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*

☆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，公開されているデータセットは希少である．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を試みてきた [4]．構音障害者の発話の特性に合わせた手法を提案し，一定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．

音声認識システムの性能を改善するための手法の 1 つとして，言語モデルがある [5]．音声認識モデルが入力音声から文字系列を推定するのに対して，言語モデルは入力された文字系列が言語として妥当であるかどうかを評価するものであり，音声認識モデルと組み合わせることでより正確な認識を可能にする．また，言語モデルを発展させたアプローチとして，誤り訂正モデルがある [6]．これは単に正しい言語を学習させるのではなく，入力された音声認識結果の誤りを正解に近くなるよう修正させることを主眼に置いたものである．

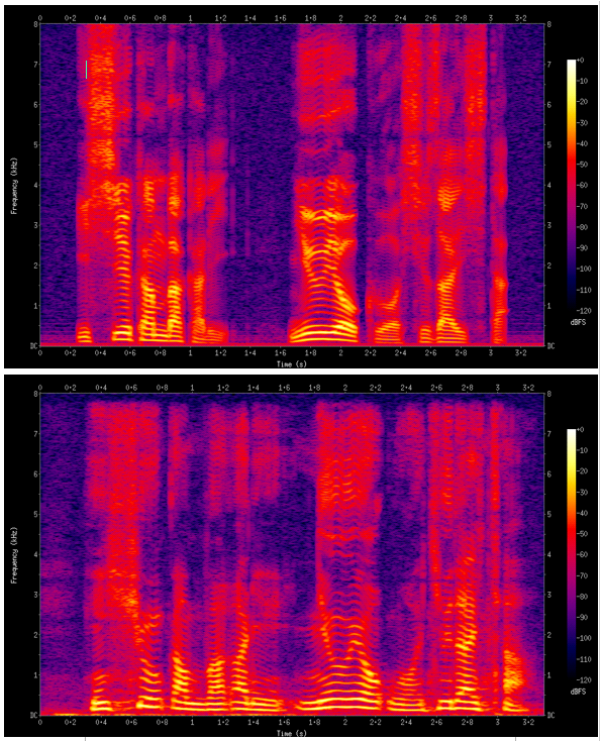

Fig. 1　Example of spectrogram uttered for /i q sh u: k a N b a k a r i ny u: y o: k u o sh u z a i sh i t a/ of a physically unimpaired person (top) and a person with cleft lip and cleft palate (bottom).

器質性構音障害者の音声が健常者に比べて不明瞭になるのは，症状の特徴が声に反映されているためである．口唇口蓋裂者を例にすると，裂によって口腔内圧が低くなる代償として，声門を強く締めることで音を発する現象が見られる [7]．構音障害者音声を用いて音声認識システムを構築する際，このような話者ごとの発話の特性は学習にも反映され，認識精度の低下をもたらすと考えられる．つまり，認識結果

*An investigation of error correction method for improving 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)

の誤りにはある程度話者ごとに異なる傾向が存在すると考えられる．

そこで本研究では，話者ごとに傾向の異なる誤りを補正する誤り訂正モデルを用いることで，音声認識の精度向上を検討する．一般に誤り訂正モデルの学習には，誤りを含む文と正解文のペアが必要であり，構音障害者音声のデータが限られている現状では誤り訂正モデルの学習データも限られることになる．この課題に対し，本研究では音声認識結果の誤り傾向を分析し，それを反映した疑似誤り文を生成することによって学習データを増量する．特定話者音声認識モデルと特定話者誤り訂正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

Mani ら [6] は，既製の音声認識モデルを用いて認識した結果を，誤り訂正モデルを用いてより高精度にする手法を提案した．文献 [6] では健常者を対象にした，単語単位の音声認識モデルを利用している．単語単位の音声認識では，認識精度が語彙の種類に左右されるという問題が起きる．例えば，医療現場での会話を認識する際，薬の名前などの専門用語がモデルの学習データに含まれていないために，発音が似ている別の単語として認識されてしまう場合がある．そこで文献 [6] では，医療現場での会話を認識した誤りを含む結果を正解文に直す誤り訂正モデルを学習した．これにより，医療用語に対する認識精度が改善されることを示した．

本研究では，上記のような手法を参考に，音素単位の音声認識システムを構築する．構音障害者の音声認識では，学習データが少量であるため，単語単位の音声認識モデルを学習することが困難である．また，構音障害者それぞれが持つ誤りの傾向は，細かい発音の違いによるものであるという仮説に基づくと，誤り訂正モデルによる補正を行うには音素単位でモデルを構築することが効果的だと考えられる．

3　提案手法

Fig. 2 に，提案する音声認識システム構築の流れを示す．まず，音素認識モデルを対象話者の音声で学習する．学習を終えたあと，この音素認識モデルのテストを行い，その結果に含まれる誤りの傾向を確認する．続いて，この誤り傾向を元に日本語音素文から擬似誤り文を生成する．この時，学習のパターンを増やすため，まず正解文に対してデータ拡張を行い，それによって増量した各文に対して個別に誤り文の生成を行う．最後に，データ拡張によって作られた正解文とそれに対する疑似誤り文のペアを用いて誤り訂正モデルを学習する．

