複数データベースを使用した end-to-end 構音障害者音声認識*

☆高島悠樹, 滝口哲也 (神戸大/JST さきがけ), 有木康雄 (神戸大)

1 はじめに

現在，我が国の障害者手帳を持つ人口は 500 万人を超えており，音声・言語・そしゃくに関する障害者の数は 6 万人とされている [1]．障害者差別解消法も施行され，障害者への支援技術の開発がこれまで以上に求められてきている．本研究では，アテトーゼ型の脳性麻痺による構音障害者を対象としている．アテトーゼ型の脳性麻痺では，意図的な動作を行う際に筋肉の不随意運動が発生するため，発話時に筋肉の緊張が起こり正しく構音できない場合がある．そのため，周囲の人とのコミュニケーションに支障をきたす．発話が困難な方でも，手話認識や音声合成システムを使用することでコミュニケーションをとることは可能であるが，脳性麻痺患者の多くは手足が不自由であり，音声に頼るしかない状況が考えられる．そのため，構音障害者のための音声による支援ツールには十分なニーズがあり，研究の必要性があるといえる．

音声認識システムは広く普及し，携帯電話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ている．しかし，これらのシステムは障害のない人を対象としており，言語障害者などの障害者が利用することは難しい．構音障害者の発話スタイルは，筋肉の不随意運動により健常者と大きく異なり，安定した構音が難しく，特に子音は非常に不明瞭になる．Fig. 1 に健常者と構音障害者の発話 “うちあわせ” のスペクトログラムを示す．図に示すように，構音障害者の音声は高周波成分が欠落する傾向があり，健常者のものと大きく異なるため，一般的な音声認識システムは役に立たない．構音障害者を対象とした音声認識システムとして，障害者の特定モデルを用いる手法 [2, 3] が提案されているが，これらの手法は比較的大量の学習データ量を必要とする．しかし，構音障害者は発話による身体への負担が大きいため，大量の音声データを収録することは難しい．障害者音声のデータ拡張として，声質変換が考えられる．相原ら [4] は，partial least squares ベースの声質変換において音素識別的特徴量を用いることで，障害者発話を非障害者発話と変換した．Jiao ら [5] は，敵対的生成ネットワークベースの声質変換により，健常者の音声を障害者のものへ変換することによるデータ拡張を提案した．本研究では，評価対象である日本人構音障害者の音声だけでなく，外国人構音障害者と日本人健常者の音声を用いることによるデータ拡張を提案する．

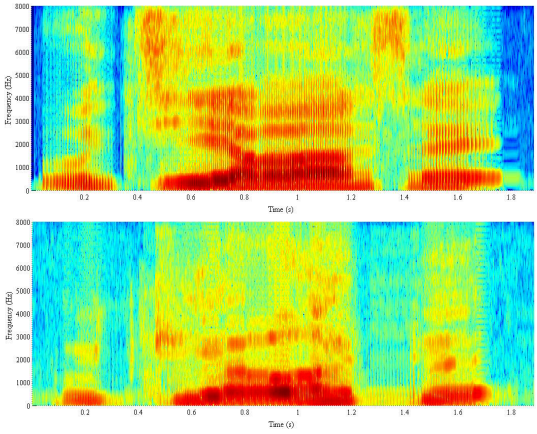

Fig. 1 Example of spectrogram uttered for /u ch i a w a s e/ of a physically unimpaired person (top) and a person with an articulation disorder (bottom)

