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*

☆澤佑哉, 冨士原健斗 (神戸大), 相原龍 (三菱電機),
高島遼一, 滝口哲也 (神戸大), △今井良枝 (三菱電機)

1 はじめに

構音障害とは，発話器官の障害や脳性麻痺などの運動機能障害によって正しい発音が困難となる症状である．本研究で対象としているアテトーゼ型脳性麻痺に起因する構音障害者は，意図した動作時に筋肉の不随意運動を伴い，この不随意運動が発話器官の筋肉に対して発生することで，正しく発音できないことがある．脳性麻痺患者は手足の動作が不自由であることが多く，手話や筆談といった音声コミュニケーションの代替手段が取れない場合が多いと考えられる．そのため，構音障害者の音声認識には高いニーズがあり，研究の必要性があると言える．

近年の音声認識技術の発展に伴って，構音障害者音声認識の分野でも様々な研究が行われている．構音障害者は発話時に身体への負担が大きく，十分な量の発話データを録音することが困難であることから，構音障害者音声認識においては利用できるデータ数が少ないという点が大きな問題となる．従来研究においても主にデータ量不足の問題に取り組んでおり，構音障害者音声を擬似的に生成するデータ拡張のアプローチ [1, 2] や，大量の健常者音声を用いて学習した不特定健常者音声認識モデルを少量の構音障害者音声を用いて再学習させるモデル適応のアプローチ [3, 4]，構音障害者音声の複数データベースを使用するアプローチ [5] などが提案されている．

本研究では，日常生活の場面等における自由発話音声を音声認識に活用することを検討する．自由発話音声の録音は，台本の読み上げによる収録と比較して構音障害者にとって身体への負担が小さいため，比較的容易に多くのデータを収集できると考えられる．しかし，構音障害者の発話スタイルは健常者とは異なり，人手により発話内容を認識し文字起こしを行うことは困難であるため，ラベルの無い音声データの活用方法が求められている．

ラベルの無い音声データを音声認識に活用するアプローチとしては，音声認識によりラベル無し音声にラベルを付与する擬似ラベリングの手法 [6, 7] や，ラベル無し音声のみで学習できるタスクにより特徴表現学習を行い，その後ラベル付き音声でファインチューニングを行う手法 [8, 9] などがある．これらの手法が健常者を対象にした音声認識において認識性能の向上に寄与している一方で，構音障害者音声認識においてはどの学習方法が効果的であるか，また各手法を効果的に組み合わせる事ができるかについては明らかになっていない．そこで本研究では，構音障害者音声認識において擬似ラベリングと特徴表現学習を使用する場合の音声認識性能の比較を行い，さらに両方の手法を併用することで音声認識性能を向上させることを試みる．

2 ラベル無し音声の音声認識への活用

2.1 擬似ラベリング

擬似ラベリングは，ラベルの無い音声を音声認識によりラベル付けを行うことで，音声認識に利用できる訓練データを擬似的に増やす手法である．文献 [6, 7] では，ラベル無し音声に擬似ラベルを付与した後，擬似ラベル付き音声を用いて音声認識モデルを再訓練することで，音声認識性能が向上することが報告されている．健常者音声認識において優れた効果を発揮する一方で，構音障害者音声認識においては擬似ラベルの精度が低いことから，音声認識モデルの訓練時に誤った音素列ラベルを与えてしまうという問題があると考えられる．

2.2 特徴表現学習

特徴表現学習は，目的のタスクに有効なデータの特徴表現を事前に擬似的なタスクを解くことにより獲得する手法であり，入力データに対して自動生成できる情報を教師ラベルとしてモデルの学習を行う．本研究では，特徴表現学習の手法として Autoregressive Predictive Coding (APC) [10] を使用する．APC モデルは Unidirectional Recurrent Neural Network (RNN) とその後の全結合層から構成され，RNN によって集約された現在までのフレーム情報から，将来のフレームを予測する．将来のフレームの予測は，音声フレームの局所的な範囲における類似性に頼らず，より大域的な範囲からフレーム予測を行うために，n ステップ先の予測フレームを推測する．モデルの学習は，入力系列 $\boldsymbol{x} = (x_1, x_2, ..., x_T)$ と予測出力系列 $\boldsymbol{y} = (y_1, y_2, ..., y_T)$ の間の，以下で表される L1 損失

