器質性構音障害者音声認識のための簡易ラベルによる中間層ロスの導入 *

☆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] などのアプローチを試みてきた．

構音障害者用の音声認識システムを構築するにあたって，澤ら [6] は話者ごとの誤認識の傾向の重要性を指摘した．脳性麻痺患者を評価話者とし，話者に合わせて発話辞書を改変することによって音声認識精度を向上させられる可能性を示した．器質性構音障害者においては，発話器官の異常があるために，話者ごとに大きく異なる固有の発話スタイルを獲得していると考えられる．その結果，発音しにくい音としやすい音の間で，誤認識の頻度も大きく異なると考えられる．本研究では，この話者ごとの誤認識の傾向を誤り傾向と呼称し，音声認識システムの精度を向上させるために活用する．

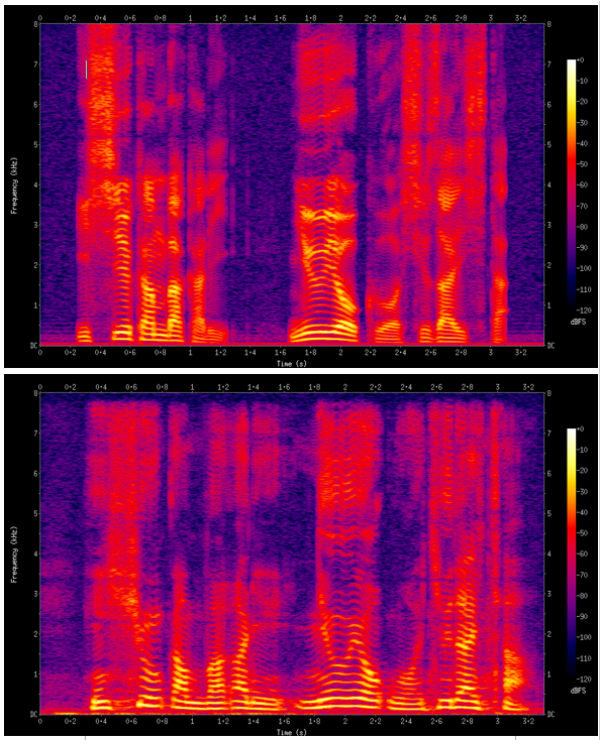

Fig. 1　Example of spectrogram uttered for /i q sh u: k a N b a k a r i ny u: y o: k u o sh u z a i sh i t a/ of a physically unimpaired person (top) and a person with cleft lip and cleft palate (bottom).

野崎ら [7] は，CTC 損失 [8] に基づく音声認識システムの性能を改善するための手法として，モデルの中間層から出力を得て損失を計算する方法を提案した．損失を求めるためにトークン数の次元へ変換された値は，元の次元に戻してから隠れ層の出力と足し合わせて次の層へ送られる．これにより，損失を求めるだけでなく，後に続く層にトークン系列の情報を制約として与えることが出来る．文献 [7] では，比較的大規模な Conformer モデルを用いて健常者音声に対する精度向上が確認されている．本論文では，このような手法を中間層ロスと呼称し，構音障害者音声認識に対する適用を検討する．

* Intermediate loss based on simple labels 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)

中間層ロスを利用する場合，音声認識システムはモデルの最終層と中間層でそれぞれ異なる損失を計算することになる．本研究では，この二つの損失を全く同じ教師ラベルで学習するのではなく，異なる教師ラベルで学習することを検討する．具体的には，最終層においては音声データの収録に利用した台本そのままの一般的な教師ラベルを用いる．一方，中間層においては，話者ごとの誤り傾向を考慮して台本から内容を簡略化させたラベルを用いる．本研究ではこれらをそれぞれ台本ラベル，簡易ラベルと呼称する．このように異なるラベルを用いて学習する場合，音声認識システムは中間層ロスを計算する部分までは簡易ラベルに対して最適化され，それより後の部分は台本ラベルによって最適化される．そのため，音声認識システムは話者ごとの誤り傾向を学びながらも，本来想定されている健常者と同じような認識結果を出力することが可能になると考えられる．

本研究では，中間層ロスにおいて簡易ラベルを用いる有効性を確認するため，音素単位の連続音声認識タスクによる実験を行う。実験においては，まず話者ごとの誤り傾向を確認するために各評価話者の音声で特定話者音声認識モデルを構築する．続いて，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．得られた誤認識パターンに基づいて台本ラベルを簡略化することによって簡易ラベルを作成する．作成した簡易ラベルを用いて，提案法による特定話者音声認識モデルを学習し，簡易ラベルを用いずに学習した場合と精度を比較する．

