社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

[ポスター講演] End-to-end 構音障害者音声認識のための複数データベースを用いたデータ拡張

高島　悠樹　　滝口　哲也　　有木　康雄

神戸大学大学院システム情報学研究科　〒657–8501 兵庫県神戸市灘区六甲台町 1–1
E-mail: takashima@stu.kobe-u.ac.jp, {takigu,ariki}@kobe-u.ac.jp

あらまし　本報告では，アテトーゼ型脳性麻痺による日本人構音障害者の音声認識を行う．意図的な動作時や緊張状態にある場合に筋肉の不随意運動が生じるため，彼らの発話は不安定となる．そのため，一般的な音声認識システムでは彼らの発話を認識することは困難となる．近年，深層学習の発展により様々な技術が提案されているが，そのほとんどは大量の学習データを必要とする．しかし，アテトーゼ型脳性麻痺による構音障害者は発話による身体への負担が大きいために，大量の学習データを用意することは困難である．そのため，限られた学習データ量で構築できるシステムが求められる．本研究では，構音障害を持つ英語話者と日本人健常者の音声を用いたデータ拡張を提案する．また，本稿では，listen, attend and spell (LAS) モデルを用いた end-to-end 音声認識を行う．LAS モデルは enocder-decoder モデルであり，encoder が音響モデル，decoder が言語モデルに相当する．音響モデルモジュールを構音障害を持つ英語話者と，言語モデルモジュールを日本人健常者と共有することにより，対象とする日本人構音障害者固有のパラメータを抑えたモデルを提案する．単語認識実験により，提案手法の有効性を確認した．
キーワード　音声認識，マルチリンガル，障害者支援，end-to-end モデル，構音障害

Data Augmentation Using Multiple Databases for End-to-end Dysarthric Speech Recognition

Yuki TAKASHIMA, Tetsuya TAKIGUCHI, and Yasuo ARIKI

Graduate School of System Informatics, Kobe University　1–1 Rokkodai, Nada, Kobe, 657–8501 Japan
E-mail: takashima@stu.kobe-u.ac.jp, {takigu,ariki}@kobe-u.ac.jp

Abstract　We present in this paper an end-to-end speech recognition system for a Japanese person with an articulation disorder resulting from athetoid cerebral palsy. The movements of such speakers are limited by their athetoid symptoms, and their utterances are often unstable or unclear, which makes it difficult for them to communicate. Therefore, the performance of automatic speech recognition (ASR) systems for people with an articulation disorder degrades significantly. Recently, deep learning approaches to speech recognition have seen much progress. These techniques require a large amount of training data, however, the amount of data from people with articulation disorder is limited due to their athetoid symptoms. This paper proposes a data augmentation method using not only the speech data of a Japanese person with an articulation disorder but also the speech data of a physically unimpaired Japanese person and a non-Japanese person with an articulation disorder. We employ an end-to-end ASR model based on the listen, attend and spell (LAS) model which has an acoustic module and a language module. In our proposed model, the acoustic module is shared between people with dysarthria, and a language module is assigned to each language regardless of dysarthria. The effectiveness of this approach was confirmed through word-recognition experiments where our proposed method outperformed a method based on the conventional LAS model.
Key words　Speech recognition, multilingual, assistive technology, end-to-end model, dysarthria

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．

脳性麻痺とは，筋肉の動きをつかさどる脳の部分が受けた損傷が原因で筋肉の制御ができなくなり，けいれんや麻痺，ほかの神経障害が起こる症状のことである．それらの原因は多様であり，出生前・出生時・出生直後の脳への酸素供給，出生前の胎内感染，妊娠中毒症，分娩時の外傷，仮死状態，未熟出生，出生後の脳を覆う組織の炎症や外傷性損傷などがあげられている．また，脳性麻痺は脳の損傷部分によって痙直型 (大脳皮質)，アテトーゼ型 (中脳もしくは脳基底核)，失調型 (小脳)，混合型 (脳の広範囲) に分類される [2]．それぞれ，痙直型は正常な筋の伸張反射が過度になる，アテトーゼ型はアテトー ゼと呼ばれる筋肉の不随意運動を伴う，失調型は協調運動の障害が現れ，混合型はそれぞれの症状が混合して現れるというような症状が見られる．

