異なる疾患の障害者音声を用いた器質性構音障害者音声認識モデルの学習 *

☆冨士原健斗, 高島遼一 (神戸大), 杉山千尋, 田中信和,
野原幹司, 野崎一徳 (大阪大), 滝口哲也 (神戸大)

1　はじめに

構音障害は，病気やケガなどが原因で言葉をうまく発することができない状態のことを指す．このうち器質性構音障害は，音を作る際に使う器官の異常による構音障害である．例えば，口唇口蓋裂の患者であれば，唇や口の中の天井部分が裂けているために空気の流れを制御しにくくなる．舌癌の患者であれば，手術によって舌を切除すると舌が動かせなくなってしまう．Fig. 1 に健常者 (上図) と口唇口蓋裂者 (下図) の発話「一週間ばかり，ニューヨークを取材した」のスペクトログラムを示す．このような構音障害者の音声は，発声に多大な負担がかかっているだけでなく，フォルマントが異常な値を示すなどの特性を持ち [1]，聞き取ることが難しくなる．

近年，機械学習の発展を背景に，音声認識技術がスマートフォンのアプリやスマートスピーカーなど生活の様々な場面で利用されるようになっている．しかし，一般的な音声認識システムは健常者を対象として作られたものであるため，健常者と異なる特性を持つ構音障害者の音声はうまく認識できず，利用に不都合が生じる．したがって，構音障害者の音声を高精度に認識できるシステムを構築することが求められている．

音声認識システムを構築するためには，人間の音声を収録した学習データが必要不可欠である．健常者の音声については，日本語話し言葉コーパス (CSJ) [2]，LibriSpeech [3] など数百時間に及ぶ大規模なデータセットが公開されている．一方，構音障害者には発声の負担やプライバシーなどの問題があるため，大量のデータを収集することが難しい．そこで，構音障害者用の音声認識システムの構築は，健常者に比べて少量の学習データで行うことが求められる．少量の学習データから効果的な学習を行うために，我々はデータ拡張によるデータの増量 [4]，誤り訂正による精度の向上 [5] を試みてきた．これらの手法により，ある程度の改善は見られたものの，依然として音声認識精度は低いままであり，更なる改善の必要があった．

少量データのためのニューラルネットワークの代表的な学習手法の 1 つとして，転移学習が挙げられる [6]．Fig. 2 に，一般的な転移学習のイメージを示す．予め大規模なデータセットでモデルを事前学習し，普遍的な知識を獲得させた後，目的のドメインでの再学習などを行うことで，モデルのパラメーターを適応させる．これにより，目的のドメインで収集可能なデータが少量な場合でも，比較的高い精度のモデルを構築することが可能である．

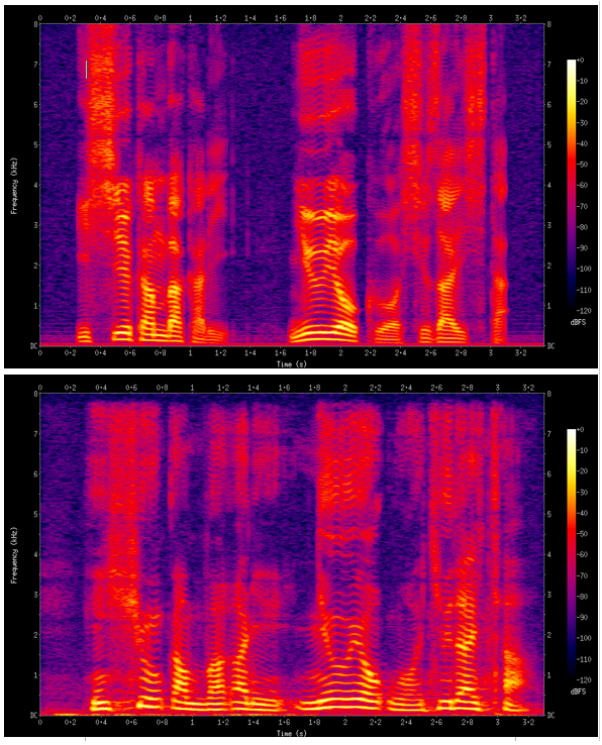

Fig. 1　Example of spectrogram uttered for /i q sh u: k a N b a k a r i ny u: y o: k u o sh u z a i sh i t a/ of a physically unimpaired person (top) and a person with cleft lip and cleft palate (bottom).