End-to-end モデルは簡潔な構造で高性能なモデル化が実現できるため，近年盛んに研究が行われている．そこで，本研究でも end-to-end 音声認識モデルを利用する．本稿では，end-to-end 音声認識モデルとして，listen, attend and spell (LAS) モデル [6] を使用する．LAS モデルは，listener モジュールと speller モジュールで構成されている．Listener モジュールは，音声特徴量系列を中間表現へと変換する encoder であり，speller モジュールはそれらから入力系列に対応する音素の生成確率系列を出力する decoder である．Listener モジュールが音響モデル，speller モジュールが言語モデルに相当する．また，マルチリンガル音声認識システムは言語間での知識転移により性能を向上させ，各言語の学習データ量を減らすことができることが示唆されており [7]，LAS モデルを用いたマルチリンガル音声認識システムにおいても，その有効性が示されている [8]．ここで，以下の 2 つを仮定する．1 つは，構音障害者は発話スタイルが異なるため，障害者専用の音響モデルが必要である．2 つ目は，障害の有無に関わらず，言語が同じであれば言語モデルは共有できる．これらの仮定より，提案手法では，構音障害を持つ日本人話者と外国人話者で listener モジュールを共有し，日本人障害者と健常者で speller モジュールを共有する構造を持つ LAS モデルを構築する．これにより，日本人構音障害者固有のモジュールを持たず，データ量を抑えることができると期待できる．

*End-to-End dysarthric speech recognition using multiple databases, by Yuki Takashima, Tetsuya Takiguchi (Kobe University/JST PRESTO), Yasuo Ariki (Kobe University)

2 Listen, attend and spell model

LAS モデル [6] は，encoder-decoder モデルであり，encoder，decoder はそれぞれ，listener，speller と呼ばれる．音響特徴量系列が与えられた時，LAS モデルは音素系列の生成確率を以下のように計算する．

$$P(\boldsymbol{y}|\boldsymbol{x}) = \prod_s P(\boldsymbol{y}_s|\boldsymbol{x}, \boldsymbol{y}_{<s}), \tag{1}$$

ここで，$\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_t, ..., \boldsymbol{x}_T)$ と $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_s, ..., \boldsymbol{y}_S)$ はそれぞれ，音響特徴量と音素の系列を表す．$\boldsymbol{x}_t$，$\boldsymbol{y}_s$，T，S はそれぞれ，t フレームにおける音響特徴量，s 番目の出力音素，入力特徴量のフレーム数，出力音素の数を表す．Listener は encoder-recurrent neural network (RNN) であり，入力音響特徴量系列 $\boldsymbol{x}$ を中間表現 $\boldsymbol{h} = (\boldsymbol{h}_1, ..., \boldsymbol{h}_u, ..., \boldsymbol{h}_U)$ へ変換する．ここで，$\boldsymbol{h}_u$，$U \leq T$ はそれぞれ，u 番目の listener 出力，listener の出力系列数を表す．Speller は decoder-RNN であり，$\boldsymbol{h}$ を入力とし，音素系列の生成確率 $\boldsymbol{y}$ を推定する．

Listener は pyramid bidirectional long short term memory (pBLSTM) により構成される．ピラミッド構造は，計算の複雑さと収束時間を軽減させ，より短い時間ステップから重要な情報を抽出することを可能にする．Listener による出力は以下の式で表現される．

$$\boldsymbol{h} = \mathrm{Listen}(\boldsymbol{x}; \theta_{\mathrm{Lis}}), \tag{2}$$

ここで，θ_{Lis} は listener のパラメータである．

Speller は注意機構を用いた単方向 LSTM により構成される．各時刻において，speller はそれまでの出力に条件付けられた当該時刻の出力分布を生成する．さらに，注意機構により，入力音響特徴量系列の情報を効果的に集約しながら出力を生成することができる．Speller による出力は以下の式で表現される．

$$P(\boldsymbol{y}|\boldsymbol{x}) = P(\boldsymbol{y}|\boldsymbol{h}; \theta_{Spl}) = \prod_s P(\boldsymbol{y}_s|\boldsymbol{h}, \boldsymbol{y}_{<s}; \theta_{Spl}) \tag{3}$$

$$= \mathrm{Spell}(\boldsymbol{h}; \theta_{\mathrm{Spl}}), \tag{4}$$

ここで，θ_{Spl} は speller のパラメータである．

LAS モデルは以下の識別損失を最適化するように学習される．

$$\mathcal{L}_{LAS}(\mathcal{D}, \theta_{Lis}, \theta_{Spl}) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{D}}[-\log(P(\boldsymbol{y}|\boldsymbol{x}))], \tag{5}$$