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．

文献 [3] において我々は，構音障害者の音声認識率を改善するため，multiple acoustic frames を提案した．また，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 [4], [5] を提案してきたが，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手法として，声質変換が考えられる．相原ら [6], [7] 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へと変換した．Jiao ら [8] 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．これらの手法は，評価話者の音声のみ，あるいは単一コーパスのデータ (単一言語) のみを用いていた．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マルチリンガルデータ拡張を提案する．

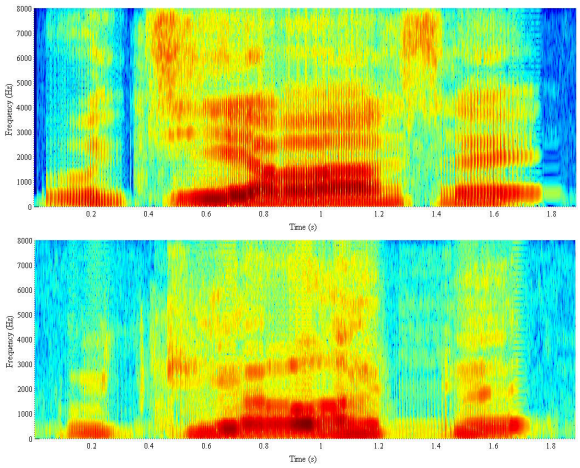

図 1　発話/u ch i a w a s e/のスペクトル．上：健常者．下：構音障害者．

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 [9], [10]．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 [11] 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，音響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 [12]，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 [13]．

我々の先行研究において，構音障害者は発話スタイルは健常者のものと異なるため，障害者専用の音響モデルを必要としてきた．しかし，障害者専用のモデルの構築には大量の学習データを必要とするため，障害のレベルが重く極めて少量の学習データしか用意できないような話者についてはモデルの構築が困難となる問題点がある．そこで本研究では，評価話者と異なる言語の音声の利用を試みる．具体的に，構音障害を持つ英語話者に関しては，いくつかの利用可能なコーパスが公開されている [14]〜[16]．本稿では，このうち TORGO コーパス [16] を利用する．言語が異なっていても，構音障害者であれば音響的

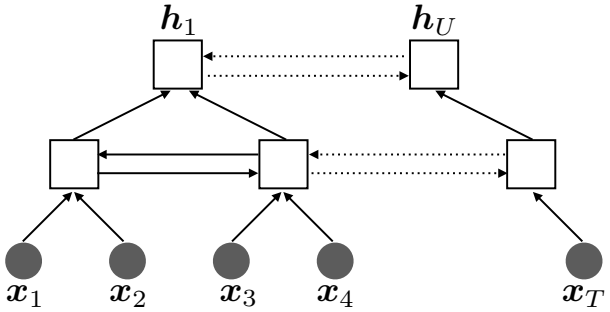

図 2　Listener のネットワーク構造.
Fig. 2　Network structure of the listener.

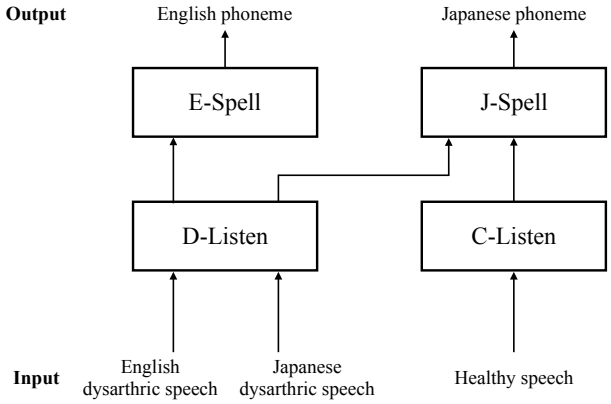

図 3　提案モデルの概要.
Fig. 3　Overview of our proposed model.

