構音障害者を対象とした日本語大語彙連続音声認識の検討*

◎高島遼一, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

深層学習技術の発展により，音声認識技術の精度は人間と同等レベルと報告されるほどに大きく向上している [1]。音声認識技術は既にスマートスピーカや音声翻訳，車載デバイスなどに応用されている。音声認識はハンズフリーでの入力が可能となるため，障害者のための入力インターフェースとしての応用が期待される。

本研究で扱うアテトーゼ型脳性麻痺は，脳の障害により，筋肉の不随意運動が発生し，身体の動作が不自由になる障害の一種である。この障害を持つ患者には手話や筆記を困難とする方も多いため，ハンズフリーによるコミュニケーションツールが望まれており，音声認識技術への期待が寄せられている。しかしながら不随意運動は顔や舌の筋肉に影響することもあるため，発話障害も併発している脳性麻痺患者も多く存在する。文献 [2] の調査では，脳性麻痺の障害を持つ児童のうち半数以上が発話障害も伴っていることが報告されている。アテトーゼ型脳性麻痺に起因する発話障害者の音声は不安定になりやすく，従来の健常者で学習した音声認識モデルでは認識が困難である。

本研究では，アテトーゼ型脳性麻痺による発話障害者の特定話者音声認識モデルを作成することを目的とする。発話障害者の学習データを十分量収録することは困難なため，少量データから音声認識モデルを学習する必要がある。少量データによるモデル学習のアプローチとして，大きく 2 種類存在する。一つは Data augmentation と呼ばれるもので，オリジナルの学習データに対して，例えばノイズを加えたり時間方向に伸縮させたりすることで学習データを増やすというアプローチである [3]。Data augmentation を用いることで構音障害者音声認識精度を向上させる研究が既にされている [4]。もう一つはモデル適応と呼ばれるもので，既存のモデルに対して，目的のドメインで得られた少量データ（適応データ）を使って fine-tuning することで，目的のドメインに適応するというアプローチである。本研究では，既存のモデルが再利用可能となるモデル適応のアプローチを用いて，既存の不特定健常者モデルから特定構音障害者モデルを構築する。

これまでモデル適応の研究が多くされているが，基本的なアプローチは，既存のモデルのパラメータを，適応データを用いて更新することである [5, 6]。ただし少量のデータを用いてパラメータの更新を行うと over-fitting が発生しやすいため，これを防ぐために様々な制約や正則化が使われている。ここでの制約は，適応元のドメインと適応先のドメインとの間で，入力の分布が大きく離れていないという仮説に基づいている。しかしながら，健常者と構音障害者の発話スタイルが大きく異なる場合，この仮説は必ずしも正しくなく，我々の実験においても，隠れ層の学習率を小さくする制約 [7] や適応層を加える制約手法 [8] では性能改善が見られなかった。

本稿では，制約を加えずに特定構音障害者モデルを学習するための，2 段階モデル適応手法を提案する。1 段階目では，複数の構音障害者音声を用いて，不特定健常者モデルを不特定構音障害者音声へ適応する。2 段階目では，適応させた不特定構音障害者モデルを，特定の構音障害者音声へさらに適応する。構音障害者の発話スタイルは，各障害者の症状によって様々ではあるが，例えば子音が不明瞭になる点や，各音素の継続長が長くなりやすい点など，障害者で共通する特徴も存在する。本手法を用いることで，1 段階目の適応では，構音障害者共通の特徴へ適応し，2 段階目の適応で話者固有の特徴へ適応が可能になると期待される。大語彙連続音声認識の実験により，提案手法の有効性を確認する。

2 手法

2.1 ベースモデル適応手法

本研究では，音響モデルとして Lattice-free maximum mutual information (LF-MMI) モデル [9] を使用する。LF-MMI モデルの適応手法はいくつか提案されているが，本研究では実装が簡単な転移学習のアプローチ [7] を使用した。本手法では，まず大量の健常者音声を用いて不特定健常者モデルを学習しておく。そして不特定健常者モデルのパラメータを初期値として，構音障害者音声を用いてパラメータを更新することで，構音障害者モデルを学習する。一般に，over-fitting を防ぐため，出力層を除く各層 (転移層) のパラメータは固定するか，出力層よりも小さい学習率でパラメータ更新を行うが，本研究の実験では，転移層の学習率を小さくしても音声認識率の改善が

*An investigation of Japanese large-vocabulary continuous speech recognition for dysarthric speakers. by Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki (Kobe University)