2　手法

2.1　音素認識モデルによる音素の誤り傾向分析

初めに話者ごとの誤り傾向を分析する．誤り傾向を分析する方法としては，専門家による聴取判定を行うことや，該当する疾患に認められる一般的な傾向に基づいて特定する方法などが考えられる．しかし，このような人手による分析は，音声認識モデルに入力した場合に発生する実際の誤りとは異なる可能性も存在する．そこで，本研究においては人手に依存しない分析を行う．

Fig. 2 は提案手法の概要を示している．まず，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，簡易ラベルを使用しない．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おり，単純に学習を行うだけでは誤りを修正することが不可能であるためだと考えられる．本研究では，そのような健常者との差異を音声認識結果から読み取り，話者の特徴として反映した簡易ラベルを作成する．

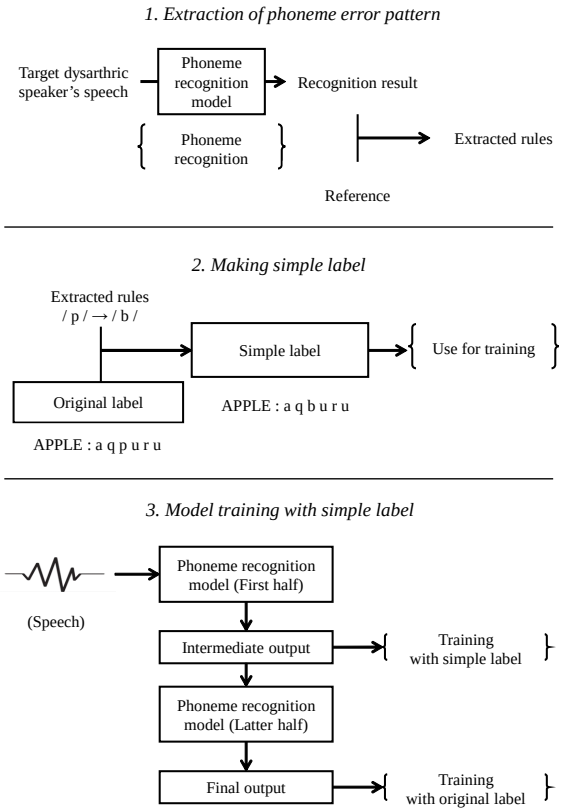

Fig. 2　Overview of our proposed method.

2.2　簡易ラベルの作成

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して台本ラベルから簡易ラベルを作成する．

Table. 1 に簡易ラベルの作成例を示した．例えば，/k/→/t/という誤認識パターンを抽出する場合，台本ラベル (Original script) に記述されている全ての/k/を/t/に置き換える．これにより，簡易ラベル (Simple script) では見かけ上の音素の種類が減少しており，台本ラベルに比べて音素の分類を行うことが簡単になる．このように作成した簡易ラベルは各話者の誤りや

Table 1　An example of change from original label to simple label.

Substitution rules	/k/→/t/, /z/→/g/
Original script	ny u: y o: k u o sh u z a i sh i t a
Simple script	ny u: y o: t u o sh u g a i sh i t a

すい音素の傾向を反映して簡略化される．このため，中間層ロスの計算に簡易ラベルを用いれば，音声認識モデルの学習を安定させるとともに，話者の特徴に関する事前情報を与えることが出来ると考えられる．

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 4 名 (TC1-4), 女性 1 名 (TC5) 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 [9] 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に使用した．

実験では CTC モデルを複数作成した．Fig. 3 に各モデルの構造を示す．モデル A-1 は，話者ごとの誤り傾向を調べるために作成した．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ている．また，パラメータ数による性能の変化を調べるため，同様の構造で 8 層の双方向 GRU を持つモデル A-2 を作成した．さらに，中間層ロスを利用するモデルを二種類作成した．これらはモデル A 二種類と同様の双方向 GRU を 8 層持ち，中間層ロス用の出力層となる全結合層を追加している．簡易ラベルの効果を調べるため，中間層ロスを台本ラベルで利用する場合をモデル B-1，中間層ロスを提案の簡易ラベルで利用する場合をモデル B-2 とする．

全モデル共通の実験条件として，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．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．簡易ラベルを作成する際は，各話者の特定話者音声認識モデルにおける誤認識パターンから，発生率 $Rate_{i\to j}$ が上位の 5 組を抽出した．なお，テスト時の出力は音声認識モデルの最終層から得たものであり，台本ラベルに基づいて精度を評価した．

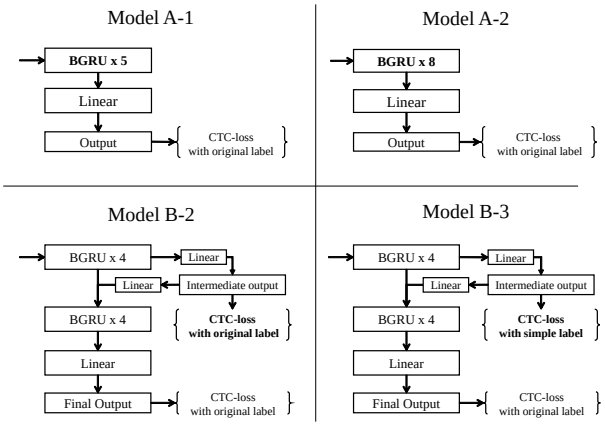

Fig. 3　Overview of each model.