な特徴は同じであると仮定し，LAS モデルの音響モデル部を構音障害を持つ英語話者と日本語話者で共有しながら学習する．また，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と仮定し，LAS モデルの言語モデル部を日本人健常者と日本人障害者で共有しながら学習する．これらのモデル共有メカニズムにより，評価話者固有のパラメータを持たず，さらに学習データ量も増加させることができ，評価話者である日本人障害者のデータ量を抑えることが期待できる．

以下，第 2 章で LAS モデルの概要を述べ，第 3 章で提案手法について説明する．第 4 章で評価実験を行い，第 5 章で本稿をまとめる．

2. Listen, attend and spell model

LAS モデル [11] 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 Fig. 2 に示すように，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 $\boldsymbol{y}_{<s}$ 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．

3. 提 案 手 法

Fig. 3 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．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るため，障害者専用の音響モデルが必要だと考えられる．LAS モデルにおいて，listener モジュールは音響モデルに相当するため，これを 2 つ用意し，それぞれ構音障害者用データベース，日本語健常者用データベースを用いて学習する．そして，D-Listen を，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”) で共有する．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，英語と日本語の 2 つの言語を扱う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 は健常者専用音響モデルであり，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と考えられるため，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．E-Spell は，英語障害者の音声を入力とした D-Listen の出力を，英語音素の生成確率系列へ変換する．D-Listen は言語の違いに関わらず共通の潜在特徴空間を持ち，E-Spell 及び J-Spell はそこから英語及び日本語の音素系列を生成できると仮定する．

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と単語音声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 5 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 [17] に含まれる 216 単語を複数回発話したものを収録した．Table 1 に話者とデータ量の関係を示す．発話単語数及び繰り返し回数は，障害の程度により話者ごとに異なる．1 発話目を評価，残りの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 (“MAU”) 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 [16] 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単語認識率と音素誤り率 (PER; phoneme error rate) を用いた．辞書サイズを 216 単語とし，推定された音素系列と最も PER が小さくなる単語を認識単語とし認識率を算出した．

表 1　日本人構音障害者のデータ量
Table 1　Dataset statistics of Japanese people with the articulation disorder.

Speaker	# words	# repetitions	# utterances
MK01	204	3	612
MM03	210	5	1,050
MK04	216	5	1,080
MY05	215	3	645
MN06	213	3	639

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 1 つの LAS のみを用いるモデル，2 つ目は，1 つ目とモデルは同じだが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，512 ユニットの 2 層 pBLSTM 層を用いた．Speller モジュールには，注意機構を持つ 512 ユニットの単方向 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 [18] を用いた．学習率は 1e-4 とした．

また，単語認識実験の従来法として，Gaussian mixture model-hidden Markov model (GMM-HMM) による認識結果と比較した．音響特徴量は上述のものを用いた．音響モデルは，54 音素の monophone-HMM で，各 HMM の状態数は 5，状態あたりの混合分布数は 6 である．

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．PER はより低い方，単語認識率はより高い方が良い精度を表す．Table 2 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，障害者の認識精度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 3 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．1 つのモデルの学習に用いる日本語構音障害者音声は 1 名分とした．目標話者のデータも用いることで，Table 2 に比べて，性能が向上することがわかる．これは，障害者音声の特徴も考慮した学習が行われたためだと考えられる．話者 MM03 と MK04 は他の話者に比べて高い性能が得られた．この 2 話者は学習データとして 4 回発話分の音声

表 2　日本人健常者の音声のみを用いて学習されたモデルの認識結果
Table 2　PER (%) and word recognition accuracy (%) of a model trained on speech data of a physically unimpaired person only.

