構音障害者音声認識における発話辞書適応の検討*

☆澤佑哉, 高島遼一, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

近年の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のレベルまで近づいている．音声認識デバイスには，音声命令によるハンズフリーな操作が可能であるという利点があるため，障害者への利用が求められている．本研究の対象としている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である．この障害を持つ患者の多くは手足が不自由であるため，音声認識デバイスの利用が期待される．しかし，筋肉の不随意運動は手足のみならず顔や舌の筋肉にも現れるため，正しい発音が困難となる構音障害を併発している患者も存在する．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．

構音障害者は発話時における身体への負担が大きい等の理由から，学習データを十分量収録することが困難である．そのため，少量の構音障害者音声から音声認識システムを学習する必要がある．これまでの構音障害者音声認識に関する研究では，Data augmentation を用いて学習データを増やすアプローチ [1] や，健常者音声で事前学習したモデルを障害者音声に適応させるモデル適応 [2] の手法が取られている．これらの研究では，音声認識モデルとして DNN-HMM ハイブリッドモデルや，End-to-End モデルが用いられている．多くの音響モデルは，通常音素単位でモデル化されており，音素列から単語列への変換には単語の発音情報を定義した発話辞書が用いられる．しかし，構音障害者は発話辞書に定義された通りに正しく発音することが困難である場合が多い．したがって，一般的に用いられている発話辞書は構音障害者の発音を反映しておらず，構音障害者音声認識に用いるのは必ずしも適切ではないと考えられる．このような不正確な発話辞書は，訓練時には誤った音素列ラベルを与えてしまい，認識時には誤った単語への転写を行なってしまうことに繋がる．

健常者を対象とした音声認識では，Acoustic-to-word モデルと呼ばれる End-to-End 音声認識モデルの研究も行われている [3]．Acoustic-to-word モデルはニューラルネットワークを用いて，音響特徴量から直接単語列への変換を行い，発話辞書を使わずに簡潔な構造で高精度なモデル化を実現できる．しかし，このモデルは従来の辞書ベースの音声認識モデルと比較して，大量の学習データが必要となる．そのため，構音障害者音声認識には適しておらず，本研究では従来の音素レベルで定義された辞書ベースの音声認識モデルを用いる．

本研究では，構音障害者のための発話辞書の適応を検討する．構音障害者の発話スタイルは各話者の症状によって異なるため，構音障害者ごとに発話辞書を適応させる．初めに，発話辞書を適応させるために，音素認識モデルでの認識結果から音素の誤認識パターンを分析し，発話辞書を修正するルールを決定する．決定したルールに従って，発話辞書に発音を追加することで，対象の構音障害者に発話辞書を適応させる．続いて，発話辞書適応の有効性を確認するために，大語彙連続音声認識タスクでの実験を行う．適応させた辞書を用いて特定話者音声認識モデルを学習し，単語認識で評価を行う．

2 音声認識における発話辞書

Fig. 1 は，音声認識における発話辞書の役割を示した図である．音響モデルは通常，音素レベルで細分化された単位で定義されている．ここで，音声認識の出力を単語列とした場合，単語に対する発音の情報が記された発話辞書を用いて音素列から単語列への変換を行う．また，音響モデルを学習する際には，逆に単語列を音素列へ変換し，音素列を正解ラベルとして扱う．この時，辞書に誤った発音が含まれている場合，誤ったラベルを用いた学習を行なってしまう．加えて認識時には，音素列から誤った単語列へ変換してしまうことに繋がる．一般的な発話辞書は健常者の発話を反映したものであり，これを健常者と発

*An investigation of adaptation of a pronunciation dictionary for dysarthric speech recognition.
by Yuya Sawa, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki (Kobe University)

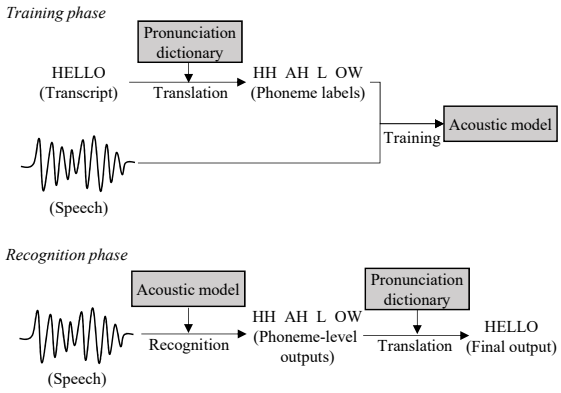

Fig. 1 Role of the pronunciation dictionary during the training and recognition phases of speech recognition.