3.1　正解文に対するデータ拡張

疑似誤り文の生成前に行う，正解文に対するデータ拡張について詳細を述べる．自然言語処理におけるデータ拡張としては，入力文・正解文の語彙をランダムに置換する SwitchOut [8] などがある．本研究では，文献 [8] と同様に，正解文の音素をランダムに置換する処理を行う．

まず，正解文の各音素について，置換するかどうかを一定の確率で決定する．置換を行う場合，置換先の音素は他の音素から等確率で選択するが，日本語として不自然な音素列が出来ると学習に悪影響を与える可能性があるため，選択可能な音素を制限する．具体的には，子音に対応する音素であれば子音の音素だけを選択可能にする．母音についても同様にし，撥音と促音は相互に置換される．なお,「a i」が「a a」になるなど，音素の置換によって同じ母音が複数並んだ場合は 1 つの長母音にまとめる．さらに,「py i」など，本来の文で使用されていない子音と母音の組合せが生じた場合は，子音を他の適切な音素になるよう選択し直す．

3.2　疑似誤り生成

データ拡張の後に行う疑似誤り生成について詳細を述べる．Fig. 3 に，疑似誤り生成の例を示す．音声認識時の誤りには，置換・削除・挿入の三種類があるが，今回生成する誤りは置換誤りのみを対象にする．音声認識モデルのテストデータに対する認識結果 (Recognition result) について，正解文 (Reference) と比較して誤り数を音素ごとに数え上げると，その音素の出現数から見た置換誤りの発生率と，誤りとして出現する音素の確率分布を求められる．これらを用いて，データ拡張によって生成された誤り訂正学習用の文 (Augmented text) に含まれる音素を 1 つずつ置換し，擬似的な置換誤りを含む文 (Pseudo error text) を生成する．この 2 つの文をペアにし，疑似誤り文の誤りを修正するように誤り訂正モデルを訓練することによって，誤り訂正モデルが音声認識モデルの認識結果をうまく正解文に近づけられるようになることが期待される．

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/CLP2)，舌切除後の舌癌患者男性 1 名 (TC) を対象にした．音声データとして，ATR 研究用日本

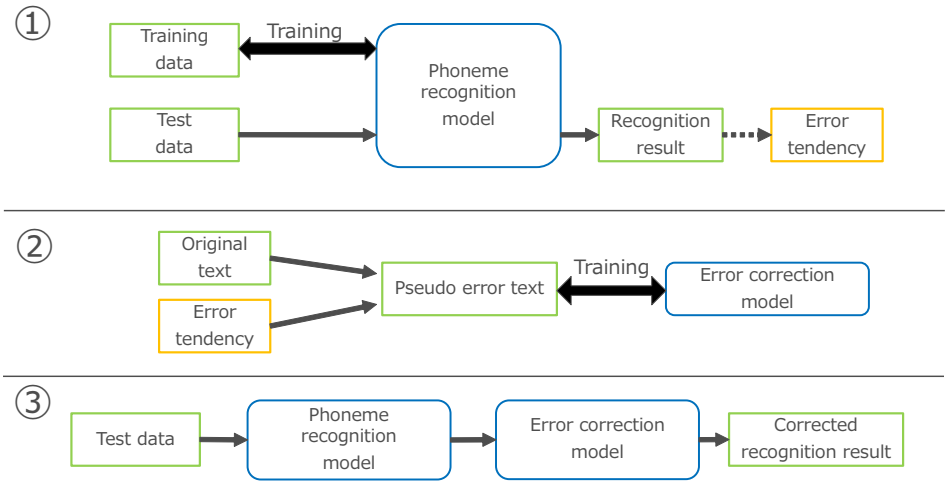

Fig. 2　Overview of our proposed method.

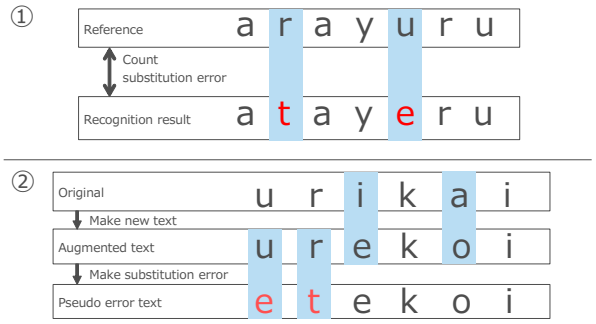

Fig. 3　Overview of our pseudo error generation.

