器質性構音障害者向け音声認識モデルにおける発話辞書適応方式の比較検討 *

☆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．口腔腫瘍の患者であれば，治療のために舌の大部分を切除することで子音の区別が難しくなってしまう．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．そのため，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] などのアプローチを検討してきた．

このような構音障害者の音声認識に関して，澤ら [6] は発話辞書の適応を提案している．多くの音響モデルでは，音素単位のモデル化を行い，単語の発音情報を定義した発話辞書によって音素列から単語列への変換を行う．しかし，構音障害者は健常者と異なる発話スタイルを持っているため，健常者の発話スタイルに沿って定義された発話辞書をそのまま利用する場合，音素列から単語列の変換において誤った転写を行ってしまう可能性が高くなる．そのため，各構音障害者の発話スタイルに合わせて発話辞書を適応させることによって，音声認識精度を向上させることが可能だと考えられる．

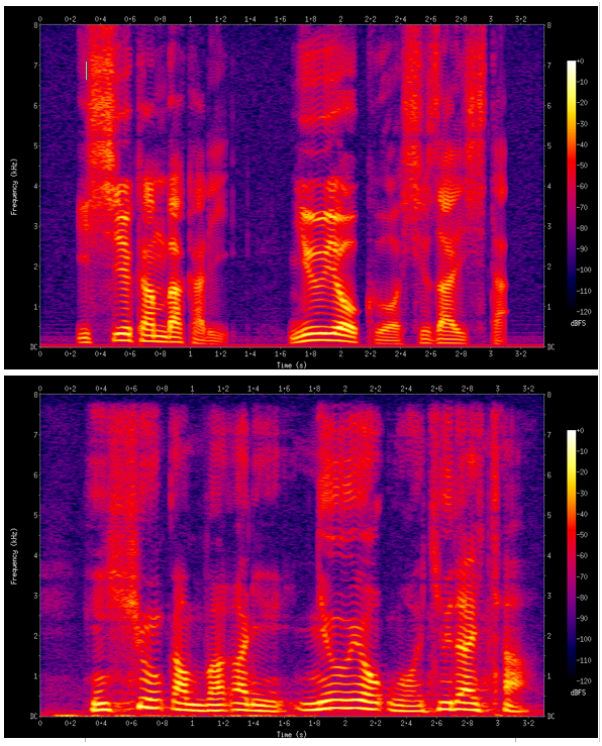

Fig. 1　Example of spectrogram uttered for /i q sh u: k a N b a k a r i ny u: y o: k u o sh u z a i sh i t a/ of a physically unimpaired person (top) and a person with cleft lip and cleft palate (bottom).

先行研究 [6] では脳性麻痺による構音障害者を対象として検証していたが，脳性麻痺者は発声時の不随意運動によって構音障害を引き起こすため，その発話スタイルは一定していなかった．一方、本研究で対象とする器質性構音障害者については，器質的な異常によって発音が難しくなるため，話者ごとの発話スタイルが明確に現れやすいと考えられる．そこで，本研究では参考文献 [6] の手法を元に，器質性構音障害者の音声認識に対する発話辞書の有効性を検証する．

初めに評価話者の音声で音素認識モデルを構築し，その認識結果から音素単位の誤認識パターンを分析する．分析結果を元に発話辞書に発音の定義を追加することで，評価話者に適応した発話辞書を作成する．その後，発話辞書適応の有効性を確認するため，

* A comparison of adaptation methods of a pronunciation dictionary 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)

