複数種類の構音障害音声を用いたwav2vec音声認識モデル学習の検討*

☆大谷赳[1]，相原龍[2]，高島遼一[1]，滝口哲也[1]，△田口進也[2]
[1]神戸大学，[2]三菱電機

1 はじめに

構音障害とは，発話器官の損傷や運動障害などによって明瞭な発話ができなくなる障害である。構音障害者音声は健常者音声と特徴が大きく異なるため，健常者音声を前提とした音声認識システムでの認識が困難である。本研究では，脳性麻痺による運動障害性構音障害者，口唇口蓋裂および舌切除による器質性構音障害者を対象とした音声認識を目的とする。

構音障害者音声認識においては，障害者ユーザ本人の音声を用いて特定話者モデルを学習することが有効であるが，構音障害者から大量の学習データを収録することは困難である。学習データ不足の問題に対して，先行研究ではデータ増強 [1] やモデル適応 [2] などの手法が提案されている。

本研究では，利用可能な構音障害者音声そのものを増やすことを目的に，2種類の音声データの利用に着目する。一つは日常の自由発話のようなラベルの無い音声である。従来の台本読み上げ音声と比較して自由発話は収録の負担が小さいため，比較的多くの音声が収録可能と期待される。しかし，聞き取りが困難な構音障害者音声から教師ラベルを書き起こすことは困難である。そこで本研究では，自己教師あり学習の枠組みでラベル無し音声をモデル学習に利用する。

もう一つは認識ターゲットである障害者以外の構音障害者の音声である。本研究では，自己教師あり学習時に，認識対象である構音障害者の音声の他に，同じ種類の構音障害を持つ別話者の音声や，異なる種類の構音障害を持つ話者の音声も利用する。構音障害者音声の特徴は話者や障害の種類によって異なるが，一部の音素の欠落や音声全体の間延びといった特徴は，多くの構音障害音声に共通すると考えられる。そこで，構音障害者それぞれから収録可能な音声が少量であったとしても，複数人の障害者音声を用いることで，比較的多量の音声から構音障害音声の共通特長を学習でき，音声認識の性能向上に寄与できると期待される。本稿では脳性麻痺者，口唇口蓋裂者，舌摘出者を対象とした特定話者音声認識実験により，提案手法の有効性を示す。

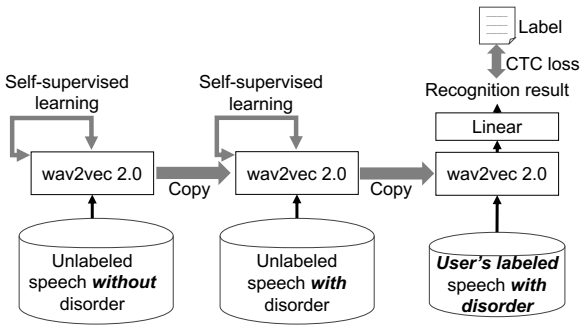

Fig. 1 The proposed model training procedure.

2 提案手法

本研究では，自己教師あり学習モデルとしてwav2vec 2.0 [3]（以降 wav2vec と呼ぶ）を用いる。構音障害者音声認識に対する wav2vec の有効性はいくつかの先行研究（例えば文献 [4]）で示されているが，これらの先行研究では健常者音声のみを用いて wav2vec を学習している。我々の先行研究 [5] では，健常者音声で学習した wav2vec を，さらに構音障害者のラベル無し音声を用いて追加学習して構音障害者の wav2vec を構築することで，認識性能の改善を確認している。先行研究 [5] では wav2vec の学習時に認識対象者の音声のみを用いていたが，本研究では認識対象者以外の構音障害者音声も用いる。

提案する学習方法を Fig. 1 に示す。まず健常者音声を用いて wav2vec を事前学習する。本稿では日本語話し言葉コーパス (CSJ) より 660 時間の健常者音声を用いて wav2vec の事前学習をしている。次に，構音障害者の音声を用いて wav2vec の自己教師あり学習の追加学習を行う。最後に，wav2vec の出力層に音声認識のための線形層を追加し，認識対象である構音障害者のラベル有り音声を用いてネットワーク全体を CTC 損失関数によりファインチューニングする。

3 実験

3.1 実験条件

本実験では 4 名の脳性麻痺者 (cerebral palsy; CP1〜4)，3 名の口唇口蓋裂者 (cleft lip and palate; CLP1〜3)，3 名の舌切除者 (tongue resection; TR1〜3) の計 10 名の構音障害話者音声を用いた。各話者は ATR

*Training of wav2vec model using speech with multiple types of speech disorder. by OTANI, Takeru[1], AIHARA, Ryo[2], TAKASHIMA, Ryoichi[1], TAKIGUCHI, Tetsuya[1], and TAGUCHI Shinya[2] ([1]Kobe University, [2]Mitsubishi Electric Corporation)

Table 1　PERs [%] for each speaker when wav2vec 2.0 was trained with and without unlabeled disordered speech.

