構音障害者音声認識における自己教師あり学習と疑似ラベリングの動的重み付きマルチタスク学習*

☆澤佑哉 (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いため，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．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，構音障害者音声を疑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．

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．

ラベルの無い音声データを音声認識に活用するアプローチとして，疑似ラベリングや自己教師あり学習を用いた手法が提案されている．疑似ラベリングは，ラベルの無い音声に対して音声認識によりラベルを付与することで，訓練データとして利用可能にする手法であり，健常者音声認識において有効性が確認されている [6, 7]．しかし構音障害者音声認識においては疑似ラベルの精度が低いため，多くの誤った正解ラベルを付与した音声データで音声認識モデルを学習することの悪影響が懸念される．一方，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疑似的なタスクを解くことにより獲得する手法であり，近年盛んに研究が行われている [8, 9]．自己教師あり学習では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ため，ラベル情報に依存しない学習が可能である．

我々の以前の研究 [10] では，構音障害者音声認識において疑似ラベルの誤りによる悪影響を緩和するため，音声認識時にラベルに非依存なタスクである自己教師あり学習とのマルチタスク学習を実施する手法を検討し，提案手法による性能改善を確認した．文献 [10] では，マルチタスク学習における各タスクの重みパラメータについて，入力された発話データに関わらず常に一定の値を付与していた．本研究では，入力された発話データに基づいてマルチタスク学習の重みを動的に制御することで，提案手法の更なる改善を試みる．

2 自己教師あり学習と音声認識のマルチタスク学習

本章では，以前に我々が提案した手法 [10] について説明する．本研究では，自己教師あり学習の手法として Autoregressive Predictive Coding (APC) [11] 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

*Multi-task learning with dynamic weights of self-supervised learning and pseudo-labeling for dysarthric speech recognition. by Yuya Sawa, (Kobe University), Ryo Aihara (Mitsubishi Electric Corporation), Ryoichi Takashima, Tetsuya Takiguchi (Kobe University), and Yoshie Imai (Mitsubishi Electric Corporation)

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，以下の式 (1) で表される L1 損失を最小化するように行われる．

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

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 APC 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．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．

APC モデルを用いた自己教師あり学習の後，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，自己教師あり学習に使用したラベル無し音声に疑似ラベルを付与したデータも使用する．文献 [12] では，自己教師あり学習されたモデルと音声認識モデルに対して，疑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，構音障害者音声においては疑似ラベルの精度が健常者と比較して低く，誤った正解ラベルが多く付与されるという問題がある．そこで，音声認識モデルの学習時にラベルに依存しない自己教師あり学習を同時に実施する事で，この問題を緩和する．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 (2) 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 [13] 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．

3 疑似ラベルの信頼度に基づいたマルチタスク学習重みの動的変更

以前の手法 [10] では，全ての学習データに対して固定のパラメータ λ でマルチタスク学習を行っていた．しかし，正解ラベルを持つ発話や疑似ラベルの誤りが少ない発話に対しては，音声認識の損失を重視した方が良いと考えられる．そこで本研究では，疑似ラベルデータが入力された際に，疑似ラベルの誤りの多少に基づいてマルチタスク学習重み λ を決定する手法を検討する．疑似ラベルの誤りが多い場合は λ の値を大きく設定し，反対に誤りが少ない場合は λ の値を小さく設定する．本論文では，疑似ラベルの誤りの多少を表す指標を信頼度 (Confidence Score) と呼ぶ．

本研究では，疑似ラベル生成時の CTC の出力確率分布を用いて信頼度の推定を行い，各フレームにおける最大の出力確率の平均を発話全体の信頼度とする．音響特徴量のフレーム数を T，時刻 t におけるトークン i の確率を y_i^t とすると，疑似ラベルの信頼度 CS は以下の式 (3) で表される．

$$CS = \frac{1}{T} \sum_{t=1}^{T} \max_i y_i^t \tag{3}$$

ただし，最大の出力確率が空白記号 ⟨blank⟩ となるフレームに関しては平均の計算から除外し，文字記号を出力した場合のみ信頼度の計算に含める．

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 [14] 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) [15] 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．

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

Method	Multi-task weight λ									
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Baseline	18.05	–	–	–	–	–	–	–	–	–
MTL	–	17.83	17.73	17.57	17.52	17.42	17.44	17.56	17.72	18.38
MTL(dynamic)	–	17.75	17.50	17.48	17.23	**17.19**	17.22	17.27	17.31	17.39
MTL(dynamic) + CS	–	17.53	17.40	17.24	17.18	**17.08**	17.19	17.34	17.39	17.39

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 [16] を用いて，Hybrid CTC/attention モデル [13] 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．

マルチタスク学習重みの制御については，以下の 3 つの条件を検証した．(dynamic) の表記がある手法は，マルチタスク学習重み λ の動的変更を行うことを意味する．

