吃音者の音声認識における連発ラベル導入による連発箇所の検出*

☆松坂勇樹, 高島遼一 (神戸大), △安井美鈴 (大阪人間科学大), 滝口哲也 (神戸大)

1　はじめに

吃音とは，言葉が滑らかに話せない発話障害の一つである [1]．吃音者の音声特徴自体は健常者のものに近いものの，発話の際に吃音症状が生じることで滑らかな発話が困難となり，日常生活において円滑なコミュニケーションをとることが難しくなる．吃音が生じてしまう明確な原因は現状解明されておらず，吃音を完治する方法は確立されていないが，訓練などによって緩和させることは可能である．例えば，言語聴覚士の指導により吃音症状の緩和を促す方法などが考えられる．

吃音緩和のための訓練を行う際，発話内容における吃音の発生箇所など，分析のために記録することが望ましいが，手動による記録は訓練者にとって負担が大きいため困難である．そこで発話から自動的に吃音箇所を検出することができれば，訓練を円滑に進めることが可能と考えられる．そこで本研究では音声認識を活用して，発話内容の書き起こしに加え，吃音者の発話における吃音箇所の検出を目指す．

吃音の特徴はいくつかあるが，大きな特徴として Fig. 1 のように「連発」と「伸発」と「難発」の 3 つがある．連発は，同じ言葉を繰り返してしまう症状であり，特に話し始める際に起きやすい症状である．伸発は，言葉の一部を引き伸ばしてしまう症状である．難発は，言葉を話す際に，詰まってしまう症状である．本稿では初期検討として，吃音の特徴の中でも「連発」に着目し，音声認識による連発箇所の検出を目指す．

本研究では吃音者の発話における連発箇所を検出するために，2 つの手法を提案する．1 つ目は，吃音者の収録した音声において，連発が発生した箇所に連発ラベルを付与した上で音声認識の学習を行うことによって，連発箇所の検出をする手法である．2 つ目は，1 つ目の手法に加えて，連発箇所の検出精度を高めるために，連発ラベルを含めた健常者事前学習を行う手法である．

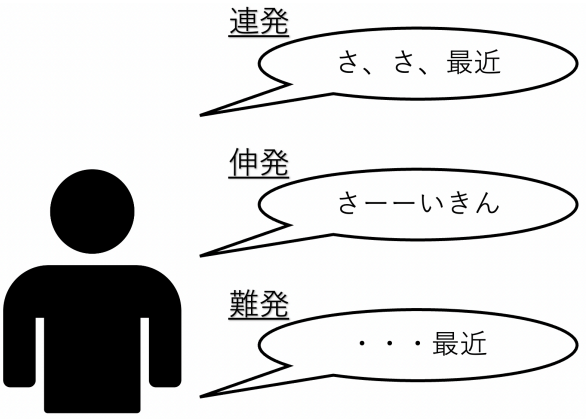

Fig. 1　Characteristics of stuttering.

2　吃音箇所への連発ラベルの付与

本研究では，音声認識によって吃音者の発話における連発箇所を検出するために，連発ラベル「#」を導入する．Fig. 2 に連発ラベルを付与した例を示す．本実験では，実際に吃音者の読み上げ発話を収録し，連発が発生した箇所に連発ラベルを手動で付与した．連発箇所に連発ラベルを付与した上で音声認識の学習を行うことで，実際の発話時における連発箇所を検出することを目指す．実用的なシステムにするためには，連発箇所の検出に加え，その連発箇所の発話内容も表示すべきだが，本研究では第一段階として検出のみに絞って検証を行う．

発話内容：
「ひひ品の良い横顔がささみしそうだった」
連発ラベルの付与：
「＃＃品の良い横顔が＃さみしそうだった」

Fig. 2　An example of the assignment of consecutive labels in the speech of a person who stutters.

