自己教師あり学習によるラベル無し自由発話を用い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), 本山信明 (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，Deep Neural Network (DNN) を用いた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．構音障害者音声認識においては，利用可能なデータ数が少ないという点が大きな問題となる．構音障害者は発話時に身体への負担が大きく，発話データを十分に収録することが難しい．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 Data Augmentation 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]，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．

これまでの研究で使用されてきた発話データは，構音障害者があらかじめ用意された台本の文章を読み上げ，その発話音声を録音したものである．このような収集方法は構音障害者にとって負担が大きいため，大量のデータを集めることが困難である．より多くの発話データを収集する方法としては，自由発話を収録するという手法がある．日常生活の場面等における自由発話を収録する方法は，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，データの収集が比較的容易であると考えられる．しかし，構音障害者の発話スタイルは健常者と異なることから，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り，ラベルの無い音声データの活用方法が求められている．

ラベルの無いデータをモデルの学習に用いるアプローチとして，自己教師あり学習が挙げられる．自己教師あり学習は，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ものであり，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，ラベルの無い構音障害者の自由発話を用いて自己教師あり学習を行い，学習したモデルを音声認識に流用することで音声認識精度の向上を行う．

2 Autoregressive Predictive Coding

本研究では，自己教師あり学習の手法として Autoregressive Predictive Coding (APC) [6] を使用する．Fig. 1 は，APC モデルの概略図を示している．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 損失を最小化するように行われる．

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

3 構音障害者音声の自己教師あり学習

Fig. 2 は，提案手法の概要を示している．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，将来フレームを予測する自己教師あり学習により APC モデルを学習する．本研究では，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．ラベル無しの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．そのため，構音障害者音声のみで自己教師あり学習を行った場合，データ量不足によりモデルが有効な特徴量表現を十分に獲得で

*Dysarthric speech recognition using unlabeled speech with self-supervised learning. by Yuya Sawa, Kento Fujiwara (Kobe University), Ryo Aihara (Mitsubishi Electric Corporation), Ryoichi Takashima, Tetsuya Takiguchi (Kobe University), and Nobuaki Motoyama (Mitsubishi Electric Corporation)

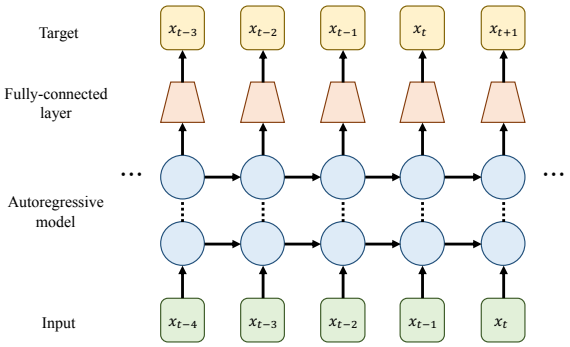

Fig. 1　Overview of autoregressive predictive coding ($n = 1$).

きない可能性がある．少量データからモデルを学習させる際のアプローチとして，既存の学習済みモデルに対して目的ドメインの少量データを用いて Fine tuning をする，モデル適応の手法が考えられる．そこで本研究では，健常者の不特定話者モデルから構音障害者の特定話者モデルを構築するモデル適応を行う．

自己教師あり学習後，構音障害者のラベル付き音声を用いて音声認識タスクを行う．本研究では，音声認識モデルは Connectionist Temporal Classification (CTC) と，Attention 機構を用いた Encoder-Decoder モデルを組み合わせた，Hybrid CTC/attention[7] モデルを使用する．共有 Encoder の前に学習した APC モデルの GRU 層部分を取り付け，特徴量抽出機として使用する．これにより，一般的な特徴量を用いた場合よりも良い特徴表現による学習が可能になると期待される．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，健常者によるモデルの事前学習を行う．特徴量抽出機部分である GRU 層は自己教師あり学習で学習したモデルのパラメータを初期値として使用し，学習時には GRU 層の重みが更新されるように設定した．