*Dysarthric speech recognition using pseudo-labeling and feature learning. by Yuya Sawa, Kento Fujiwara (Kobe University), Ryo Aihara (Mitsubishi Electric Corporation), Ryoichi Takashima, Tetsuya Takiguchi (Kobe University), and Yoshie Imai (Mitsubishi Electric Corporation)

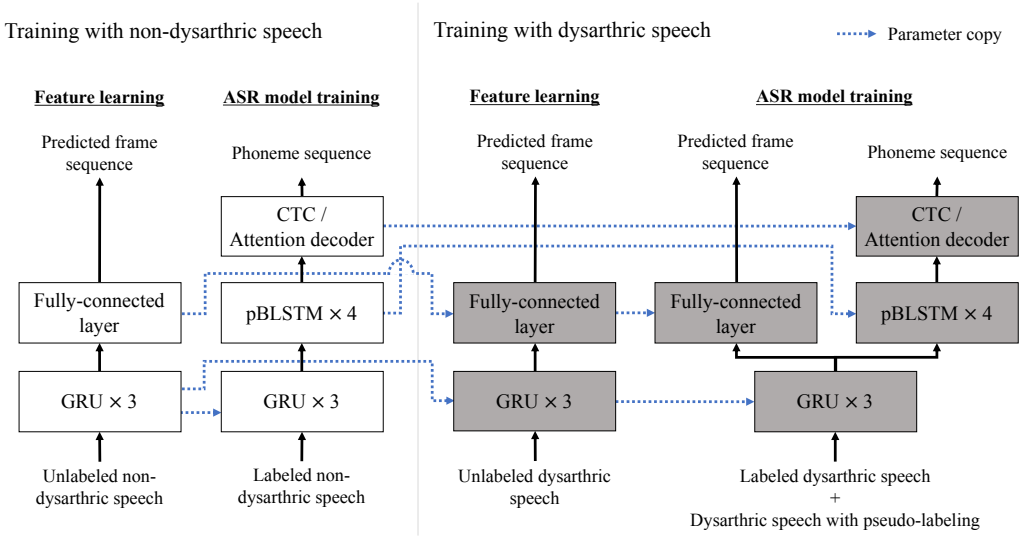

Fig. 1　Training procedure of the automatic speech recognition (ASR) model.

を最小化するように行われる．

$$L_{APC} = \sum_{i=1}^{T-n} |x_{i+n} - y_i| \tag{1}$$

3　特徴表現学習と擬似ラベリングの併用

本研究では，擬似ラベルの使用と特徴表現学習を同時に行うことで，音声認識性能の向上を試みる．Fig. 1 に提案手法の概要を示す．まず初めに，構音障害者のラベル無し自由発話音声を用いて，APC モデルによる特徴表現学習を行う．APC モデルは 3 層の Unidirectional Gated Recurrent Unit (GRU) と，その後に続く 1 層の全結合層から構成される．本研究では，構音障害者音声による APC モデルの学習時に，健常者音声で学習したモデルのパラメータを初期値としてファインチューニングをする，モデル適応のアプローチを取っている．ラベル無し自由発話は比較的多くのデータを収集することができるが，構音障害者の音声に関しては健常者と比べると依然少ない．構音障害者音声のみで APC モデルの学習を行った場合，モデルが有効な特徴表現を十分に獲得できない可能性があるため，モデル適応によりこれを回避する．Fig. 1 の破線の矢印は，学習されたモデルのパラメータをコピーすることを表す．

APC モデルを用いた特徴表現学習の後，音声認識モデルの学習を行う．音声認識モデルの学習にはラベル付きデータに加えて，特徴表現学習に使用したラベル無し音声に擬似ラベルを付与したデータも使用する．文献 [11] では，特徴表現学習されたモデルと音声認識モデルに対して，擬似ラベルを用いて学習を行う手法が提案されている．しかし，構音障害者音声の擬似ラベルの精度は健常者と比較して低く，多くの誤った正解ラベルが付与されるという問題がある．そこで提案手法では，音声認識モデルの学習時にラベルに依存しない特徴表現学習を同時に実施する事で，この問題を緩和することを試みる．具体的には，音声認識の損失関数 L_{ASR} と APC モデルの損失関数 L_{APC} の線形和である以下の式を最小化するように学習が行われる．

$$L = (1-\lambda)L_{ASR} + \lambda L_{APC} \tag{2}$$

本研究において L_{ASR} は，Connectionist Temporal Classification (CTC) と Attention 機構を用いた Encoder-Decoder モデルのマルチタスク学習損失関数 [12] を使用している．λ は L_{ASR} と L_{APC} の重みを決定するパラメータである．また音声認識タスクにおいてもデータ量不足の問題を緩和するため，健常者モデルからのモデル適応を行っている．