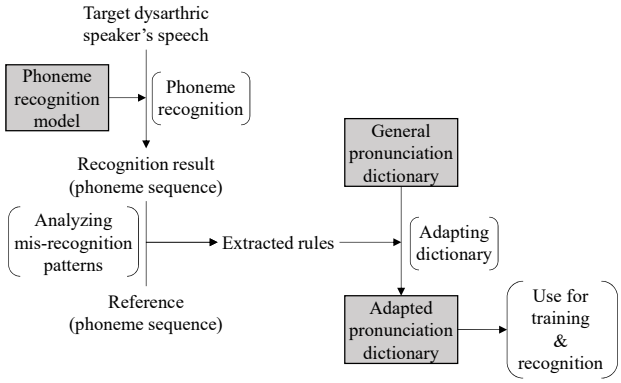

Fig. 2 System overview.

話スタイルが異なる構音障害者の音声認識にそのまま使用することは，必ずしも適切ではない．また，構音障害者の実際の発音は症状によっても異なる．本研究では，健常者用の一般的な辞書を個々の構音障害者に適応させることを試みる．

3 手法

Fig. 2 は提案手法の概要を示している．まず初めに，各構音障害者がどのように発音しているかを，音素認識タスクにより明らかにする．続いて，音素認識結果の誤認識パターンを分析し，発話辞書を修正するためのルールを決定する．最後に，ルールに基づいて修正された辞書を用いて特定話者モデルを学習し，単語認識タスクで評価を行う．

3.1 音素認識モデルによる音素の誤り傾向分析

音素の誤り傾向を分析するために，構音障害者音声を音素認識モデルで学習・認識させる．このモデルの学習の時点では，対象話者の実際の発音は分からないので，一般的な発話辞書に基づく正解ラベルを使用することに留意する．本研究では，一般的な発話辞書を用いることによる弊害を軽減するために，大量の健常者音声を用いたモデルの事前学習を行う．一般的な発話辞書は健常者の発話を対象にしているので，事前学習されたモデルは発話と音素ラベルの対応関係を適切に学習する．この事前学習したモデルを各構音障害者へと適応するために，各話者の発話に対してモデルを fine-tuning させる．

音素認識モデルには，End-to-End モデルの一種である Hybrid CTC/attention[4] モデルを用いている．このモデルは，CTC と注意機構付きのエンコーダデコーダモデルの 2 種類の End-to-End モデルを組み合わせたものである．DNN-HMM モデルでは，前後にある音素を考慮してモデル化を行い，各音素をさらに細分化して HMM の状態として定義する．このような詳細な音素のモデル化は，一般的な発話辞書を使用する際に影響を受けやすいと考えられる．一方の End-to-End モデルは，音素を単純なモノフォンとして定義するため，このような影響に対してより頑健なモデルになると期待される．

3.2 適応辞書の作成

本節では，音素認識モデルにおける認識結果から適応辞書を作成する手順を述べる．まず，評価データに 10 回以上出現する音素に着目し，それ以外の音素は信頼性が低いため分析対象から除外する．各音素について，すべての誤認識パターンの出現割合を以下のように計算する．

$$Rate_{i \to j} = \frac{Occ_{i \to j}}{Occ_i} \tag{1}$$

ここで，$Occ_{i \to j}$ および $Rate_{i \to j}$ は，音素 i を音素 j と誤認識した回数，割合をそれぞれ表している．また，Occ_i は評価データ内の音素 i の出現回数を表す．続いて，すべての誤認識パターンの中で発生率 $Rate_{i \to j}$ が最も高い上位 5 つの組み合わせを抽出する．この組み合わせに従って，一般的な発話辞書に登録されている単語の発音を修正し，新たな発音セットを追加する．例えば Fig. 3 に示すように，/p/ → /b/と置換する場合，発話辞書に登録されているすべての/p/を/b/に置き換えたパターンの発音を新たに追加する．また，