ここで，$\mathcal{D}$ は入力系列及び対応する音素系列の集合である．

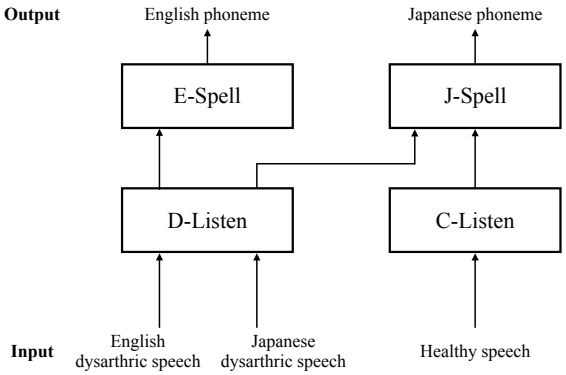

Fig. 2 Overview of our proposed method. Japanese dysarthric speech is uttered by one target speaker in this paper.

3 提案手法

Fig. 2 に提案モデルの概要を示す．提案モデルは LAS モデルをベースとしており，2 つの listener と 2 つの speller により構成される．本研究では，日本人構音障害者を評価対象としており，以降，目標話者とはこれを参照する．

3.1 モデリング

提案モデルは，障害者用 listener “D-Listen” と健常者用 listener “C-Listen” を持つ．D-Listen は構音障害を持つ日本語話者 (“日本語障害者”) と英語話者 (“英語障害者”) で共有される．これにより，目標話者である日本語障害者の学習データが少量であっても，英語障害者の音声データを大量に用意することで，十分に学習された listener を得ることができる．また，マルチリンガル学習により listener がより良い特徴表現を獲得することが期待できる．C-Listen は構音障害のない日本語話者 (“日本人健常者”) の音声のみを入力として扱う．

さらに，提案モデルは英語 speller “E-Spell” と日本語 speller “J-Spell” を持つ．これらは listener モジュールの出力を入力とし，それぞれ，英語と日本語の音素の事後確率を生成する．日本語障害者の音声を入力とした D-Listen の出力，及び日本語健常者の音声を入力とした C-Listen の出力は，J-Spell を通して音素の生成確率系列へ変換される．言語モデルは障害の有無に関わらず日本語発話で共有できるため，健常者の音声を用いることで，十分に学習された speller を獲得することができると期待できる．

3.2 学習

学習は音響特徴量系列が入力されたときに，その正解音素列の生成確率を最大化するようにモデルパラメータの推定を行う．日本語障害者のデータセット

として，$\mathcal{D}_{JD}$ を入力系列及び対応する音素系列の集合とする．$\mathcal{D}_{ED}$ と $\mathcal{D}_{JC}$ も同様に，英語障害者と日本語健常者のデータセットとしてそれぞれ定義する．提案モデルは以下の目的関数を最小化するように最適化される．

$$\mathcal{L}_{LAS}(\mathcal{D}_{ED}, \theta_{D\text{-}Lis}, \theta_{E\text{-}Spl}) + \mathcal{L}_{LAS}(\mathcal{D}_{JD}, \theta_{D\text{-}Lis}, \theta_{J\text{-}Spl}) + \mathcal{L}_{LAS}(\mathcal{D}_{JC}, \theta_{C\text{-}Lis}, \theta_{J\text{-}Spl}), \quad (6)$$

ここで，$\theta_{D\text{-}Lis}$, $\theta_{C\text{-}Lis}$, $\theta_{E\text{-}Spl}$, $\theta_{J\text{-}Spl}$ はそれぞれ，D-Listen, C-Listen, E-Spell, J-Spell のパラメータである．このモデルにおいて，全てのモジュールが同時に最適化される．