3　健常者音声による連発ラベルの事前学習

本研究では，吃音者の発話を収録し，収録音声に対して手動で連発ラベルを付与している．しか

*Detection of consecutive parts of speech by inserting consecutive labels in speech recognition for people who stutters. by Yuki Matsuzaka, Ryoichi Takashima (Kobe University), Misuzu Yasui (Osaka University of Human Sciences), Tetsuya Takiguchi (Kobe University)

し，この方法には，吃音者 1 名の収録音声だけで音声認識モデルを学習することが困難という課題がある．連発箇所を正しく検出するためには，音声認識の学習に使用する吃音者の音声データはより多く必要となるが，1 名の話者から収録音声を大量に入手することは難しく，音声認識モデルの学習には不十分である．

本研究では，上記の課題を改善するために，健常者音声による連発ラベルを含めた事前学習を提案する．吃音者に限らず健常者であっても，即興で話す場合や急いで話す場合などは，吃音者の連発と同様の症状が生じることがある．そこで，本研究では非流暢ラベルを用いた音声認識の先行研究 [2] を参考に，日本語話し言葉コーパス (CSJ) [3] を利用した健常者事前学習を実施する．Fig. 3 に示すように，CSJ のデータには連発に近い特徴を持つ言い淀み箇所が多く含まれており，その部分を連発ラベルに変更することで，CSJ による連発ラベルを含めた事前学習が可能となる．また，CSJ では言い淀みなどの箇所には付加情報のタグが割り当てられているため，手動ではなく自動で連発ラベルに変更できるという利点がある．

発話内容：
「ん形式張ったことは一つもなくてえそそこにあのーお記録されてるは発言を見ますとまーあのそれぞれそうそうたるもう打てば響くような学者」
連発ラベルの付与：
「ん形式張ったことは一つもなくてえ#そこにあのーお記録されてる#発言を見ますとまーあのそれぞれそうそうたるもう打てば響くような学者」

Fig. 3　An example of the assignment of consecutive labels in normal speech (CSJ).

4　評価実験

4.1　データ設定

本研究で評価対象とする話者は吃音症状を持つ日本人女性 1 名である．収録音声として ATR 日本語音声データベース [4] に含まれる音素バランス文 503 文の読み上げ発話を収録しており，そのうち 403 発話を学習データ，50 発話を検証データ，50 発話を評価データとした．ただし，収録した 503 発話の中で連発ラベルは 306 個しか含まれておらず，評価データにおけるサンプルも少量となるため，本研究では 10-fold の交差検証により評価を行う．また，話速変化のデータ拡張である Speed Perturbation [5] も実施しており，速度因数を 0.9, 1.0, 1.1 に設定した．

音声認識モデルを事前学習するための健常者音声として，本研究では 2 種類用意した．一つは前述した日本語話し言葉コーパス (CSJ) である．CSJ の付加情報における言い淀みのタグ (‘D’, ‘D2’) を連発ラベル「#」に変更した．また，比較用に言い淀みが含まれない健常者音声として，JSUT コーパス [6] を用意した．

本研究では音声認識の認識単位として,「かな」と「文字」の 2 つを扱う．より実用的な認識単位である「文字」認識に加え,「かな」認識も行う理由としては,「文字」認識であると通常の音声認識の誤りも比較的多く含まれ，その誤認識が連発箇所の検出に悪影響を及ぼす恐れがあるためである．より簡単な認識タスクである「かな」を行うことで，音声認識の誤りが少ない条件下でも評価を行うことを目的としている．

4.2　モデル設定

音声認識のモデルの学習には Espnet [7] を使用した．入力特徴量は 80 次元の対数メルフィルタバンク特徴量を使用している．

使用した音声認識モデルの構造は「かな」認識と「文字」認識で異なっている．かな認識では，5 層の BLSTM によるエンコーダと，最終層の線形層 1 層で構成された CTC [8] のモデルを使用しており，最適化には AdaDelta を使用した．かな認識は文字認識と比較して簡単なタスクのため，小規模のモデルを使用している．文字認識では，Joint CTC-Attention [9] による Transformer [10] の ASR モデルを使用しており，12 層の Transformer エンコーダと 6 層の Transformer デコーダで構成されている．また，最適化には Adam を使用している．