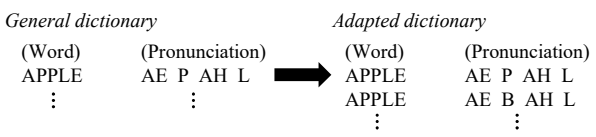

Fig. 3 An example of modifying dictionary based on a mis-recognition pattern /p/→/b/.

複数の置換パターンが考えられる場合は，すべての置換パターンを辞書に追加する．

発話辞書を適応させた後，各構音障害者に対して，適応辞書を用いて単語認識タスクを行う．単語認識モデルとしては，End-to-End モデルではなく DNN-HMM ハイブリッドモデルを使用する．単語レベルの End-to-End モデルは学習に大量のデータを必要とするため，発話辞書を適切に適応できれば DNN-HMM ハイブリッドモデルの方が認識に適していると考えられる．

4 評価実験

4.1 実験条件

提案手法の有効性を評価するため，アテトーゼ型脳性麻痺を持つ構音障害者 2 名 (SPK1/SPK2) を対象に，日本語大語彙連続音声認識の実験を実施した．話者 SPK1 と SPK2 について，ATR 日本語データベース [5] に含まれる音素バランス文を，それぞれ 429 文と 501 文収録した．各話者について，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，音素認識および単語認識モデルの学習に使用した．発話辞書は，CSJ コーパス [6] を用いて構築した．ATR データセットには存在するが CSJ データセットに存在しない単語については，形態素解析エンジン MeCab[7] を用いて，ATR データセットのスクリプトに対して形態素解析を行い，未知語を辞書に登録した．辞書の総語彙数は約 73,000 語であった．

音素認識には，End-to-End 音声認識ツールキット ESPnet[8] を用いて，Hybrid CTC/attention モデルの学習を行なった．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にピッチ特徴を加えた計 83 次元の特徴量を用いた．出力次元数は，音素 39 種類に未知文<unk>，・ 始端記号<sos>，・終端記号<eos>，を加えた 42 次元とした．Hybrid CTC/attention モデルのエンコーダは，320 次元の隠れ層を持つ 4 層からなる pBLSTM とした．デコーダは 300 次元の隠れ層を持つ 1 層からなる単方向 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．注意機構には，location-aware attention を使用し，最適化には AdaDelta を用いた．マルチタスク学習時には CTC の損失関数の重みを 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．音素認識モデルについては，CSJ データセットに収録されている約 240 時間の健常者音声を用いて事前学習をした後に，構音障害者の訓練データでモデル適応を行なった．

単語認識モデルの学習および評価には，音声認識ツールキット Kaldi[9] を用いた．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を適用した．音響モデルは，全結合層と time-delay neural network (TDNN) 層からなる．隠れ層のノード数は 625 であり，活性化関数には ReLU を用いた．音響モデルは LF-MMI 基準 [10] を用いて学習した．初期学習率は 0.001 とし，クロスエントロピー正規化を 0.1 の重みで適用し，エポック数は 4 とした．言語モデルとして，tri-gram モデルを CSJ データセットのテキストを用いて学習した．

4.2 実験結果

4.2.1 音素誤り傾向の分析

Fig. 4 は，各話者の音素認識における混同行列である．また Table 1 は，この混同行列から得られた誤認識率が高い上位 5 つの音素のペアを示している．各話者ともに子音の誤認識が多く見られ，特に/p/や/f/のような息を強く吹く無声子音が認識されにくいことが分かる．その一方で，話者特有の誤認識パターン (SPK1 の/ts/と/s/，SPK2 の/ch/と/e/) もあり，話者ごとの発話辞書適応が必要であると言える．

4.2.2 適応辞書を用いた単語認識

得られた誤認識パターンに従って，各話者に発話辞書を適応させた．Table 2 は，一般的な発話辞書と適応辞書をそれぞれ用いた場合の単語認識モデルにおける単語誤り率 (word error rate; WER) を示している．各構音障害者に発話辞書を適応することで，SPK1 で 9.27%，SPK2 で 3.93%の相対性能改善を達成した．これらの結果から，構

Table 1 The extracted substitution rules with their occurrence rates $Rate_{i\to j}$ of each dysarthric speaker.