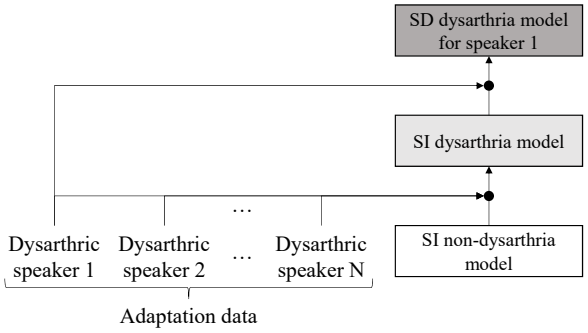

Fig. 1　Two-step model adaptation

見られなかったため，出力層と同じ学習率で更新を行うことにした。また，転移層を固定し，出力層と転移層の間に適応層を加える linear hidden network [8] も評価したが，性能は改善されなかった。

2.2　2 段階モデル適応

構音障害者と健常者の発話スタイルは大きく異なるため，不特定健常者モデルを直接特定障害者へ適応する従来手法では，十分な適応が行われないと考えられる。そこで提案手法では，一旦不特定構音障害者へ適応してから特定構音障害者へ適応する 2 段階のモデル適応を提案する。図 1 に提案するモデル適応手法を示す。1 段階目では，目的の構音障害者も含む複数の構音障害者の音声を用いて，不特定健常者モデルから不特定構音障害者モデルへ適応を行う。2 段階目では，目的の構音障害者の音声を用いて，適応した不特定構音障害者モデルを特定構音障害者モデルへ，さらに適応を行う。本手法の狙いは，1 段階目で子音の不明瞭化といった構音障害者共通の特徴へと適応し，健常者モデルを構音障害者音声の分布へ近づけておくことで，2 段階目の特定構音障害者への適応をより正確に行えるようにすることである。

3　実験

3.1　実験条件

3.1.1　音声認識モデル

提案手法の有効性を確認するため，構音障害者の日本語大語彙連続音声認識の実験を実施した。モデル学習および評価は Kaldi ツールキットを用いて行った。ベースラインの不特定健常者音響モデルおよび言語モデルは，日本語話し言葉コーパス (CSJ) から約 240 時間の音声およびテキストを用いて，CSJ 向け Kaldi レシピに基づいて学習した。

音響モデルのネットワーク構造を表 1 に示す。入力として，40 次元の MFCC を前後 1 フレーム分結合し，さらに 100 次元の iVector を加えて線形判別分析

Table 1　Acoustic model structure. TDNN ($-i$, 0, $+j$) means a TDNN layer splicing inputs of $(t-i)$-th, t-th, and $(t+j)$-th frames, where t denotes the index of the current frame.

ID	Layer
1	Fully-connected + ReLU
2	TDNN (-1, 0, +1) + ReLU
3	Fully-connected + ReLU
4	TDNN (-1, 0, +1) + ReLU
5	Fully-connected + ReLU
6	TDNN (-3, 0, +3) + ReLU
7	TDNN (-3, 0, +3) + ReLU
8	TDNN (-3, 0, +3) + ReLU
9	TDNN (-3, 0, +3) + ReLU
10	Fully-connected + ReLU
11	Output

を適用した 220 次元の特徴量を用いた。音響モデルは全結合層と time-delay neural network (TDNN) 層からなる。各隠れ層のノード数は 625，活性化関数は ReLU とし，batch normalization を適用した。出力は約 3,900 個の bi-phone で定義した。音響モデルは LF-MMI 基準を用いて学習し，初期学習率は 0.001，クロスエントロピーによる正規化重みは 0.1，エポック数は 4 である。

言語モデルとして，tri-gram モデルを CSJ のテキストを用いて学習した。語彙サイズは約 71,800 単語である。

3.1.2　適応および評価データ

適応および評価データとして，4 名のアテトーゼ型脳性麻痺による構音障害者の音声を収録した。構音障害者の音声は，ATR 日本語データベースのテキスト 503 文を読み上げたものである。ただし，症状による理由から，503 文全てを読み上げていない被験者も存在する。表 2 に適応および評価データの詳細を示す。各構音障害者の収録音声 (DYS-SPK1–4) のうち，200 文を適応データに，残りを評価データとして用いた。健常者音声との性能比較を行うため，ATR データベースに収録された健常者音声 (ATR-FKN, FYM, MSH, MTK) も評価した。さらに，参考として，CSJ の評価セット (CSJ-eval1–3) も評価した。

表 3 に各評価セットの未知語 (out-of-vocabulary; OOV) 率を示す。本実験では CSJ を用いて辞書を作成しているため，ATR の評価セットおよび ATR のテキストを読み上げた構音障害者の評価セットは OOV 率が高い。各評価セットにおいて OOV 率が音声認識の単語誤り率 (word error rate; WER) の下限になることに注意されたい。

Table 2　Adaptation and evaluation data.