4 評価実験

4.1 実験条件

提案手法は音素認識タスクで評価を行った．評価話者として，構音障害者男性 2 名の音声を使用した．構音障害者音声は，ATR 研究用日本語音声データベース [9] に含まれる 216 単語を収録した．1 人目の障害者 (“Dysarthric speaker1”) は 216 単語の 5 回連続発話を収録した．2 人目の障害者 (“Dysarthric speaker2”) は 216 単語中 6 単語を発話できなかったため，残りの 210 単語の 5 回連続発話を収録した．1 発話目を評価，残りの 4 発話分を学習に使用した．健常者音声は，ATR 研究用日本語音声データベースから男性 1 名を選択し，5,240 単語を学習，構音障害者音声と同一の 216 単語を評価に使用した．英語障害者音声として，TORGO データベース [10] を使用した．女性 3 名，男性 4 名の音声から，2,726 単語を選択した．各データベースの学習データのうち，95%をモデルの学習セットとし，残りを開発セットとして早期終了のために使用した．音響特徴量は，フレームシフト 10ms，窓幅 25ms で抽出された 13 次元の MFCC に Δ，ΔΔ 成分を加えた 39 次元のベクトルを用いた．出力音素数は E-Spell が 59，J-Spell が 56 (始端・終端記号を含む) とした．認識性能の評価には文字誤り率 (CER: character error rate) を用いた．

ベースラインシステムとして，従来の LAS モデルに基づく 2 つのモデルを調査した．1 つ目は日本語健常者の音声のみを用いて学習された LAS モデル，2 つ目は日本語障害者の音声も加えて学習したモデルである．Listener モジュールには，BLSTM 層の後に 512 ユニットの 2 層 pBLSTM 層を加えたものを用いた．Speller モジュールには，注意機構を持つ 512 ユニットの LSTM 層の後に，出力層としてソフトマックス層を導入したモデルを用いた．バッチサイズ 64 とし，ラベルスムージングを行い，最適化には Adam を用いた．学習率は 1e-4 とした．

4.2 実験結果

最初に，ベースラインモデルの評価結果について述べる．Table 1 に，日本語健常者の音声のみを用いて学習したモデルの認識結果を示す．健常者に比べて，障害者の認識率は著しく劣化することが見て取れる．これは，障害者の発話スタイルが健常者と大きく異なるためであると考えられる．Table 2 に，日本語健常者と日本語障害者の両方の音声を用いて学習したモデルの認識結果を示す．障害者のデータも用いることで，Table 1 に比べて，より低い CER が得られた．障害者の音声も考慮した学習が行われたためだと考えられる．

Table 1 CER (%) of a model trained on speech data of a physically unimpaired person only.

Test speaker	Top-1	Top-3
Controlled speaker	12.2	9.2
Dysarthric speaker1	72.2	69.2
Dysarthric speaker2	76.3	75.3

Table 2 CER (%) of a model trained on joint speech data of a physically unimpaired person and a person with an articulation disorder.

Test speaker	Top-1	Top-3
Controlled speaker	11.2	7.2
Dysarthric speaker1	27.2	22.2
Controlled speaker	11.2	9.2
Dysarthric speaker2	32.1	27.0

次に，提案手法の評価結果について述べる．Table 3 に示すように，提案手法はベースラインモデルよりも高い性能が得られた．提案モデルは，Table 2 で評価したモデルと同じだけの日本語障害者の学習データ量にも関わらず精度が向上しており，これは提案モデルが英語障害者のデータを用いることで，listener モジュールがより優れた特徴空間を学習できたことを示す．先行研究 [8] と同様に，LAS モデルを用いたフレームワークは障害者音声に対しても有効であることが示された．

Table 3 CER (%) of our proposed method.

Target speaker	Top-1	Top-3
Dysarthric speaker1	23.2	19.2
Dysarthric speaker2	30.0	24.9

Table 4 Example of ASR results for dysarthric speaker1 predicted from our proposed method. “pau” indicates the pause.

