音素 ASR モデルによって得られた音素認識結果を文字テキストに変換する Phoneme-to-Character (P2C) テキストモデルを使用する．このテキストモデルは学習に脳性麻痺者のデータを必要としない利点がある．

P2C テキストモデルの構造は，Transformer Encoder-Decoder モデルであり，Transformer エンコーダ 3 層，Transformer デコーダ 3 層で構成している．また，損失関数は Cross-Entropy Loss を使用している．学習データには CSJ のテキストデータを使用しており，モデル入力を音素テキスト，モデル出力を文字テキストとして学習を行った．

Table 1 Recognition error rates for a person with cerebral palsy (CP) using combination of self-supervised learning (SSL) in wav2vec 2.0 and pseudo-labeling (PL).

SSL (Normal)	SSL (CP)	PL	PER[%] (ASR)	CER[%] (P2C)
Librispeech			13.37	44.65
		✓	12.16	41.95
	✓		12.13	41.56
		✓	11.08	39.92
CSJ			9.97	34.08
		✓	8.88	32.25
	✓		8.73	31.76
		✓	**8.61**	**31.69**

4.4 実験結果

wav2vec 2.0 の自己教師あり学習と疑似ラベリングの併用により，ラベル無し音声を活用したときの脳性麻痺者の音声認識における実験結果を Table 1 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，さらに疑似ラベリングの有無による結果を比較している．評価指標は二つ用意しており，一つは ASR モデルによる音素認識結果を音素誤り率 (Phoneme Error Rate; PER) で評価したもの，二つ目は音素認識結果を P2C テキストモデルで文字テキストに変換し，文字誤り率 (Character Error Rate; CER) で評価したものである．

まず音素誤り率 (PER) を中心に結果を確認していく．実験結果より，健常者音声として英語音声である Librispeech より日本語音声である CSJ の方が認識性能が優れていることがわかる．これは評価話者が日本語音声を発話しているので，健常者音声も日本語音声が適しているためと考えられる．次に健常者音声として CSJ を用いた場合において，疑似ラベリングを行わずに (w/o PL)，脳性麻痺者のラベル無し音声による自己教師あり学習 (SSL(CP)) の有無を比較すると，実施することで認識性能が向上していることがわかる (9.97%→8.73%)．これより，ラベル無し音声を用いた自己教師あり学習の単体での効果を確認できる．また，脳性麻痺者のラベル無し音声による自己教師あり学習を行わずに (w/o SSL(CP))，疑似ラベリング (PL) の有無を比較すると，実施することで認識性能が向上していることがわかる (9.97%→8.88%)．これより，疑似ラベリング単体での効果も確認できる．

[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．

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

Table 2　PERs of pseudo-labels (PL).

SSL (Normal)	SSL (CP)	PER[%] (PL)
Librispeech		25.45
	✓	22.33
CSJ		20.04
	✓	**16.25**

そして，ラベル無し音声による自己教師あり学習と疑似ラベリングを併用することで，認識性能がより向上していることが確認できる (9.97%→8.61%)．これらの効果は，健常者音声として Librispeech を用いた場合でも同様に確認できる．

また，文字誤り率 (CER) で評価した場合，音素誤り率の場合と同様にラベル無し音声を活用することで，認識性能が向上することが確認でき，自己教師あり学習と疑似ラベリングを併用した場合が最も良い結果となった．これは，音素誤り率が小さいほど，正確な文字に変換しやすいと考えられ，今後の改良によって ASR モデル (音素認識) の認識性能を向上させることで，さらに文字誤り率を改善することが可能と思われる．

4.5　疑似ラベルの精度

Fig. 1 の Step3 で生成した疑似ラベルを用いることで，Fig. 2 の Step4 のように ASR モデルの学習においてラベル無し音声を活用することができた．ここでは，Step3 で生成した疑似ラベルの精度を確認する．

Table 2 に疑似ラベルの精度を音素誤り率 (PER) で評価した結果を示す．自己教師あり学習の学習手順や使用したデータによって比較している．結果より，自己教師あり学習に脳性麻痺者のラベル無し音声を活用した方が良い性能となっている．疑似ラベルは誤りを含めたラベルなので，ASR の学習に使用する上でラベルの誤りは少ないことが望ましいが，自己教師あり学習におけるラベル無し音声の活用によって，ラベルの誤りが改善していることがわかる．

5　おわりに

本研究では，脳性麻痺者の音声認識におけるデータ不足の問題を改善するために，脳性麻痺者のラベル無し音声を学習に活用することを検討した．我々の先行研究における wav2vec 2.0 による自己教師あり学習に加え，今回新たに疑似ラベリングによる方法を加え，2 つの方法を併用して脳性麻痺者のラベル無し音声を活用することで認識性能が向上することを確認した．また，P2C テキストモデルを用いて音素から文字に変換することで，文字認識による評価も実施した．今後の拡張としては，文字認識を行うための P2C モデルに関して，入力音素の誤りに頑健なモデルに改良することを検討する．

謝辞　本研究の一部は，JST さきがけ JPMJPR23I7 および 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] Y. Matsuzaka *et al.*, “Data augmentation for dysarthric speech recognition based on text-to-speech synthesis,” in *LifeTech*, pp. 399-400, 2022.

[3] 松坂勇樹 他, “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, pp. 1317-1320, 2022.

[4] Q. Xu *et al.*, “Iterative pseudo-labeling for speech recognition,” in *Interspeech*, pp. 1006-1010, 2020.

[5] A. Baevski *et al.*, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *NeurIPS*, vol. 33, pp.12449-12460 , 2020.

[6] Q. Xu *et al.*, “Self-training and pre-training are complementary for speech recognition,” in *ICASSP*, pp. 3030-3034, 2021.

[7] 澤佑哉 他, “擬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.

[8] Y.-A. Chung *et al.*, “An unsupervised autoregressive model for speech representation learning,” in *Interspeech*, pp. 146-150, 2019.

[9] 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.

[10] 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.

[11] V. Panayotov *et al.*, “Librispeech: An asr corpus based on public domain audio books,” in *ICASSP*, pp.5206-5210, 2015.

[12] T. Ko *et al.*, “Audio augmentation for speech recognition,” in *Interspeech*, pp.3586-3589, 2015.