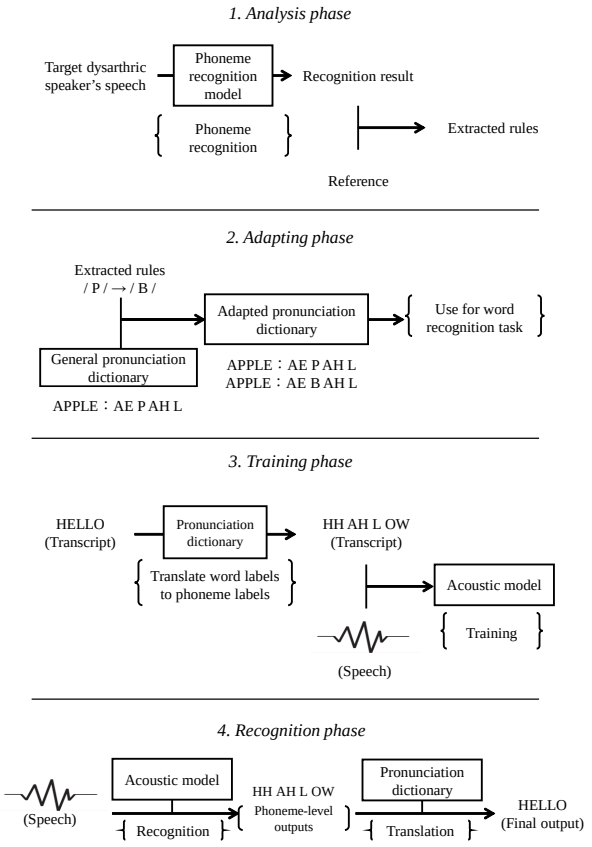

Fig. 2　Overview of our proposed method.

大語彙連続音声認識タスクによる実験を行う．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．辞書の適応では複数の異なる条件を試行し，効果的な手法を検討する．

2　手法

2.1　音素認識モデルによる音素の誤り傾向分析

Fig. 2 は提案手法の概要を示している．初めに，音素の誤り傾向を分析するために，構音障害者音声を学習させた特定話者音素認識モデルを構築する．このモデルの学習の時点では，対象話者の発音の特徴は分からないため，一般的な発話辞書に基づく正解ラベルを使用する．この後，学習データを対象にして認識結果を確認する．学習済みのデータに対して音素の誤りが発生する要因は，一般的な発音と対象話者の発音が異なっているためだと考えられる．本研究では，そのような音素の誤りを軽減するために発話辞書の適応を試みる．

2.2　適応辞書の作成

音素認識モデルの認識結果を分析し，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 $Rate_{i\to j}$ が高い上位の組み合わせを抽出する．この組み合わせを元に，発話辞書に登録されている単語の発音を修正した新たな発音セットを追加する．

例えば，/t/→/k/という誤認識パターンを抽出する場合，発話辞書に登録されているすべての/t/を/k/に置き換えたパターンの発音を新たに追加する．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．なお，サンプル数が少ない音素は誤認識パターンの信頼性も低いと考えられるため，分析対象は学習データに 10 回以上出現する音素のみとしている．発音の追加を行った辞書 (適応辞書) を作成した後は，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．

3　評価実験

3.1　実験条件

評価話者として，口唇口蓋裂者男性 2 名 (CLP1-2), 舌切除後の口腔腫瘍患者男性 2 名 (TC1-3), 女性 1 名 (TC4) を対象にした．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．それぞれ 503 文を 1 回ずつ収録し，このうち 50 文を開発データ，50 文をテストデータとした．残りを学習データとして，音素認識モデルの学習および単語認識モデルの fine-tuning に使用した．

発話辞書は，CSJ コーパスを用いて構築した．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 [9] 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．この時点で，辞書の総語彙数は約 73,000 語であった．

音素認識実験には，音素を出力単位とする CTC モデルを用いた．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．モデルは 320 次元の隠れ層を持つ 5 層の双方向 GRU [10] と，出力層にあたる全結合層で構成され，1 層目と 2 層目で入力フレームを 2 分の 1 にサブサンプリングした．出力次元数は，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 The extracted substitution rules with their occurrence rates $Rate_{i\to j}$ of each dysarthric speaker.

Speaker	1	2	3	4	5
CLP1	$gy \to j$ 0.21	$z \to g$ 0.11	$hy \to sh$ 0.09	$a :\to a$ 0.06	$ch \to k$ 0.05
CLP2	$gy \to j$ 0.23	$a :\to a$ 0.06	$ch \to k$ 0.05	$z \to d$ 0.05	$ky \to ch$ 0.05
TC1	$by \to d$ 0.15	$ry \to y$ 0.16	$p \to k$ 0.16	$gy \to j$ 0.12	$hy \to sh$ 0.09
TC2	$p \to k$ 0.10	$gy \to j$ 0.09	$by \to gy$ 0.08	$ry \to y$ 0.06	$a :\to a$ 0.06
TC3	$gy \to j$ 0.07	$a :\to a$ 0.04	$ky \to ch$ 0.04	$ry \to j$ 0.03	$z \to r$ 0.02
TC4	$a :\to a$ 0.40	$i :\to i$ 0.29	$e :\to e$ 0.15	$by \to b$ 0.11	$ny \to n$ 0.08