4　評価実験

4.1　実験条件

構音障害者の音声データは，アテトーゼ型脳性麻痺による構音障害を持つ日本人男性 1 名の収録音声を使用する．構音障害者のラベル付き音声は，ATR 日本語データベース [13] に含まれる音素バランス文 503 文のうち 429 文を読み上げたものである．自由発話音声には，構音障害者が大学で講演を行った際の収録音声と，新聞の文章を読み上げ発話の収録音声の合計 1,460 文を使用した．ATR 読み上げ音声は 50 文を評価データ，50 文を開発データ，残りを訓練データに分割し，自由発話音声は 76 文を評価データ，59 文を開発データ，残りを訓練データに分割した．音声認識実験における評価データは，ATR 読み上げ音声

Table 1　Experimental results in terms of PER [%].

Method	**Pseudo labeling**	**Feature learning**	**Multitask learning**	**PER** [%]
Baseline				**22.0**
Pseudo-labeling (PL)	✓			**19.0**
Feature learning (FL)		✓		**18.8**
PL + FL	✓	✓		**18.3**
PL + FL + Multitask	✓	✓	✓	**17.4**

と自由発話音声のそれぞれの評価データを合計した126文を使用した．モデルの事前学習に使用する健常者音声は，日本語話し言葉コーパス (CSJ) [14] に含まれる約660時間の音声を使用した．

入力音響特徴量として，80次元のメルフィルタバンク特徴を用いた．特徴表現学習は Autoregressive Predictive Coding を使用し，モデルは512次元の隠れ層を持つ3層からなる Unidirectional GRU と1層の全結合層で構成される．本実験では，予測先フレームを1に設定した．最適化には Adam を使用し，学習率は1e-4，エポック数は50とした．

音声認識は音素単位での認識を行い，出力音素次元数は音素39種類に未知文字<unk>・始端記号<sos>・終端記号<eos>を加えた42次元とした．音声認識モデルは，End-to-End 音声認識ツールキット ESPnet [15] を用いて，Hybrid CTC/attention モデル [12] の学習を行った．共有の Encoder は，320次元の隠れ層を持つ4層から成る Pyramid 型 Bidirectional Long-Short Term Memory (LSTM) とした．Attention 機構を用いた Decoder は320次元の隠れ層を持つ1層から成る Unidirectional LSTM と，その後の42次元のノードを持つ Softmax の出力層から構成される．CTC と Attention 機構付き Encoder-Decoder のマルチタスク学習では，CTC 損失関数の重みを0.5に設定し，認識時の CTC の出力確率の重みも同じく0.5とした．最適化には Adadelta を使用し，学習率は1e-8，エポック数は50とした．提案手法の音声認識と特徴表現学習のマルチタスク学習においては，APC モデルの損失関数重み λ を0.1に設定した．

4.2　実験結果

4.2.1　擬似ラベリングと特徴表現学習および提案手法の性能比較

Table 1 に，各実験における音素誤り率 (Phoneme Error Rate; PER) を示す．Baseline は特徴表現学習を行わずにラベル付きデータのみを用いて，3層の GRU，4層の pBLSTM，Hybrid CTC/attention モデルから成る音声認識モデルの学習を行った場合の結果である．Pseudo-labeling (PL) は，Baseline モデルを用いて擬似ラベルを付与したラベル無し音声を，音声認識の学習データに使用した際の結果である．Feature-learning (FL) は，ラベル無し音声を用いて特徴表現学習を行った後，特徴量抽出器 (3層の GRU) を音声認識に流用し，ラベル付き音声のみで音声認識モデルの学習を行った場合の結果である．PL+FL は FL の音声認識の学習データに擬似ラベル付き音声を混合した場合の結果であり，PL+FL+Multitask は提案手法である音声認識時に特徴表現学習とのマルチタスク学習を実施した場合の結果である．なお，Baseline モデルにより付与された擬似ラベルの音素誤り率は26.1 %であった．

