Hybrid CTC/attention モデルを用いた構音障害者音声認識の検討 *

☆澤佑哉，高島遼一，滝口哲也，有木康雄 (神戸大)

1　はじめに

本研究ではアテトーゼ型の脳性麻痺患者の音声認識を検討する．アテトーゼ型脳性麻痺患者は，意図した動作時に筋肉の不随意運動を伴い，発話を行う際に筋肉の緊張により正しく構音ができない場合がある．また，脳性麻痺患者の多くは手足の動作が不自由であり手話や筆談といった代替手段が取れず，音声によるコミュニケーションに頼らざるを得ない場面も多い．そのため，構音障害者の音声認識には高いニーズがあり，研究の必要性があると言える．

近年，音声認識技術の発展に伴って，様々な環境下での音声認識システムの利用が期待されている．従来の音声認識システムの多くは健常者の発話を対象としており，構音障害をはじめとする言語障害者を対象としたものは少ない．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の発話とは大きく異なり，安定した発話が困難である．例として，Fig. 1 に健常者と構音障害者の発話のスペクトログラムを示す．構音障害者の発話は高周波成分の欠落や，子音の欠落，間延びが発生するなどの特徴がある．そのため，従来の DNN-HMM ハイブリッドモデルを用いた場合では，学習に必要なアライメント情報を得ることが困難であるという課題がある．そこで本研究では，フレーム単位のアライメント情報を必要としない End-to-End 音声認識モデルを用いて構音障害者の音声認識を行う．End-to-End 音声認識モデルはニューラルネットワークを用いて音響特徴量から音素や文字列などの記号列に直接変換するモデルであり，簡潔な構造で高精度なモデル化を実現できる．これにより，構音障害者に特有な発話の特徴がニューラルネットワークの枠組みの中で自動的に学習されることが期待できる．

一般的に End-to-End モデルの学習には大量のデータが必要であるが，構音障害者の発話データを十分に収録することは困難であるため，少量の発話データからモデルを学習しなければならないという問題がある．少量データからモデルを学習する際のアプローチとして，既存の学習済みモデルに対して目的ドメインの少量データを用いて Fine tuning をする，モデル適応の手法が考えられる [1]．本研究では，健常者の不特定話者モデルから構音障害者の特定話者モデルを構築するモデル適応を行う．

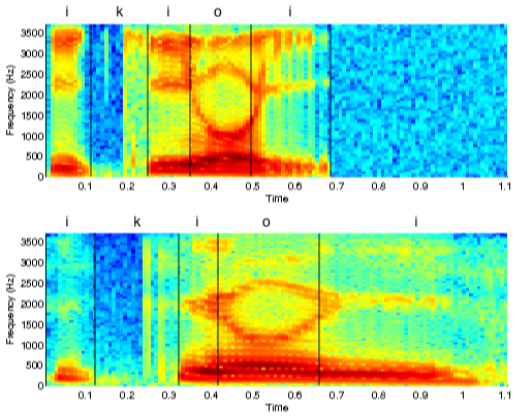

Fig. 1　Sample spectrograms of a physically unimpaired person (top) and a person with dysarthria (bottom)

また，データ数の不足の問題と関連して，出力記号の種類の問題がある．漢字を含む日本語文字単位の音声認識では，出力のノード数が数千以上と非常に大きく，文字の出現頻度が偏ってしまう．そのため，多くの文字記号が学習データに含まれず認識が困難となる．そこで本研究では，出力する文字記号のレベルを音素単位とし，出力ノード数を高々50 個程度に抑えることでこの問題を回避する．

2　モデル構造

End-to-End 音声認識を実現するアプローチとして，主に Connectionist Temporal Classification (CTC) [2] と，Attention 機構を用いた Encoder-Decoder モデル [3, 4] の 2 つが挙げられる．本項では，それぞれの手法の概要を述べる．

以下では，音響特徴量の系列を $\boldsymbol{x} = (x_1, ..., x_T)$ とし，ネットワークの出力記号系列 (本研究では音素系列) を $\boldsymbol{y} = (y_1, ..., y_L)$ とする．

---

* An investigation of speech recognition using Hybrid CTC/attention model for dysarthric speakers.
by Yuya Sawa, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki (Kobe University)