Set	Dysarthria	Script	#Utts for adaptation	#Utts for evaluation
CSJ-eval1	No	CSJ	0	1,272
CSJ-eval2				1,292
CSJ-eval3				1,385
ATR-FKN		ATR		303
ATR-FYM				303
ATR-MSH				303
ATR-MTK				303
DYS-SPK1	Yes		200	229
DYS-SPK2			200	300
DYS-SPK3			200	303
DYS-SPK4			200	301

Table 3　OOV rates of evaluation datasets.

Set	OOV rates [%]
CSJ-eval1	4.87
CSJ-eval2	5.02
CSJ-eval3	7.61
ATR-FKN, FYM, MSH, MTK	14.85
DYS-SPK1	15.06
DYS-SPK2	14.51
DYS-SPK3	14.85
DYS-SPK4	14.23

3.2　実験結果

3.2.1　不特定健常者モデルの評価結果

まずベースラインである不特定健常者モデルの性能を評価した。表 4 に健常者音声に対する WER を示す。CSJ の評価セットに対して ATR の評価セットは WER が高い。この理由として，まず 3.1.2 節で述べた通り，ATR 評価セットは CSJ 評価セットに比べて OOV 率が高い点が挙げられる。また，CSJ は講演音声が主に収録されているのに対して ATR はバランス音素文が収録されているため，言語モデルのドメインのミスマッチも WER が高い理由として考えられる。

次に，不特定健常者モデルを用いて構音障害者音声を認識した結果を表 5 に示す。構音障害者音声に対する WER は健常者音声と比べて大きく増加しており，このことから，構音障害者と健常者の発話スタイルが大きく異なることが示唆される。

3.2.2　構音障害者モデルの評価結果

本節では不特定構音障害者モデルおよび特定構音障害者モデルの評価結果について述べる。図 2 に評価した適応方法を示す。構音障害者モデルは，別のモデルからの適応か，あるいはスクラッチから学習し，

Table 4　WERs [%] of non-dysarthric evaluation sets on the baseline SI non-dysarthria model.

Set	WER [%]
CSJ-eval1	10.79
CSJ-eval2	8.86
CSJ-eval3	10.55
ATR-FKN	28.00
ATR-FYM	23.67
ATR-MSH	23.39
ATR-MTK	27.21

Table 5　WERs [%] of dysarthric evaluation sets on the baseline SI non-dysarthria model.

Set	WER [%]
DYS-SPK1	82.28
DYS-SPK2	132.50
DYS-SPK3	116.48
DYS-SPK4	95.88

評価した。スクラッチから学習する際はエポック数を 4 に，適応を行う際はエポック数を 2 とした。その他の学習条件は 3.1 節で述べたものと同様である。

不特定構音障害者モデルの構築について，不特定健常者モデルから適応する方法と，スクラッチから学習する方法の二通りを比較した。表 6 に，二通りの学習方法における評価結果を示す。表 5 と比較して，学習方法によらず構音障害者モデルを構築することで，WER が大幅に改善することが分かる。さらに，スクラッチから学習するよりも，不特定健常者モデルからの適応により学習した方が WER が改善されることが分かる。このことから，構音障害者と健常者では発話スタイルは大きく異なるとはいえ，健常者音声から学習した知識は構音障害者の学習データが少ない場合において有効であることが示唆される。

次に，特定構音障害者モデルの構築については，ス

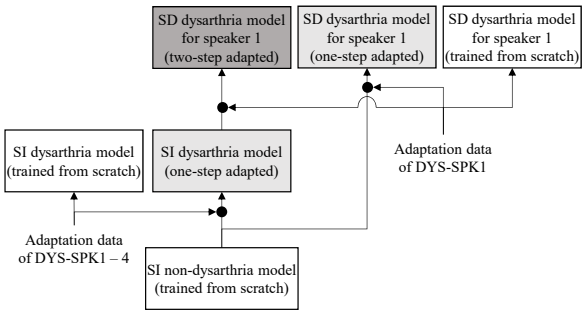

Fig. 2　SI/SD dysarthria model training approaches.

Table 6　WERs [%] of two approaches to train the SI dysarthria model.