Speaker	1	2	3	4	5
SPK1	z→d	f→s	ts→sh	ts→ch	s→y
	0.190	0.143	0.143	0.107	0.098
SPK2	f→s	p→t	z→d	ch→k	e→i
	0.294	0.250	0.185	0.179	0.169

Table 2 WERs [%] of word-recognition models with general dictionary and those with adapted dictionary.

Speaker	WER	
	General dictionary	Adapted dictionary
SPK1	51.88	47.07
SPK2	64.57	62.03

音障害者に発話辞書を適応させることで，特定話者認識の性能が向上することが分かる．

5 おわりに

本研究では，構音障害者音声認識を対象にした発話辞書の適応を行なった．発話辞書を修正するルールを選定するため，音素認識結果における誤認識パターンを分析し，特に無声子音が誤認識されやすい傾向があることが明らかになった．分析結果を用いて発話辞書を各構音障害者に適応させることで，単語の誤認識率が減少することを確認した．今後は，単純な置換誤りの他にも挿入誤りや削除誤り，隣接する音素への依存性といった，より複雑なルールを考慮した発話辞書適応を検討する．

参考文献

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

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

[3] S. Ueno *et al.*, "Acoustic-to-word attention-based model complemented with character-level CTC-based model," in *ICASSP 2018*, 2018, pp. 5804–5808.

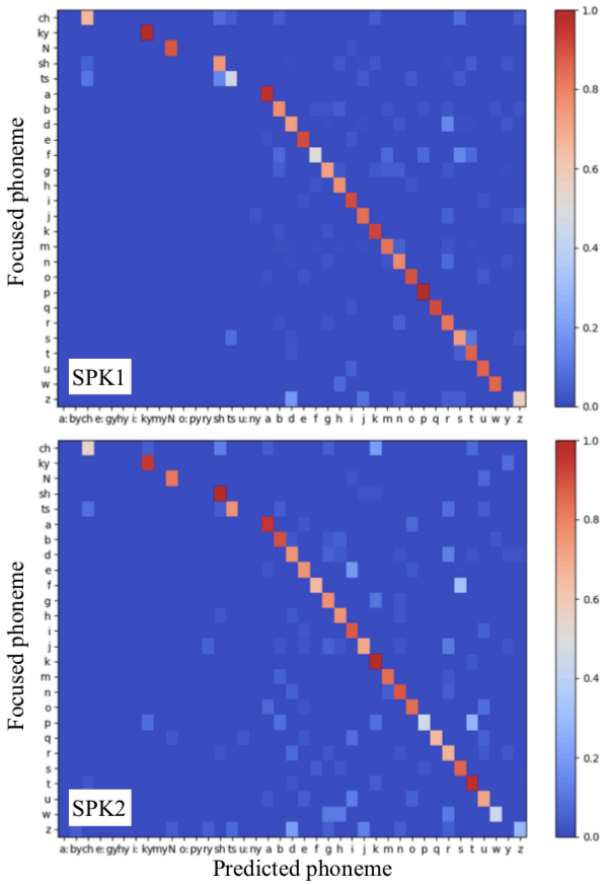

Fig. 4 Confusion matrices of phoneme recognition. Each cell in the confusion matrix reflects the occurrence rate $Rate_{i\to j}$ in Equation (1).

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

[5] 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.

[6] K. Maekawa, "Corpus of spontaneous Japanese: its design and evaluation," *Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pp. 7–12, 2003.

[7] T. Kudo *et al.*, "Applying conditional random fields to Japanese morphological analysis," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 230–237.

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

[9] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE Signal Processing Society, 2011.

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