3.2　実験結果

Table 2 は，音素認識タスクにおける各モデルの音素誤り率 (phoneme error rate: PER) を示している．このうち，モデル A-1 とモデル A-2 を比較すると，CLP1-2, TC1-3 についてモデル A-2 の方が悪化していることが分かる．双方向 GRU を増やしたモデル A-2 はパラメータ数が大きく増加しているが，今回の学習データ量に対しては学習が不安定になってしまったと考えられる．

一方，モデル A-2 に中間層ロスの機構を追加したモデル B-1 は，モデル A 二種に比べてほとんどの話者で精度が改善している．先行研究 [7] の手法は，比較的小規模な CTC モデルで構音障害者音声を学習した場合にも有効であることが確認できた．なお，TC3 についてのみモデル B-1 における精度が悪化している．この話者は音声認識制度が最も低いことから，音声の明瞭度が著しく低い話者については，中間層ロスによる制約の追加によって，学習が却って難しくなると考えられる．

また，中間層ロスを台本ラベルではなく提案の簡易ラベルで計算したモデル B-2 については，モデル B-1 に比べて TC2-3 で精度の改善が確認できた．ただし，TC3 に対する精度はモデル A-2 より低い性能に留まっている．

この結果に付随して，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 Table 3 に示す．これらのペアは今回の実験において簡易ラベルへ反映されたものである．簡易ラベルによってモデル B-1 よりモデル B-2 の精度が改善した話者 (TC2-3) に着目すると，各音素の誤認識率が他の話者に比べて低い値になっていることが確認できる．このことから，今回の手法は，偏った誤認識を引き起こすような発音パターンを持つ話者よりも，

Table 2 Phoneme error rates [%] of phoneme-recognition models.

Speaker	Model A-1	Model A-2	Model B-1	Model B-2
CLP1	15.47	15.51	13.95	14.36
CLP2	16.82	16.9	15.7	15.72
TC1	25.32	25.34	24.07	24.12
TC2	19.62	20.05	18.75	18.17
TC3	33.01	32.99	36.76	34.75
TC4	28.61	28.07	26.52	27.11
TC5	14.84	14.74	13.35	13.55

Table 3 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	$i :\to i$ 0.11	$a :\to a$ 0.32	$py \to t$ 0.09	$my \to m$ 0.32	$ry \to y$ 0.18
TC5	$a :\to a$ 0.40	$i :\to i$ 0.29	$e :\to e$ 0.15	$by \to b$ 0.11	$ny \to n$ 0.08

幅広い音素に誤りを引き起こすような発音パターンを持つ話者に対して有効性が高いことが示唆される．

今回作成した簡易ラベルは，誤り率の高かった音素に対して一律に置換を行うことで作成した．この手法では，誤りが発生する可能性が高い部分に限らず，正しく認識できる可能性が高い部分も置換されることになる．音声認識モデルは中間層ロスで簡易ラベルを用いて学習するが，最終層で正しい認識結果を出力するためには，この音素の置換によって生じる簡易ラベルと台本ラベルとの食い違いを修正する必要がある．簡易ラベルに反映された音素の誤認識率が低い話者に関して簡易ラベルの効果が認められたのは，食い違いの修正が比較的簡単に行えたからではないかと考えられる．

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中間層ロスの適用方式を検討した．話者ごとの誤り傾向を学習に取り入れるため，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，台本ラベルを簡略化した簡易ラベルを作成した．簡易ラベルを中間層ロスの計算に利用することで，話者によって精度を改善できる可能性があることを確認できた．今後は更に有効性の高い簡易ラベルを作成するためのアルゴリズムについて検討する予定である．

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] Y. Sawa *et al.*, “Adaptation of a Pronunciation Dictionary for Dysarthric Speech Recognition,” in 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech), pp.612-616, 2022.

[7] J. Nozaki and T. Komatsu, “Relaxing the Conditional Independence Assumption of CTC-based ASR by Conditioning on Intermediate Predictions,” INTERSPEECH, 2021.

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

[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] 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.