構音障害音声を用いたwav2vecモデル学習における障害種別についての比較*

☆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，発話したい内容を理解しているのにも関わらず，発話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。構音障害者音声は健常者音声と発話特徴が大きく異なることから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，脳性麻痺者，口唇口蓋裂者，舌切除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，障害者本人の音声を用いた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集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を増やすことを目的に，ラベルのない音声や認識対象者以外の構音障害者の音声を効果的に活用する方法について検討する。日常会話などの音声は，従来の台本読み上げ音声と比較して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，聞き取りが困難なため，教師ラベルを書き起こすことは困難である。そこで，自己教師あり学習の枠組みでこのようなラベルのない音声をモデル学習に利用する。構音障害者音声の特徴である，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，複数人の構音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の特徴を学習でき，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，舌切除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検証する。

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を用いた実験も行う。

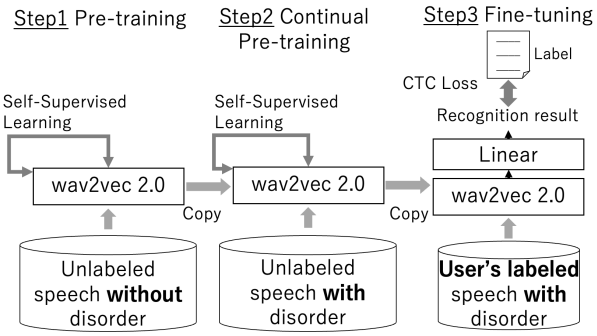

Fig. 1　The proposed model training procedure.

提案する学習方法を Fig. 1 に示す。

Step1 健常者による事前学習

まず健常者音声を用いて自己教師あり学習の枠組みにより wav2vec モデルの事前学習を実施する。本研究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。

Step2 構音障害者音声による自己教師あり学習

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。

Step3 音声認識タスクのための教師あり学習

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。

また，Step2 の wav2vec の追加学習において，以下の 4 種類の方法で構音障害者音声データを活用し，それぞれの認識性能を比較する。A. 構音障害者の音声を用いない (Step2 の処理をスキップする) 場合，B. 認識対象者本人のラベルなし音声のみを用いる場合，

*A comparative study of wav2vec model training on speech from different types of speech disorders.
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 method when wav2vec 2.0 was trained disordered speech in four different ways.

Unlabeled disordered speech	CP1	CP2	CLP1	CLP2	CLP3	TR1	TR2	TR3	TR4
A. Not used (Skipping step2)	9.5	16.3	5.6	5.1	4.9	10.4	11.9	6.7	5.2
B. unlabeled speech of the user	8.5	-	4.7	-	-	-	-	-	-
C. same type as the target user	9.7	16.3	5.1	5.6	6.1	11.2	13.4	7.4	5.4
D. multiple types	11.0	18.6	6.4	6.7	7.1	13.0	15.7	9.1	7.0

C. 認識対象者と同じ種別の障害を持つ構音障害者の音声データを用いる場合，D. C に加えて，認識対象者と異なる種別の障害も含めた複数種類の構音障害者音声データを用いる場合である。

3　実験

3.1　実験条件

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)，4 名の舌切除者 (tongue resection: TR1〜4) の計 11 名の構音障害話者音声を用いた。各話者は ATR 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。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 40 種類の音素で定義し，音素単位での認識と評価を行った。Step3 では，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全体の 10%を開発，10%を評価，残りを学習に使用した。

3.2　実験結果

Table 1 に，Step2 の wav2vec の追加学習の際，A. 構音障害者の音声を用いない (Step2 の処理をスキップする) 場合，B. 認識対象者本人のラベルなし音声のみを用いる場合，C. 認識対象者と同じ種別の障害を持つ構音障害者の音声データを用いる場合，D. C に加えて，認識対象者と異なる種別の障害も含めた複数種類の構音障害者音声データを用いる場合のそれぞれの音素誤り率（phoneme error rate: PER）を示す。表より，wav2vec の追加学習に認識対象者のラベルなし音声を用いると PER が低くなるが，認識対象者以外の構音障害者音声を用いた場合の PER は，CLP1 を除いたすべての話者で高くなることがわかる。CLP1 の実験に関しては，CLP2，CLP3 の音声が CLP1 の音声に比べて比較的少数であることから，CLP1 のみの音声を活用した場合と似た実験条件となり，PER が低くなった可能性がある。wav2vec2 の追加学習に本人の音声を用いることの有効性を確認することができたが，認識対象者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なかった。

4　おわりに

本稿では，構音障害者音声の活用方法の検証として，wav2vec の学習時に活用する構音障害者音声の種類を変えて実験を行い，性能比較を行った。対象者以外の構音障害者音声を活用することの有効性は確認できなかった。今後は，認識対象者以外の有効活用できなかった原因を検証する。また，有効活用することを目的とした，コードブックの改善など，モデルの改善などを行っていく。

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.