このような手法は，利用可能なデータが少ない器質性構音障害者のための音声認識モデル構築にも有効であることが期待される．具体的には，健常者音声のデータセットを用いて事前学習を行い，対象にする話者の音声で再学習を行うという手順になる．また，健常者音声だけでなく，対象話者以外の障害者が収録した音声を事前学習に利用することも考えられる．明瞭度が低い障害者音声であっても，複数の話者によって蓄積されたデータを利用することにより，健常者音声と同様の効果が期待できる。また，明瞭度が低い音声に対して，モデルの頑健性が高まることが期待できる。

* A training method using various dysarthric speech for speech recognition of organic dysarthria, by Kento Fujiwara, Ryoichi Takashima (Kobe University), Chihiro Sugiyama, Nobukazu Tanaka, Kanji Nohara, Kazunori Nozaki (Osaka University), Tetsuya Takiguchi (Kobe University)

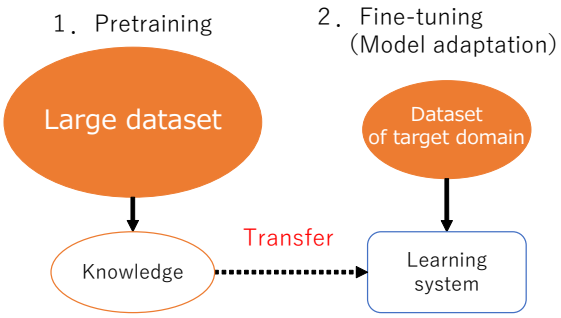

Fig. 2　Overview of general transfer learning.

従来の構音障害者音声認識の研究では，対象話者と同じ疾患を持つ障害者の音声を転移学習に利用した例があり，性能改善が報告されている [7]．しかし，構音障害者はその原因となる疾患によって特徴に大きく差があり，対象話者と異なる疾患を持つ障害者の音声は利用されていなかった．そこで本研究では，疾患によらず可能な限り大量の障害者音声を転移学習に利用することを検討する．

その場合，モデルが多様な音声を事前学習することから，対象話者の音声認識には不要な知識が転移されてしまう可能性が懸念される．そこで本研究では，モデル適応の効果を更に高めるため，話者ごとの差異が大きくなると考えられるデコーダーのみを事前学習から取り除くことを提案する．特定話者音声認識モデル，従来の転移学習に基づく音声認識モデル，そして提案する転移学習に基づく音声認識モデルをそれぞれ学習し，音素認識で評価を行い，提案法の有効性を確認する．

2　関連研究

高島ら [7] は，アテトーゼ型脳性麻痺患者を対象にした日本語音声認識システムを提案した．脳性麻痺者は，本研究で対象にしている器質性構音障害者とはその症状が異なるものの，音声が不明瞭になるなど構音障害者としての問題点は似通っている．

Fig. 3 に，文献 [7] で提案された音声認識システム構築の流れを示す．ここでは，健常者音声から障害の有無に依らない日本語音声の特徴を，英語の障害者音声から言語非依存な障害者特有の特徴を学習させるという転移学習手法を提案している．モデルのデコーダーを日本語用と英語用に分けることで，エンコーダー部分に共通の知識を蓄積すると共に，最終的には不要になる英語音声のデコードに関する知識を捨てることを可能にしている．

本研究では，このような手法を参考に，障害の有無や種類，話者の個性に依らない日本語音声の特徴をモデルに学習させるための手法を検討する．日本語と

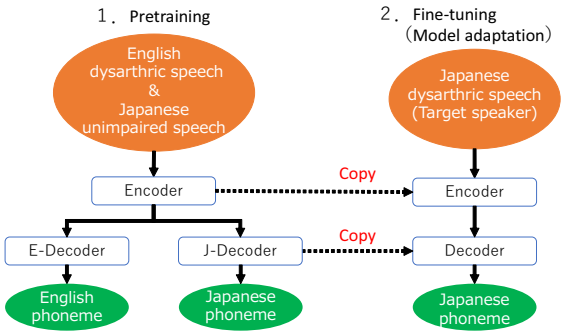

Fig. 3　Overview of transfer learning in previous work.

英語では発音の方法が大きく異なることは明らかであるが，同じ日本語の間でも話者によって発音の特徴は異なる．特に障害者音声では，その傾向は顕著になると言える．多様な障害者音声を転移学習に利用する場合は，話者ごとの差異がモデル適応に悪影響を与えないようにすることが重要になると考えられる．

3　提案手法

Fig. 4 に，提案する音声認識システム構築の流れを示す．まず，音素認識モデルを複数話者の音声で事前学習する．ここでは，話者ごとに個別のエンコーダーやデコーダーを構築することはせず，共通のパラメーターで学習を行う．これにより，モデルが話者に依らない一般的な音声の知識を獲得すると考えられる．

ここで，以下の 2 つの仮説を考える．

- 音声認識にとって重要な音声特徴量は，障害の有無や種類に関係なく，どんな話者でも共通である．
- 各音素に対する音声特徴量の分布は，個別の障害が持つ特性により，異なる可能性がある．