実験の結果，擬似ラベリング使用の手法 (PL) と特徴表現学習を使用する手法 (FL) の両方で，Baseline からそれぞれ13.6 %，14.5 %の相対性能改善が確認された．加えて，それぞれのアプローチを同時に使用する手法 (PL+FL) により更に性能が向上し，Baseline から16.8 %の相対性能改善を達成した．擬似ラベリングにより利用可能なデータ数が増えたことと，特徴表現学習により音声認識に有効な特徴量表現が獲得できたことが，性能向上に寄与したと考えられる．

提案手法 (PL+FL+Multitask) である擬似ラベリングの使用と特徴表現学習を同時に実施することで，Baseline から20.9 %の相対性能改善となった．擬似ラベリングにより利用可能なデータ数が増える一方で，構音障害者音声認識においては健常者と比較して擬似ラベルの精度が低いという問題がある．精度が低いラベルを使用する際に，ラベルに依存しない特徴表現学習を組み込むことで，ラベルの低信頼性を補完するように機能していると考えられる．

4.2.2　使用するデータ量に関する比較

Table 2 に，音声認識に使用するデータ量に対する性能の変化を示す．訓練データ数には音声認識モデルの学習に使用したラベル付きデータとラベル無し

Table 2　The correlation between PERs [%] and the number of training data for ASR.

Number of utterances	labeled data	100	200	329	329	329	329	329
	unlabeled data	–	–	–	171	671	1,171	1,320
Method	Baseline	36.8	28.2	**22.0**	–	–	–	–
	Pseudo-labeling (PL)	–	–	–	21.2	20.6	19.8	**19.0**
	PL + FL	–	–	–	20.2	20.2	18.3	**18.3**
	PL + FL + Multitask	–	–	–	17.7	17.7	17.6	**17.4**

データの発話数の内訳を示しており，擬似ラベル使用時にはラベル付きデータ全 329 発話も併せて学習に使用している．なお，特徴表現学習時には全ての自由発話音声を使用して学習が行われている．提案手法である音声認識と特徴表現学習のマルチタスク学習を実施する場合は，データ数が少ない場合でも改善率が高い結果となった．このことから，提案手法は利用可能なデータ数が少ない時により有効な手法になると言える．

5　おわりに

本研究では，構音障害者音声認識において擬似ラベリングと特徴表現学習のアプローチを用いて，音声認識性能の向上を試みた．実験の結果，擬似ラベリングと特徴表現学習の各手法の使用により，従来の教師あり学習と比較して音声認識性能が向上することが確認された．加えて，本研究における提案手法である音声認識時の特徴表現学習とのマルチタスク学習の実施により，更なる性能向上を達成した．今後は，これらの学習方法をより効果的に統合できる手法を模索する．

参考文献

[1] B. Vachhani *et al.*, "Data augmentation using healthy speech for dysarthric speech recognition," in *Interspeech*, pp. 471–475, 2018.

[2] F. Xiong *et al.*, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *ICASSP*, pp. 5836–5840, 2019.

[3] J. Shor *et al.*, "Personalizing ASR for dysarthric and accented speech with limited data," in *Interspeech*, pp. 784–788, 2019.

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

[5] Y. Takashima *et al.*, "End-to-end dysarthric speech recognition using multiple databases," in *ICASSP*, pp. 6395–6399, 2019.

[6] Q. Xu *et al.*, "Iterative pseudo-labeling for speech recognition," in *Interspeech*, pp. 1006–1010, 2020.

[7] D. Park *et al.*, "Improved noisy student training for automatic speech recognition," in *Interspeech*, pp. 2817–2821, 2020.

[8] W. Wang *et al.*, "Unsupervised pre-training of bidirectional speech encoders via masked reconstruction," in *ICASSP*, pp. 6889–6893, 2020.

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

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

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

[12] S. Watanabe *et al.*, "Hybrid CTC/attention architecture for end-to-end speech recognition," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240–1253, 2017.

[13] 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.

[14] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pp. 7–12, 2003.

[15] S. Watanabe *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Interspeech*, pp. 2207–2211, 2018.