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 [12] を用いた．入力データとして，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 [13] 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．なお，音響モデルの学習では CSJ データセットに収録されている約 240 時間の健常者音声で事前学習を行い，評価話者の学習データで fine-tuning を行った．また，言語モデルには tri-gram モデルを用い，CSJ データセットのテキストを用いて学習した後，音響モデルの fine-tuning を行う際に適応辞書を用いて再学習を行った．

3.2 実験結果

3.2.1 音素誤り傾向の分析

Table 1 は，各話者の音素認識実験結果から得られた，誤認識率が高い上位 5 つの音素のペアを示している．どの話者についても，/a:/→/a/のように長母音と短母音の混同が見られる．口唇口蓋裂者に関しては，発声時に空気が漏れるため，/z/や/gy/のように口腔内の閉鎖を必要とする音や，/ch/のような破裂音が認識されにくい傾向がある．また，口腔腫瘍患者については，/ry/のような舌の挙上を必要とする音が認識されにくい傾向がある．

3.2.2 適応辞書を用いた単語認識

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文字誤り率 (character error rate; CER) を示している．ここで，

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

	Speaker					
N_p	CLP1	CLP2	TC1	TC2	TC3	TC4
0	**16.75**	17.22	**33.14**	28.50	23.94	14.14
5	18.51	**16.35**	33.38	28.02	23.41	14.41
10	18.26	16.65	34.23	28.92	**23.13**	**13.70**
15	17.72	17.12	33.68	**27.93**	24.31	14.38

N_p は発話辞書の作成を行う際に発生率 $Rate_{i\to j}$ が上位の組み合わせを何位まで抽出したかを示す．$N_p = 0$ の場合は一般的な発話辞書をそのまま使用している．今回構築したのは単語単位の音声認識モデルだが，区切り位置や表記揺れによる誤り率集計への悪影響を軽減するため，単語誤り率ではなく文字誤り率によって評価を行った．Table 2 によると，各構音障害者に発話辞書を適応する際，精度が改善するかどうかは話者によって異なる．また，N_p を増加させた際の誤り率の増減は一定でないことが分かる．本研究では，音素認識モデルの学習データに対する認識結果に基づいて発話辞書の適応を行った．しかし，今回の実験条件では評価話者の学習データや評価データが少量であるため，学習データと評価データの間で語彙や音素の出現数に乖離がある．このため，学習データに基づく適応辞書では，必ずしも評価データに対して有効ではないと考えられる．CLP1，TC1 のように発話辞書適応による精度の悪化が見られる話者については，特に学習データと評価データの分布の違いが大きいと考えられる．

学習データの認識結果から抽出した誤認識パターンについて，1 つ 1 つの組み合わせがもたらす影響をより詳細に調べるため，発生率 $Rate_{i\to j}$ が 1〜5 位の組をそれぞれ単独で抽出し，発話辞書を作成した．Table 3 は，それぞれの適応辞書を用いた単語認識モデルにおける文字誤り率を示している．各話者名の下に示した値は，一般的な辞書を使用した場合の誤り率である．誤認識パターンの組み合わせを個別に用いて適応辞書を作成したことにより，一般的な発話辞書を使用した場合に比べて精度が改善する組と改善しない組が混在していることが明らかになった．また，各組の発生率 $Rate_{i\to j}$ と精度への影響の大きさは比例していない．この結果から，単に発生率 $Rate_{i\to j}$ が上位の誤認識パターンから適応辞書を作成する今回の手法では，評価データの認識精度を改善するのに不適切なルールが混入してしまうことが分かる．

適応辞書による具体的な影響を調べるため，TC3 について一般的な発話辞書を用いた場合と，発生率

Table 3 Character error rates [%] of word-recognition models with adapted dictionary using single error pattern.