Method	w/o adaptation	1-step adaptation
Source model	Scratch	SI non-dysarthria model
DYS-SPK1	44.83	43.29
DYS-SPK2	59.85	59.07
DYS-SPK3	70.25	67.74
DYS-SPK4	62.08	59.19

Table 7　WERs [%] of three approaches to train the SD dysarthria model.

Method	w/o adaptation	1-step adaptation	2-step adaptation
Source model	Scratch	SI non-dysarthria model	SI dysarthria model
DYS-SPK1	52.35	40.20	36.41
DYS-SPK2	73.83	65.21	54.66
DYS-SPK3	80.28	75.12	67.30
DYS-SPK4	67.58	59.98	56.28

クラッチから学習する方法，不特定健常者モデルから適応する方法 (従来の 1 段階適応)，そして適応された不特定構音障害者モデルをさらに適応する方法 (提案の 2 段階適応) の三通りを比較した。表 7 に各方法での評価結果を示す。表 5 および表 6 と比較して，特定構音障害者モデルをスクラッチから学習した場合は，不特定健常者モデルよりは WER が低く，不特定構音障害者モデルより WER が高い結果となった。このことから，200 発話の特定構音障害者の音声が，約 240 時間 (約 159,000 発話) の健常者音声よりも有用であることが示唆される。

特定構音障害者モデルを不特定健常者モデルからの適応により学習した場合，つまり従来の 1 段階適応を行った場合では，スクラッチから学習した場合よりも低い WER を示した。また，表 6 の不特定構音障害者モデルの結果と比較して，DYS-SPK1 で低い WER を，DYS-SPK2，DYS-SPK3 で高い WER を，DYS-SPK4 で同程度の WER を示した。表 5 において，DYS-SPK1 の音声は DYS-SPK2 および DYS-SPK3 の音声よりも WER が低いことから，DYS-SPK1 の音声は DYS-SPK2 および DYS-SPK3 の音声に比べて分布が健常者音声の分布に近いと考えられる。従って，不特定健常者モデルから適応により学習する手法は，健常者音声と分布が近い構音障害者音声に対して特に有効であると推察される。

特定構音障害者モデルを不特定構音障害者モデルからの適応により学習した場合，つまり提案する 2 段階適応を行った場合，比較した 3 手法の中で最も低い WER を示し，かつ不特定構音障害者モデルよりも低い WER を示した。このことから，提案手法では，ベースラインの不特定健常者モデルからは構音障害の有無によらない共通の知識を，さらに不特定障害者モデルからは構音障害者に共通の知識を，2 段階の適応処理により特定構音障害者モデルへ転移することができ，それにより目的構音障害者音声の学習データが少量の場合において有効に働いたと考えられる。

4　おわりに

本稿では，既存の不特定健常者音響モデルを利用してモデル適応により特定構音障害者モデルを構築する手法について検討した。構音障害者音声と健常者音声の発話スタイルの差が大きいことから，従来の 1 段階のモデル適応では十分に目的の構音障害者音声に適応しきれないという課題に対して，不特定構音障害者モデルへの適応を介して特定構音障害者モデルへ適応するという 2 段階のモデル適応を提案し，1 段階のモデル適応よりも WER が低くなることを確認した。今後は，提案手法と，既に提案されている正則化手法との組み合わせについて検討していく。

参考文献

[1] W. Xiong, et al., “The microsoft 2017 conversational speech recognition system,” in *ICASSP*, pp. 5934–5938, 2018.

[2] A. Nordberg, et al., “Speech problems affect more than one in two children with cerebral palsy: Swedish population-based study,” *Acta Paediatrica*, vol. 102, pp. 161–166, 2013.

[3] T. Ko, et al., “Audio augmentation for speech recognition.,” in *Interspeech*, pp. 3586–3589, 2015.

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

[5] S. Mirsamadi and J. H. L. Hansen, “A study on deep neural network acoustic model adaptation for robust far-field speech recognition,” in *Interspeech*, pp. 2430–2434, 2015.

[6] K. Wang, et al., “Empirical evaluation of speaker adaptation on DNN based acoustic model,” in *Interspeech*, pp. 2429–2433, 2018.

[7] P. Ghahremani, et al., “Investigation of transfer learning for ASR using LF-MMI trained neural networks,” in *ASRU*, pp. 279–286, 2017.

[8] R. Gemello, et al., “Linear hidden transformations for adaptation of hybrid ANN/HMM models,” *Speech Communication*, vol. 49, no. 10, pp. 827 – 835, 2007.

[9] D Povey, et al., “Purely sequence-trained neural networks for ASR based on lattice-free mmi,” in *Interspeech*, pp. 2751–2755, 2016.