ファインチューニングに使用した健常者事前学習モデルや，最終的な評価に使用するモデルは検証損失が最小となったエポックのモデルとした．また，本実験において言語モデルは用いていない．

4.3　評価方法

本実験では音声認識の結果に関して，2 つの評価を実施する．

1 つ目は，本研究の目的でもある吃音者の収録音声における連発箇所の検出精度の評価である．検出精度の評価のために，再現率 (Recall)，適合率 (Precision)，F1 スコア (F1-score) の 3 つの指

標を用いる．

2つ目は，吃音者の収録音声の認識精度の評価である．連発ラベルの検出がうまくできても通常の音声認識性能が悪ければ実用的ではないので，合わせて評価を行う．認識精度の評価のために，誤り率を評価指標とする．ただし，評価の際には認識結果および正解スクリプトともに連発ラベルを削除した上で評価を行う．

4.4 実験結果

4.4.1 連発ラベルの検出精度

学習した ASR モデルを用いて吃音者の音声を認識した際の，連発箇所の検出精度の結果を Table. 1 に示す．認識単位が「かな」と「文字」の結果を示しており，それぞれの認識単位において，健常者事前学習の有無や使用した健常者データなどで比較を実施している．また，CSJ データに関しては，連発ラベル (C label; consecutive labels) の有無でも比較している．

まず「かな」認識においては，ベースラインである健常者事前学習を行わない場合においても，それぞれの指標で 70%以上の検出精度を達成していることがわかる。文字よりも比較的容易なタスクであるため，吃音者 1 名のデータのみでもある程度の検出精度を達成できたと考えられる．しかし，連発ラベルを用いずに JSUT や CSJ のデータで健常者事前学習を実施した場合においては，検出精度が全体的に悪化していることがわかる．そこで，連発ラベル (C label) を付与した CSJ データの事前学習を行った場合，ベースラインと比較して検出精度が全体的に向上していることがわかる．この結果より，健常者事前学習における連発ラベルの付与の効果を確認できる．

次に「文字」認識の場合においては，健常者事前学習を実施しない場合は適合率と F1 スコアが非常に悪い結果となった．また，再現率だけ高い値となっているが，これは連発箇所でない発話箇所においても連発箇所と誤認識しているためである．このような悪い検出精度になったのは，学習データが不足しているため，連発箇所に関係なく誤認識している，すなわち音声認識精度が悪いためである．連発ラベルを用いずに JSUT や CSJ による健常者事前学習を行った場合，よりデータ量の多い CSJ では高い検出性能を出したが，比較的データ量の少ない JSUT では低い検出性能となった．そして，連発ラベル (C label) を付与して CSJ データの事前学習を行うことで，検出性能がさらに向上したことが確認できる．この結果より，連発ラベルを付与することで，連発箇所を検出しやすくなることがわかる．

Table 1　Detection accuracy of consecutive labels (C label) in speech recognition.

token	Dataset for pre-training	Recall [%]	Precision [%]	F1 [%]
kana	—	74.3	79.5	76.8
	JSUT	66.0	76.3	70.8
	CSJ	67.0	81.2	73.4
	CSJ(+C label)	**84.5**	**85.9**	**85.2**
char	—	86.5	3.2	6.3
	JSUT	46.9	22.9	30.7
	CSJ	78.5	81.2	79.9
	CSJ(+C label)	**85.1**	**85.1**	**85.1**

最後に，実際の認識結果例を Fig. 4 に示す．図の認識結果は文字認識における連発ラベルを含む CSJ データの学習を行った時の結果例である．連発ラベルを検出できている例もあれば，正しく検出できずに同じ発話を繰り返して出力する誤りも確認できた．

成功例

実際の発話：
「じ自分の実力は自分がいい一番よく知っているはずだ」
認識結果：
「＃自分の実力は自分が＃＃一番よく知っているはずだ」

失敗例

実際の発話：
「夜空をああか赤い灯が点滅しながら…」
認識結果：
「夜空を＃赤赤い火が点滅しながら…」

Fig. 4　Examples of speech recognition results for a person who stutters.