- (**MTL**)：正解ラベルデータと疑似ラベルデータの両方に対して，マルチタスク学習を行う．マルチタスク学習重み λ は常に一定の値が付与される．
- (**MTL(dynamic)**)：疑似ラベルデータに対してのみマルチタスク学習を行い，正解ラベルデータに対してはマルチタスク学習を行わない (λ=0 とする)．
- (**MTL(dynamic) + CS**)：ある閾値以下の信頼度の疑似ラベルデータに対してのみマルチタスク学習を行い，それ以外のデータに対してはマルチタスク学習を行わない (λ=0 とする)．

MTL(dynamic) + CS における信頼度の閾値は，0.9 に設定した．

4.2　実験結果

4.2.1　マルチタスク学習重みの動的変更の有効性に関する比較

Table 1 は，提案手法における音素誤り率 (Phoneme Error Rate; PER) を表している．Baseline は，音声認識時に自己教師あり学習とのマルチタスク学習を行わない場合，つまりマルチタスク学習重み λ の値を常に 0.0 に設定した場合の結果である．実験の結果，音声認識時に自己教師あり学習とのマルチタスク学習を導入することで認識性能が向上した．加えて，マルチタスク学習重みの動的変更を組み込むことで更に性能が向上し，Baseline と比較して最大で 4.7%の相対性能改善を達成した．また，提案手法である疑似ラベルの信頼度に基づいてマルチタスク学習重みを動的に変更することで，Baseline と比較して最大で 5.4%の相対性能改善を達成した．提案手法では推定された疑似ラベルの信頼度を参照し，不正確な疑似ラベルに対してのみマルチタスク学習が行われたことで，より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　信頼度と音素誤り率の関係

Fig. 1 は，発話単位の信頼度と疑似ラベルデータの音素誤り率の関係を表しており，発話の信頼度と疑似ラベルの音素誤り率の間には負の相関が認められた (相関係数: −0.561)．信頼度の高い発話データほど疑似ラベル中の誤りが少ない傾向があり，信頼度が疑似ラベルの正確性を表すことができていると言える．

一方で Fig. 2 は，音素単位の信頼度と誤り率の関係を表している．音素誤り率は，削除または置換のいずれかの誤りをした割合を示している．発話単位の場合と同様に，音素単位の場合も信頼度と誤り率の間に負の相関が認められた (相関係数: −0.688)．また誤りの傾向を見ると，本実験の話者に関しては破裂音や摩擦音の認識が特に困難であることが分かる．

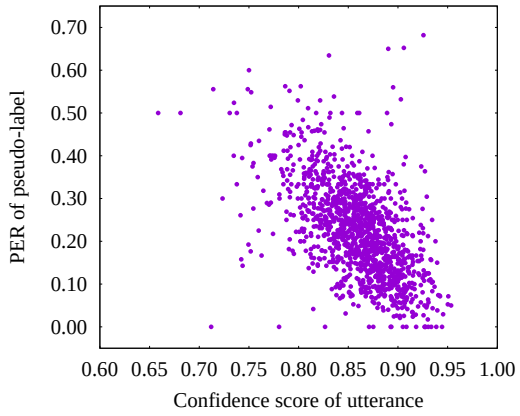

Fig. 1 The correlation between confidence score and PER of pseudo-label for each utterance.

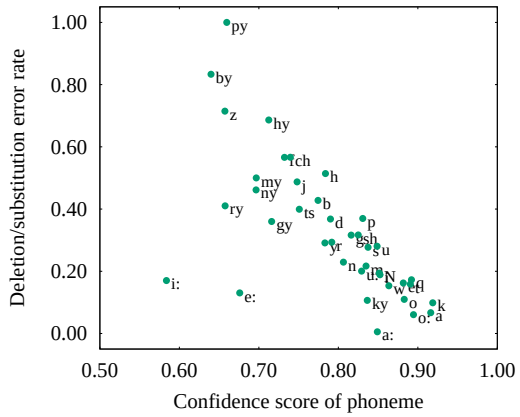

Fig. 2 The correlation between average confidence score and average deletion/substitution error rates per phoneme.

5 おわりに

本研究では，構音障害者音声認識における自己教師あり学習とのマルチタスク学習による性能向上を試みた．疑似ラベルの生成時に音声認識モデルの出力に基づき信頼度を推定し，信頼度に基づいてマルチタスク学習重みを制御する手法を検討した．実験の結果，推定された信頼度に基づきマルチタスク学習重みを動的に変更することで，音声認識性能が向上することが分かった．また，信頼度が疑似ラベルの音素誤り率を反映していることを確認した．今後は，より正確な疑似ラベルの精度を表す信頼度の推定方法を模索する．

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] 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 2021 年秋季研究発表会, pp. 847–850, 2021.

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

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

[13] 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.

[14] 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.

[15] 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.

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