Test speaker	PER	Word recognition accuracy
MAU	6.65	96.30
MK01	103.81	1.96
MM03	105.79	0.48
MK04	91.13	2.31
MY05	80.02	3.26
MN06	99.55	0.94

表 3　日本人健常者と日本人構音障害者の音声を用いて学習されたモデルの認識結果

Table 3　PER (%) and word recognition accuracy (%) of a model trained on joint speech data of a physically unimpaired person and a person with an articulation disorder.

Test speaker	PER	Word recognition accuracy
MAU	11.70	93.98
MK01	43.33	48.53
MAU	11.06	91.67
MM03	32.42	63.33
MAU	11.64	90.28
MK04	25.37	72.69
MAU	9.43	93.06
MY05	46.26	43.72
MAU	9.65	93.06
MN06	48.79	31.92

を使用しており，他の話者と比べて学習データ量が多いからだと考えられる．

次に，提案手法の評価結果について述べる．Fig. 4 に単語認識結果を示す．“Joint” は Table 3 に示した，日本語健常者と日本語障害者の音声を両方を用いて LAS モデルを学習した結果を表す．図に示すように，提案手法 “Prop” は単純に学習データを両方用いた場合よりも高い性能が得られた．提案モデルは，Table 3 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，これは提案モデルにおいて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 [13] と同様に，LAS モデルを用いたマルチモーダル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．

しかし，従来の GMM-HMM 音声認識結果と比べて，平均で 13.76% の精度劣化が見られる．提案モデルの学習に使用したデータ量は，ドメインごとにばらつきが大きく学習が偏ることが考えられる．そこで，英語障害者と日本語健常者の損失を重み付けすることで，目標話者である日本語障害者に対してより焦点を当てた学習を試みる．“Prop (α, β)” は，英語障害者の損失に α，日本語健常者の損失に β の重みをかけることを表す．図より，適切な重みをつけ学習を制御することで，モデルの性能をより向上させることが分かる．重み付け損失関数の利用により精度は向上したが，依然として従来の GMM-HMM と比べて平均で 8.86% 低い認識率となっている．この原因として，従来手法のように言語的な制約がないためだと考えられる．例えば，提案手法では子音が連続して推定される状況が起こりうる．

Table 4 に話者 MK04 の推定結果の例を示す．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．

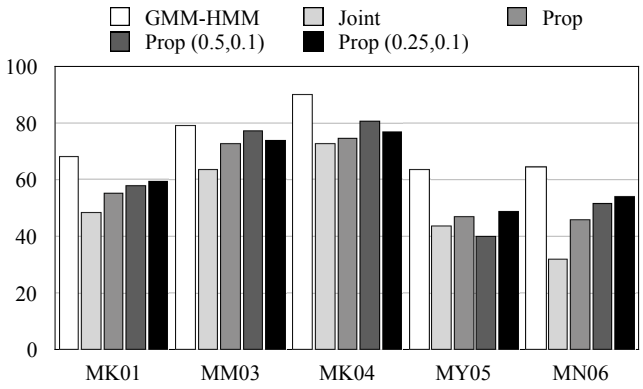

図 4　各手法に対する単語認識実験結果．

Fig. 4　Word recognition accuracy (%) for each method.

表 4　話者 MK04 の発話に対する提案手法の推定結果の例

Table 4　Example of ASR results for MK04 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

提案モデルにおいて，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少ないと，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性能が劣化し，従来の GMM-HMM と比べて，認識精度が低下したと考えられる．また，日本語 speller はこのように 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．画像処理の分野において，ドメイン適応が盛んに研究されている．これは，学習データと評価データのドメインが異なることによる精度劣化を軽減する技術であり，様々な手法が提案され，その有効性が示されている．本手法においても，ドメイン適応技術を適用することでこの軽減が期待できる．