	CP1	CP2	CP3	CLP1	CLP2	CLP3	TR1	TR2	TR3
w/o unlabeled disordered speech	10.2	16.4	13.4	6.0	5.7	4.9	11.6	13.2	7.4
w/ unlabeled disordered speech	8.7	13.1	9.7	4.6	4.8	4.3	8.4	10.5	6.4

日本語音声データベースの音素バランス文 503 文を読み上げ，429〜503 文の音声を収録した。これらの音声は wav2vec の自己教師あり学習および音声認識モデルの学習と評価に使用した。ただし，CP4 の音声については音声認識モデルの学習が収束しなかったため，自己教師あり学習にのみ用いて，評価は行わなかった。CP4 の評価については今後の課題とする。また，上記の音素バランス 503 文に加え，CP1 による新聞読み上げ音声と講演音声の計 1,460 文，CLP1 が JSUT コーパスの Basic5000 を読み上げた音声 4,973 文を収録し，自己教師あり学習に使用した。

音声認識モデルの出力は 39 種類の音素で定義し，音素単位での認識と評価を行った。健常者および複数構音障害者音声で学習した wav2vec に対して，それぞれの話者の音声で特定話者モデルを学習・評価した。このとき，音素バランス文読み上げ音声のうち 50 文を開発，50 文を評価，残りを学習に使用した。

3.2　実験結果

Table 1 に，wav2vec の学習時に健常者音声のみを用いた場合（w/o unlabeled disordered speech，つまり Fig. 1 における中央の処理をスキップした場合），および構音障害者音声も用いた場合（w/ unlabeled disordered speech）の音素誤り率（phoneme error rate; PER）を示す。表より，wav2vec の事前学習時に構音障害者音声を用いた場合のほうが，いずれの話者においても PER が低くなっていることがわかる。CP1 および CLP1 を除く話者は学習データ数が 503 文以下とわずかであるにも関わらず，全話者のデータと合わせて wav2vec の学習に用いることで，認識性能を改善できることが確認できた。

次に，CP1 の音声認識において，wav2vec の事前学習時に CP1 の音声のみを用いた場合と，全話者の音声を用いた場合で性能比較を行った。結果を Table 2 に示す。表より，CP1 の音声のみを用いた場合において PER が向上していることから，本人以外の構音障害者音声を用いることの有効性が確認できた。

Table 2　PERs [%] for CP1 when all speakers' speech was used for training wav2vec 2.0 (Proposed) and only CP1's speech was used.

	CP1
wav2vec trained with all speakers' speech	8.7
wav2vec trained with only CP1's speech	9.2

4　おわりに

本稿では，構音障害者音声認識の学習方法として，複数の構音障害者音声を用いて wav2vec を学習することを提案した。全ての評価話者について，構音障害者音声を用いて wav2vec を学習することの有効性を確認した。また，1 話者での実験において，評価話者の音声のみを使うよりも複数話者の音声を用いて wav2vec を学習したほうが性能が向上することを確認した。今後は，同じ種類の構音障害者のみを用いた場合と異なる種類の構音障害者音声を用いた場合での比較など，より詳細な検証を行っていく。

謝辞　本研究の一部は，JST さきがけ JPMJPR23I7 および JSPS 科研費 JP23K20733，JP22K12168 の支援を受けたものである。

参考文献

[1] Y. Jiao et al., "Simulating dysarthric speech for training data augmentation in clinical speech applications," *ICASSP*, pp. 6009–6013, 2018.

[2] R. Takashima et al., "Two-step acoustic model adaptation for dysarthric speech recognition," *ICASSP*, pp. 6104–6108, 2020.

[3] A. Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, pp. 12449–12460, 2020.

[4] M. K. Baskar et al., "Speaker adaptation for wav2vec2 based dysarthric ASR," *Interspeech*, pp. 3403–3407, 2022.

[5] 松坂 他, "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," 音講論（春）, pp. 873–876, 2024.