以上の仮説を元に考えると，入力音声から音声特徴量を抽出するエンコーダー部分は，複数話者の音声を利用して大量のデータによる事前学習を行い，続けてモデル適応を行うという従来の転移学習の流れが有効であると考えられる．一方，音声特徴量から音素を転写するデコーダー部分は，事前学習後にそのままモデル適応を行うことにより，対象話者の音声にはありえない発音パターンなど，不要な知識が転移してしまう可能性がある．

そこで，提案する転移学習では，事前学習後のモデル適応を行う直前に，デコーダー部分のパラメーターのみを初期化する．すなわち，対象話者の音声で再学習を行う際，エンコーダー部分は事前学習したパラメーターを fine-tuning するのに対し，デコーダー部

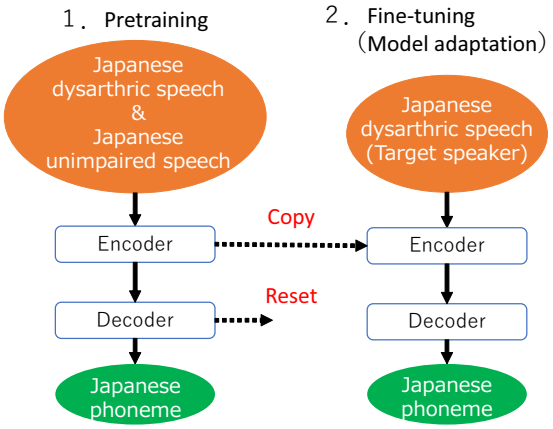

Fig. 4　Overview of our proposed transfer learning.

分は一から学習し直すということである．これにより，転移学習によってエンコーダー部分に蓄積された知識を利用しつつ，対象話者の発音しか学んでいない専用のデコーダーを構築することができ，音声認識精度の改善が期待される．

4　評価実験

4.1　実験条件

特定話者音素認識を行い，音素誤り率で評価した．評価話者として，口唇口蓋裂者男性 2 名 (CLP1-2)，舌切除後の舌癌患者男性 1 名 (TC1)，女性 1 名 (TC2) を対象にした．また，転移学習用の話者として，健常者男性 6 名，女性 4 名，アテトーゼ型脳性麻痺患者男性 4 名 (ACP1-4)，脊髄性筋萎縮症患者 (SMA 患者) 女性 2 名 (SMA1-2) の音声を利用した．アテトーゼ型脳性麻痺は，脳性麻痺による筋肉の不随意な運動が原因で構音障害を引き起こす．また，SMA は筋肉の萎縮が原因で構音障害を引き起こす．

音声データとして，ATR 研究用日本語音声データベース [8] に含まれる音素バランス文，または単語の読み上げ音声を収録した．障害の種類による音声の明瞭度の差異を音声認識率を用いて調べるため，どの話者のデータも学習データ，開発データ，テストデータの 3 つに分割した．具体的には，SMA 患者以外の話者について，それぞれ 200 文を 1 回ずつ収録し，このうち 100 文を学習データ，50 文を開発データ，50 文をテストデータとした．SMA 患者に関しては 216 単語を 5 回ずつ収録し，150 語を学習データ，30 語を開発データ，36 語をテストデータとした．

音声データのサンプリング周波数は 16kHz であり，音響特徴量として，フレームシフト 10ms，窓幅 25ms で抽出された 40 次元のメルフィルタバンク特徴を用いた．

音声認識モデルには，音素を出力単位とする CTC を用いた [9]．提案手法のエンコーダー部分にあたるモデルの中間層は，5 層の双方向 GRU [10] で構成され，各層で入力フレームを 2 分の 1 にサブサンプリングした．デコーダー部分にあたるモデルの出力層は，音素 40 種類に未知音素と CTC のブランクを加えた 42 次元とした．学習時のバッチサイズは 5，初期学習率は 0.001 とし，最適化には Adam [11] を用いた．

Table 1　Phoneme error rate [%] of the speaker-dependent dysarthria models.