4　評価実験

4.1　実験条件

本実験において使用したデータセットの概要について，Table 1 に示す．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者男性 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 [8] 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．ラベル無し自由発話音声には，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，合計 2,185 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) [9] 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．自己教師あり学習においては，ラベル無しデータ 2,185 文のうち 215 文を開発データとし，残りを訓練データとして使用した．自己教師あり学習は 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．音声認識においては，ラベル付きデータ 429 文のうち 50 文を評価データ，50 文を開発データ，残りを訓練データとして使用した．音声認識モデルは，End-to-End 音声認識ツールキット ESPnet[10] を用いて，Hybrid CTC/attention モデルの学習を行った．共有の Encoder は，320 次元の隠れ層を持つ 4 層からなる 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 層からなる Unidirectional Long-Short Term Memory 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．

Table 1　Data set for self-supervised learning and speech recognition.

Data set	Existence of label	Number of utterances
Dysarthria	Yes	429
	No	2,185
CSJ	Yes	1,213,203

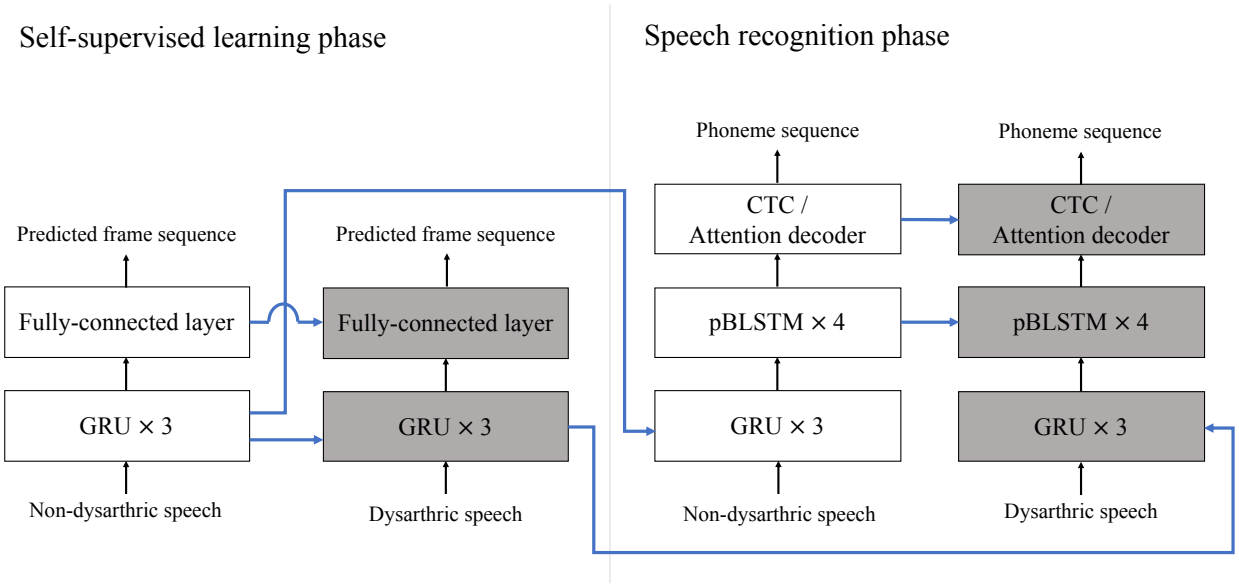

Fig. 2　System overview.