Speaker	1	2	3	4	5
CLP1	$gy \rightarrow j$	$z \rightarrow g$	$hy \rightarrow sh$	$a :\rightarrow a$	$ch \rightarrow k$
16.75	17.20	17.59	17.96	17.07	17.72
CLP2	$gy \rightarrow j$	$a :\rightarrow a$	$ch \rightarrow k$	$z \rightarrow d$	$ky \rightarrow ch$
17.22	17.00	16.55	17.42	17.23	16.52
TC1	$by \rightarrow d$	$ry \rightarrow y$	$p \rightarrow k$	$gy \rightarrow j$	$hy \rightarrow sh$
33.14	32.78	34.76	32.41	33.78	33.20
TC2	$p \rightarrow k$	$gy \rightarrow j$	$by \rightarrow gy$	$ry \rightarrow y$	$a :\rightarrow a$
28.50	29.44	28.21	28.87	27.29	28.42
TC3	$a :\rightarrow a$	$i :\rightarrow i$	$e :\rightarrow e$	$by \rightarrow b$	$ny \rightarrow n$
23.94	23.98	24.19	23.35	23.40	23.74
TC4	$gy \rightarrow j$	$a :\rightarrow a$	$ky \rightarrow ch$	$ry :\rightarrow j$	$z \rightarrow r$
14.14	13.61	14.48	13.98	14.69	14.37

Table 4 The examples of recognition results with general dictionary and those with adapted dictionary.

Reference	逆境に耐えた /gy a q ky o: n i t a e t a/
Hypothesis *Original*	弱強に耐えた /j a k u ky o: n i t a e t a/
Hypothesis $/gy/ \rightarrow /j/$	逆境に耐えた

$Rate_{i \rightarrow j}$ が 1 位の組を用いて適応辞書を作成した場合とで認識結果が変化した例を Table 4 に示す．これらを見比べると,「弱強 (j a k u ky o:)」という誤りが「逆境 (gy a q ky o:)」という正しい単語として認識出来るようになっている．これは，適応辞書が誤り組 $/gy/ \rightarrow /j/$ に対応した結果，音素/j/を発音パターンに含む単語ではなく，音素/gy/を発音パターンに含む単語が出力されやすくなったためだと考えられる．全ての誤認識パターンが精度の改善に寄与するとは限らないが，このように精度の改善に寄与するパターンが多く含まれるルールで適応辞書を作成すれば，Tabel 2 に示した実験結果を更に改善することも可能であると考えられる．

4 まとめ

本研究では，器質性構音障害者音声認識を対象にした発話辞書の適応方式を検討した．発話辞書を修正するルールを選定するため，音素認識モデルにおける学習データの誤認識パターンを分析した結果，口唇口蓋裂者と口腔腫瘍患者のそれぞれに症状と誤認識の関連性が確認された．分析結果をもとに発話辞書を適応し，評価データの認識結果を確認することで，話者や条件によっては精度が改善されることが確認できた．一方，精度の改善に寄与しない誤認識パターンの存在も確認された．このため，今後は各評価話者の個別の誤り傾向を分析するほか，発話辞書適応の有効性をさらに高めるルールを検討する予定である．

謝辞 本研究の一部は，JSPS 科研費 JP21H00906, JP22K12168 の支援を受けたものである．

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. 114-12, 2010.
[2] K. Maekawa, “Corpus of spontaneous Japanese: Its design and evaluation,” in SSPR, pp. 7-12, 2003.
[3] V. Panayotov *et al.*, “Librispeech: An ASR corpus based on public domain audio books,” in ICASSP, 2015.
[4] K. Fujiwara *et al.*, “Data augmentation based on frequency warping for recognition of cleft palate speech,” in 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,” 音講論 (秋), pp. 1081-1084, 2021.
[6] Y. Sawa *et al.*, “Adaptation of a Pronunciation Dictionary for Dysarthric Speech Recognition,” in LifeTech, pp.612-616, 2022.
[7] A. Graves *et al.*, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in ICML, 2006.
[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] T. Kudo *et al.*, “Applying conditional random fields to Japanese morphological analysis,” in EMNLP, pp. 230-237, 2004.
[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,” in ICLR, 2015.
[12] D. Povey *et al.*, “The Kaldi speech recognition toolkit,” in ASRU, 2011.
[13] D. Povey *et al.*, “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” in Interspeech, pp. 2751-2755, 2016.