4.4.2 音声認識精度

次に音声認識精度の結果を Table 2 に示す．かな認識においてはかな誤り率，文字認識においては文字誤り率として，トークン誤り率 (TER; Token Error Rate) で評価している．

かな認識の結果では，健常者事前学習を実施することで認識性能が向上するとともに，JSUT よりもデータ量が多い CSJ データを使用することでさらに向上した．また，連発ラベルを含めた CSJ データの事前学習を実施した場合でも，音声認識性能に大きな影響はなかった．文字認識に

Table 2　Speech recognition performance of detection models for consecutive labels.

token	Dataset for pre-training	TER[%]
kana	—	18.1
	JSUT	13.0
	CSJ	5.3
	CSJ (+C label)	**4.7**
char	—	84.1
	JSUT	30.8
	CSJ	12.5
	CSJ (+C label)	**12.4**

関しても同様であるが，大量の健常者音声による事前学習が必須であることがわかる．

4.4.3　健常者事前学習の言い淀み検出精度

本実験では CSJ データの言い淀み箇所に連発ラベルを付与することで，連発ラベルを用いた健常者事前学習を実施したが，事前学習モデルにおける連発ラベル「#」の検出精度を追加で調査する．CSJ の評価データ (eval1 + eval2 + eval3) を用いて評価した際の検出精度を Table 3 に示す．表より，健常者事前学習においてもある程度の検出精度を達成しているが，吃音者データの最も良かった検出結果と比較して精度が悪いことがわかる．連発ラベルを含めた大量のデータで学習したとはいえ，事前学習モデルにおける連発ラベルの検出精度が十分に高い値ではないことが課題となる．

Table 3　Detection accuracy of consecutive labels in pre-training of CSJ data.

token	Recall [%]	Precision [%]	F1 [%]
kana	60.6	66.4	63.4
char	66.8	78.3	72.1

5　おわりに

本研究では，吃音者の発話における連発箇所を，音声認識によって自動検出することを試みた．吃音者の収録音声に連発ラベルを付与して音声認識を学習する方法，そして健常者音声による連発ラベルの事前学習を行うことである程度有効な検出精度が得られた．

本実験では，連発箇所を連発ラベル「#」として認識することで検出したが，より実用的なシステムにするために，連発箇所で何を発話したのかを表示できるようにする必要があり，この点は今後の課題となる．また，吃音の特徴は連発以外にもあるため，それらの特徴も検出できるようにする必要がある．

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] 内須川洸 他, “講座言語障害治療教育 ⟨5⟩ 吃音,” 福村出版, 1982.

[2] 堀井こはる 他, “非流暢ラベルを用いた言い淀み整形 End-to-End 音声認識,” 日本音響学会春季研究発表会講演論文集, pp. 889-892, 2022.

[3] K. Maekawa, “Corpus of spontaneous Japanese: Its design and evaluation,” in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pp. 7-12, 2003.

[4] 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.

[5] T. Ko *et al.*, “Audio augmentation for speech recognition,” in *Interspeech*, pp.3586-3589, 2015.

[6] R. Sonobe, S. Takamichi, H. Saruwatari, “JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” arXiv preprint, 1711.00354, 2017.

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

[8] A. Graves *et al.*, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *ICML*, pp. 369-376, 2006.

[9] S. Watanabe *et al.*, “Hybrid CTC/attention architecture for end-to-end speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, Vol. 11, No. 8, pp. 1240-1253, 2017.

[10] A. Vaswani *et al.*, “Attention is all you need,” in *NeurIPS*, pp. 5998-6008, 2017.