4.2　実験結果

4.2.1　自己教師あり学習の有無および事前学習の有無に関する性能比較

Table 2 に，各実験における音素誤り率 (Phoneme Error Rate; PER) を示す．Baseline は音声認識において自己教師あり学習による特徴量抽出器を使用しない場合，すなわち GRU 層を取り除いた場合の結果である．提案手法の自己教師あり学習を使用する場合では，音声認識モデルの学習時に特徴量抽出部分の GRU 層のパラメータ更新の有無 (有: unfreeze，無: freeze) に関する比較と，Hybrid CTC/attention における健常者音声での事前学習の有無に関する比較を行った．Baseline と比較すると提案手法では音素誤り率が低下しており，自己教師あり学習による特徴量表現が構音障害者音声認識においても有効であることが確認された．また，自己教師あり学習において健常者音声による事前学習を行った場合に認識精度が向上しており，事前に大量の健常者音声で自己教師あり学習を行うことがより有効な特徴量表現の獲得に寄与したと考えられる．さらに，Hybrid CTC/attention の事前学習を行った場合，ネットワーク全体が健常者音声で事前学習されることにより，音素誤り率が低下していることが確認できる．GRU 層のパラメータの固定に関する比較では，Hybrid CTC/attention をランダム初期値から学習する場合，パラメータを固定した方が良い結果を示した．パラメータを更新する方法が劣った理由として，ランダム初期値から学習することで音声認識モデルのパラメータの更新が大きくなり，その影響で事前学習された GRU 層のパラメータも大きく更新されたため，自己教師あり学習で得られた効果が弱まってしまったことが原因であると考えられる．一方で音声認識モデルが事前学習されている場合は，GRU 層のパラメータを更新した場合に性能の向上が見られた．この場合は音声認識モデルのパラメータの更新が比較的小さくなることから GRU 層のパラメータも大きく更新されず，自己教師あり学習で得られた特徴を音声認識時に保持できていると考えられる．

4.2.2　自己教師あり学習に使用するデータ量に関する性能比較

自己教師あり学習に使用する構音障害者のデータ量を変更して，音声認識における性能を比較した．Table 3 に，自己教師あり学習に使用するデータ数と音素誤り率の関係を示す．この実験では自己教師あり学習においては健常者音声による事前学習を行っており，音声認識モデルの訓練時には特徴量抽出部分である GRU の重み更新は行っていない．表より，学習データに含まれる発話数を増加させるにつれて性能改善の比率が小さくなり，およそ 500 文より発話数を増やした場合にはほぼ性能が変わらないことが確認できる．このことから，自己教師あり学習において比較的少ないデータ量でも性能向上が行えることが示唆される．

5　おわりに

本研究では，構音障害者のラベルの付いていない自由発話音声を用いて自己教師あり学習を行い，学習したモデルを音声認識タスクに使用することで，音声認識精度の向上を試みた．実験の結果，自己教師

Table 2　Experimental results in terms of PER [%].

	Pre-training CTC/attention	**Data set for self-supervised learning**		**GRU freezing**	**PER** [%]
		CSJ	Dysarthria		
Baseline		–	–	–	29.9
Using self-supervised learning		✓		freeze	26.3
		✓		unfreeze	25.7
			✓	freeze	24.3
			✓	unfreeze	25.6
		✓	✓	freeze	22.8
		✓	✓	unfreeze	25.3
Baseline	✓	–	–	–	16.6
Using self-supervised learning	✓	✓		freeze	20.4
	✓	✓		unfreeze	14.7
	✓		✓	freeze	21.2
	✓		✓	unfreeze	17.5
	✓	✓	✓	freeze	19.0
	✓	✓	✓	unfreeze	13.7

Table 3　The correlation between PERs [%] and the number of training data for self-supervised learning.

Usage rate of training data	**Number of utterances**	**PER** [%]
0.125	247	25.0
0.25	493	23.6
0.5	986	23.5
0.75	1,477	23.1
1.0	1,970	22.8

あり学習で獲得した特徴量表現を音声認識タスクに使用することで認識精度が向上し，加えて自己教師あり学習において大量の健常者音声を用いた事前学習が有効であることを確認した．今後は他の自己教師あり学習の手法との比較を行い，構音障害者にとってより有効な自己教師あり学習の手法を模索する．

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

[7] 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.

[8] 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.

[9] 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.

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