語音声データベース [9] に含まれる音素バランス文の読み上げ音声を，それぞれ 495 文，503 文，503 文収録した．音声認識モデルの学習では，このうち 50 文を開発データ，50 文をテストデータとし，残る 395 文 (403 文) を学習データとした．

誤り訂正モデルの学習を行う際は，音声認識モデルの学習データと同様の 395 文 (403 文) を元に 790 文 (806 文) の疑似誤り文を生成し，開発データとして使用した．また，日本語音声コーパス JSUT に含まれる 5000 文を元に 25 万文の疑似誤り文を生成し，学習データとして使用した．

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 [10]．モデルの中間層は 5 層の双方向 GRU [11] 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．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 [12] を用いた．

誤り訂正モデルには Seq2Seq モデルを用いた．エンコーダーは 6 層の双方向 GRU で構成され，デコーダーは 1 層の Attention [13] つき単方向 LSTM [14] で構成された．出力層は音声認識モデルと同様の 42 次元とした．学習時のバッチサイズは 50，初期学習率は 0.15 とし，最適化には Adam を用いた．

4.2　実験結果

実験結果を Table 1 に示す．また，Table 2 に置換誤り (SUB)，削除誤り (DEL)，挿入誤り (INS) の内訳を示す．音声認識モデルによる認識結果（Recognition）と，それを誤り訂正モデルで修正した結果（Error correction）の音素誤り率を比較すると，どの話者についても誤り訂正による精度の改善が確認出来る．

3 人の評価話者のうち，TC は最も誤り率が高く，誤り訂正による改善も最も大きかった．口唇口蓋裂と舌の切除では，舌の切除のほうがより発声を難しくしているほか，置換誤りの傾向がよりはっきりしていると考えられる．一方，CLP2 は 3 人の中で最も誤り訂正による改善が小さかった．その要因として，音声認識結果に含まれる誤りのうち，削除誤りの占める割合が大きかったことが挙げられる．今回行った疑似誤り生成は置換誤りにしか対応していないため，相対的に誤り訂正の効果が小さなものになったと考えられる．

Table. 3 に，テストデータの正解文 (Reference)，音声認識モデルの認識結果 (Recognition result)，それを誤り訂正モデルで訂正した結果 (Correction result) の例を示した．正解文と比較して誤りが見られる音素を太字で表している．今回の実験では，全体として誤

Table 1 PER [%] of the speaker-dependent dysarthria models.

Speaker	Recognition	Error correction
CLP1	16.29	**15.67**
CLP2	19.01	**18.94**
TC	26.65	**25.68**

Table 2 PER [%] for each error type.

	Recognition			Error correction		
Speaker	SUB	DEL	INS	SUB	DEL	INS
CLP1	12.31	1.74	2.24	**11.69**	1.74	2.24
CLP2	12.57	4.15	2.29	**12.50**	4.15	2.29
TC	20.54	4.20	1.91	**19.57**	4.20	1.91

り訂正による改善は見られたものの，相対改善は小さなものになっている．例にも見られるように，認識結果に生じる置換誤りは一部しか修正されていないことが分かる．一因としては，単一の音素を考慮するmonophoneで実験を行ったため，置換誤りが発生した際の前後の音素との関係性が疑似誤りの生成やモデルの学習に反映されていない可能性が考えられる．また，学習データのほとんどは誤りではなく正しい音素となっているため，誤り訂正モデルが音素の変換をほとんど行わない消極的なモデルになってしまった可能性がある．

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の誤り訂正モデルの利用を検討した．データ拡張と疑似誤り生成を用いて誤り訂正モデルの学習を行った結果，完全に誤りを消去するには至らなかったものの，誤り訂正モデルが音声認識精度を改善するのに有効であることが確認された．今後の課題として，今回対象にしなかった削除誤り・挿入誤りを訂正するための手法や，より効果的な学習の方法などについて検討していく予定である．

Table 3 Example of test data and output.

Reference
ky u: ky u: sh a g a j u: b u N n i u g o k e z u ky u: j o s a gy o: g a o k u r e t e i r u
Recognition result
ky u: ky u: sh a g a **gy** u: b u N n i u g o k **i** z u **N ch** u: j o s a gy o: g a o k u r e t e i r u
Correction result
ky u: ky u: sh a g a j u: b u N n i u g o k **i** z u **N** ky u: j o s a gy o: g a o k u r e t e i r u

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] 冨士原健斗 他, “口唇口蓋裂者の音声認識のためのデータ拡張方式の検討,” 日本音響学会 2021 年春季研究発表会, pp. 961-964, 2021.

[5] R. Kuhn *et al.*, “A cache-based natural language model for speech recognition,” IEEE PAMI, vol. 12, pp. 570-583, 1990.

[6] A. Mani *et al.*, “ASR Error Correction and Domain Adaptation Using Machine Translation,” in ICASSP, 2020.

[7] 阿部雅子 他, “口蓋裂言語にみられる異常構音,” 音声言語医学, 29 巻, pp. 296-298, 1988.

[8] X. Wang *et al.*, “SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation,” in EMNLP, 2018.

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

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

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

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

[13] J. Chorowski *et al.*, “Attention-based models for speech recognition,” NIPS, 2015.

[14] S. Hochreiter *et al.*, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.