5. おわりに

本稿では，アテトーゼ型脳性麻痺による日本人構音障害者のための，LAS モデルに基づく end-to-end 音声認識システムを提案した．構音障害者の発話スタイルは健常者と大きく異なるため，健常者音声を用いて学習された音声認識システムは役に立たない．また，筋肉の不随意運動により，発話による身体への負担が大きいため，大量の音声を収録することは困難である．そこで，本稿では，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提案した．他のデータベースのデータとパラメータを共有することで，日本人障害者固有のパラメータを持たず，少量データによるモデル化を実現した．また，LAS モデルによる end-to-end 構音障害者音声認識システムを提案した．単語認識実験によって，提案モデルは，単純に日本語健常者と日本語障害者のデータを両方用いて LAS モデルを学習した場合と比べて，精度が向上することを示した．

提案手法は，言語的な制約を含む従来の GMM-HMM 音声認識と比べて，やや劣る結果となった．この原因の 1 つとして，学習データ量のばらつきが考えられる．本稿では損失関数に重み付けを行ったが，話者ごとに適切な重みが異なり，調整が難しいという問題点があった．今後は，学習データのばらつきを考慮できる学習方法を検討する．提案モデルでは，日本語 speller は，障害者 listener と健常者 listener の 2 つのモジュールからの入力を受けており，このことも学習を難しくしている要因の 1 つだと考えられる．今後，提案モデルにもドメイン適応を組み込み，さらなる性能の向上を試みる．

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．

文　献

[1] 厚生労働省，“平成 29 年度厚生統計要覧，” 2017. https://www.mhlw.go.jp/toukei/youran/index-kousei.html
[2] S.T. Canale and W.C. Campbell, “Campbell’s operative orthopaedics,” Technical report, Mosby-Year Book, 2002.
[3] C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, and I. Li, “Multimodal speech recognition of a person with articulation disorders using aam and maf,” IEEE International Workshop on Multimedia Signal Processing, pp.517–520, 2010.
[4] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, “Dysarthric speech recognition using a convolutive bottleneck network,” International Conference on Signal ProcessingIEEE, pp.505–509 2014.
[5] Y. Takashima, T. Nakashika, T. Takiguchi, and Y. Ariki, “Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition,” EUSIPCO, pp.1411–1415, 2015.
[6] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, “A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary,” EURASIP J. Audio, Speech and Music Processing, vol.2014, p.5, 2014.
[7] R. Aihara, T. Takiguchi, and Y. Ariki, “Phoneme-discriminative features for dysarthric speech conversion,” Interspeech, pp.3374–3378, 2017.
[8] Y. Jiao, M. Tu, V. Berisha, and J. Liss, “Simulating dysarthric speech for training data augmentation in clinical speech applications,” ICASSP, pp.6009–6013, 2018.
[9] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” ICML, vol.32, pp.1764–1772, JMLR.org, 2014.
[10] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” NIPS, pp.577–585, 2015.
[11] W. Chan, N. Jaitly, Q.V. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” ICASSP, pp.4960–4964, 2016.
[12] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, “Automatic speech recognition for under-resourced languages: A survey,” Speech Communication, vol.56, pp.85–100, 2014.
[13] S. Toshniwal, T.N. Sainath, R.J. Weiss, B. Li, P.J. Moreno, E. Weinstein, and K. Rao, “Multilingual speech recognition with a single end-to-end model,” ICASSP, pp.4904–4908, 2018.
[14] X. Menéndez-Pidal, J.B. Polikoff, S.M. Peters, J.E. Leonzio, and H.T. Bunnell, “The nemours database of dysarthric speech,” International Conference on Spoken Language Processing, pp.1962–1965, 1996.
[15] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T.S. Huang, K. Watkin, and S. Frame, “Dysarthric speech database for universal access research,” Interspeech, pp.1741–1744, 2008.
[16] F. Rudzicz, A.K. Namasivayam, and T. Wolff, “The TORGO database of acoustic and articulatory speech from speakers with dysarthria,” Language Resources and Evaluation, vol.46, no.4, pp.523–541, 2012.
[17] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol.9, no.4, pp.357–363, 1990.
[18] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.