Speaker	PER [%]
CLP1	27.35
CLP2	26.41
TC1	36.61
TC2	29.06
ACP1	53.80
ACP2	44.41
ACP3	64.60
ACP4	65.89
SMA1	84.10
SMA2	91.35

4.2　実験結果

まず，特定話者モデルを構築した場合の実験結果を Table 1 に示す．この実験では，音声認識率を指標として障害者音声の明瞭度を疾患ごとに比較するため，評価話者以外の話者 (ACP1-4 および SMA1,2) も評価した．

今回データを利用した構音障害者は，大きく分けて 4 種類の疾患を持っている．それぞれの音声認識モデルの誤り率には音声の明瞭度との相関があると考えられる．今回は SMA 患者のみ単語単位の音声データを用いて評価したため，完全に対等な条件とは言えないが，器質性構音障害者の口唇口蓋裂者と舌癌患者に比べ，転移学習用に音声を利用する脳性麻痺患者と SMA 患者の明瞭度は更に低くなっていることが分かる．このような明瞭度の違いは，重症度や発音の訓練を行っている度合いによっても影響を受けるが，やはりそれぞれの疾患の特徴に依るところが大きいと考えられる．本研究で提案する転移学習では，モデル適応を行う際にデコーダー部分を初期化しておくことで，このような疾患による差異が転移学習に悪影響を与えないようにすることが期待される．

続いて，転移学習を用いて器質性構音障害者 4 名それぞれのモデルを構築した場合の実験結果を Table 2 に示す．ここでは，事前学習を行った際に利用した音声ごとに，以下の 2 つのパターンで実験を行った．

Table 2 Phoneme error rate [%] of the speaker-dependent dysarthria models using transfer learning.

+ Physically unimpaired people + Other organic dysarthria		
Speaker	Baseline	Ours
CLP1	21.01	19.99
CLP2	23.58	21.73
TC1	31.58	30.81
TC2	25.38	24.39
+ All other speakers		
Speaker	Baseline	Ours
CLP1	20.61	19.92
CLP2	23.00	21.13
TC1	30.15	30.13
TC2	24.42	23.94

- 評価話者と健常者，他の器質性構音障害者 (口唇口蓋裂者および舌癌患者) の音声を利用した場合（+ Physically unimpaired people + Other organic dysarthria)
- 評価話者と健常者，他の器質性構音障害者，脳性麻痺患者，SMA 患者の音声を全て利用した場合（+ All other speakers）

なお，モデル適応時の学習を安定させるため，評価話者の音声は事前学習時のデータセットにも含めた．更に，従来の転移学習 (Baseline) と提案の転移学習 (Ours) をそれぞれのパターンで実験した．

どちらのパターンでも，転移学習を用いたモデル構築によって特定話者モデルよりも性能が改善していることが確認出来る．また，健常者音声や，評価話者と近い特徴を持つと考えられる他の器質性構音障害者の音声に加えて，より明瞭度の低い脳性麻痺患者や SMA 患者の音声を利用した場合，更に性能が改善していることが確認出来る．これにより，転移学習は評価話者より明瞭度の低い障害者音声を利用した場合でも有効に働くことが分かる．また，従来の転移学習で構築したモデルより，提案の転移学習で構築したモデルの方がより性能が改善していることが確認出来る．仮説から考えたように，モデル適応前にデコーダー部分を初期化することが話者ごとの差異による悪影響を防ぎ，性能改善に繋がったと考えられる．

5 まとめ

本研究では，器質性構音障害者の音声認識のために転移学習を利用することや，性質が異なると考えられる脳性麻痺患者や SMA 患者の音声を転移学習に利用することを検討した．構音障害者の音声に対するモデル適応に際して，デコーダー部分のパラメーターを初期化するという提案手法により，転移学習の有効性が更に高まることが確認された．今後の課題として，更に有効性の高い転移学習の手法や，構音障害者の発音に合わせた音声認識タスク設定の検討を行っていく予定である．

参考文献

[1] S. Sapir, “Formant Centralization Ratio: A Proposal for a New Acoustic Measure of Dysarthric Speech,” Journal of Speech Language Hearing Research, vol. 53, pp. 114-125, 2010.

[2] 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.

[3] V. Panayotov *et al.*, “Librispeech: An ASR corpus based on public domain audio books,” ICASSP, 2015.

[4] K. Fujiwara *et al.*,“ Data augmentation based on frequency warping for recognition of cleft palate speech,” APSIPA, pp.471-476, 2021.

[5] 冨士原健斗 他, “誤り訂正に基づく器質性構音障害者の音声認識精度向上の検討,” 日本音響学会 2021 年秋季研究発表会, pp. 1081-1084, 2021.

[6] S. J. Pan and Q. Yang, ”A Survey on Transfer Learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, 2010.

[7] Y. Takashima *et al.*, ”Knowledge Transferability Between the Speech Data of Persons With Dysarthria Speaking Different Languages for Dysarthric Speech Recognition,” in IEEE Access, vol. 7, pp. 164320-164326, 2019.

[8] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, pp. 357-363, 1990.

[9] A. Graves *et al.*, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” ICML, 1990.

[10] K. Cho *et al.*, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in EMNLP, pp. 1724-1734, 2014.

[11] D. Kingma *et al.*, “Adam: A method for stochastic optimization,” ICLR, 2015.