Ground Truth	pau u r a y a- m a sh ii pau
Predicted	pau u m a a a a s ii pau
Ground Truth	pau d a ky ou pau
Predicted	pau d a k ou pau

Table 4 に 1 人目の障害者の認識結果の例を示す．表より，子音が欠落しやすい傾向が確認できる．構音障害者音声は，筋肉の不随意運動により子音を発音するのが難しく，音声が聞き取りづらくなるという特徴がある．母音は正しく認識できており，モデルがこのような障害者音声の特徴を捉えていると考えられる．得られた誤りの傾向を分析することで、その話者が発話しづらい音の発見に繋がると期待される。さらに、誤り訂正技術を応用することで，より認識精度を向上させられると考えられる．

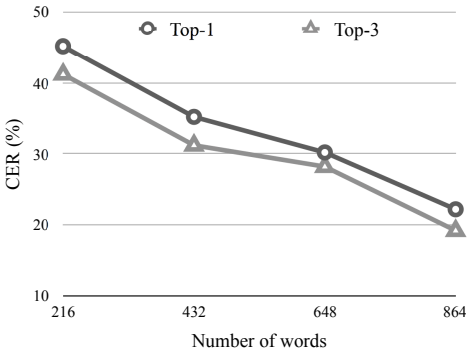

Fig. 3 The correlation between error rates and the number of training words.

最後に，学習データ量に対する提案手法の性能の評価を行った．1 人目の障害者を選択し，学習に使用するデータ量を各単語 1 回発話分から 4 回発話分まで変化させた．英語障害者及び日本語健常者の学習データ量は変化させない．開発セットは用意せず，エポック数 50 としてモデルを学習した．Fig. 3 に示すように，評価話者の学習データ量が減ると認識性能も劣化することが分かる．日本語 speller は，日本語障害者と日本語健常者のデータを入力として受ける．日本語障害者のデータ量が減少すると，相対的に日本語健常者のデータ量が大きくなり，日本語 speller が日本語健常者のデータにより注力して学習するため精度が劣化したと考えられる．また，日本語 speller は 2 つのドメインからの入力を扱うため，学習が難しくなっていると考えられる．これは，ドメイン適応などの技術を適用することで軽減が期待できる．

5 おわりに

本稿では，LAS モデルに基づく，構音障害者のための end-to-end 音声認識システムを提案した．構音障害を持つ英語話者，日本人健常者の音声を用いたデータ拡張を行い，単純に日本語健常者と日本語障害者のデータを結合して学習した場合と比べて精度が向上することを示した．

謝辞 本研究の一部は，JSPS 科研費 JP17J04380，JST さきがけ JPMJPR15D2 の助成を受けたものである．

参考文献

[1] 厚生労働省, “平成 29 年度厚生統計要覧,” 2017.
[2] T. Nakashika *et al.*, “Dysarthric speech recognition using a convolutive bottleneck network,” in *International Conference on Signal Processing*. IEEE, pp. 505–509, 2014.
[3] Y. Takashima *et al.*, “Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition,” in *EUSIPCO*, pp. 1411–1415, 2015.
[4] R. Aihara *et al.*, “Phoneme-discriminative features for dysarthric speech conversion,” in *INTERSPEECH*, pp. 3374–3378, 2017.
[5] Y. Jiao *et al.*, “Simulating dysarthric speech for training data augmentation in clinical speech applications,” in *ICASSP*, pp. 6009–6013, 2018.
[6] W. Chan *et al.*, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *ICASSP*, pp. 4960–4964, 2016.
[7] L. Besacier *et al.*, “Automatic speech recognition for under-resourced languages: A survey,” Speech Communication, vol. 56, pp. 85–100, 2014.
[8] S. Toshniwal *et al.*, “Multilingual speech recognition with a single end-to-end model,” in *ICASSP*, pp. 4904–4908, 2018.
[9] 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.
[10] F. Rudzicz *et al.*, “The TORGO database of acoustic and articulatory speech from speakers with dysarthria,” Language Resources and Evaluation, vol. 46, no. 4, pp. 523–541, 2012.