2.1 Connectionist Temporal Classification (CTC)

CTC では，出力記号列に空白文字列<blank>を挿入し，さらに同じ記号を連続で出力させることを許すことにより，出力記号系列を入力データ系列長 T に対応させている．ラベルの集合 $\boldsymbol{L}$ に<blank>を加えたラベル集合を $\boldsymbol{L}' = \{\boldsymbol{L}\} \cup \{<blank>\}$ とし，正解ラベル列を $\boldsymbol{l}$ とする．冗長なラベル列 $\boldsymbol{\pi} = (\pi_1, ..., \pi_T)$，$\pi_t \in \boldsymbol{L}'$ について，同一ラベルの繰り返しと<blank>を削除することにより，$\boldsymbol{y}$ が出力される．$\boldsymbol{\pi}$ から文字列を出力する関数を Ω とし，ラベル列 $\boldsymbol{l}$ を出力するパス $\boldsymbol{\pi}$ の集合を $\Omega^{-1}(\boldsymbol{l})$ とする．ここで，パス $\boldsymbol{\pi}$ が出力される確率は以下のようになる．

$$p(\boldsymbol{\pi}|\boldsymbol{x}) = \prod_{t=1}^{T} p(\pi_t|x_t) \tag{1}$$

ラベル列 $\boldsymbol{l}$ を出力する確率は，可能なパスの確率の総和であるから，

$$p(\boldsymbol{l}|\boldsymbol{x}) = \sum_{\boldsymbol{\pi} \in \Omega^{-1}(\boldsymbol{l})} p(\boldsymbol{\pi}|\boldsymbol{x}) \tag{2}$$

となる．CTC 損失関数はこの対数を取ることにより，

$$L_{ctc} = -\log(p(\boldsymbol{l}|\boldsymbol{x})) \tag{3}$$

と表される．また，CTC の損失関数の勾配は前向き・後ろ向きアルゴリズムにより効率的に計算される．

2.2 Attention ASR

Attention 機構を備えたモデルは，Encoder と Decoder の 2 つのモジュールから構成される．Encoder では入力音響特徴量系列 $\boldsymbol{x}$ が，RNN や LSTM を通じて中間表現 $\boldsymbol{h} = (h_1, ..., h_{T'})$ へと変換される．

$$\boldsymbol{h} = \text{Encoder}(\boldsymbol{x}) \tag{4}$$

本稿では，Encoder は pyramid bidirectional long short term memory (pBLSTM)[4] により構成されている．各層の出力は，連続した時間ステップ同士で連結され次の層へと送られる．このピラミッド構造が短い時間ステップから情報を効率的に集約することを可能にしており，計算の複雑性と学習および推論の時間を軽減する．

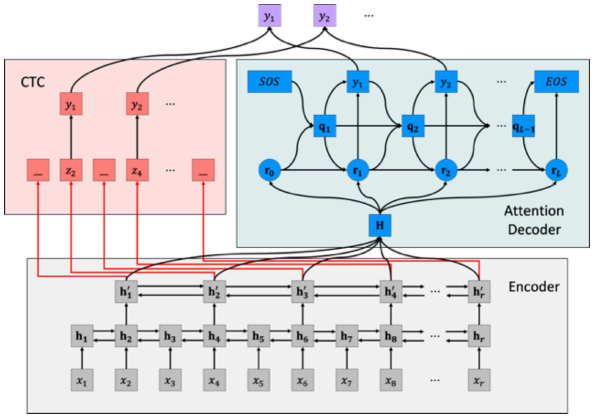

Fig. 2　Hybrid CTC/attention architecture

Decoder は 1 層の単方向 LSTM を使用する．Decoder では l 番目の隠れ状態 $\boldsymbol{q}_l$ は以下のように計算される．

$$\boldsymbol{q}_l = \text{LSTM}(\boldsymbol{r}_l, \boldsymbol{q}_{l-1}, y_{l-1}) \tag{5}$$

$$\boldsymbol{r}_l = \sum_t a_{l,t} h_t \tag{6}$$

$a_{l,t}$ は h_t の attention weight と呼ばれ，隠れ状態 h_t がどの出力 y_l に対応付けられるかを示す確率である．本稿では，$a_{l,t}$ の計算には次式で表される location-aware attention[5] を用いている．

$$e_{l,t} = \boldsymbol{w}^{\mathsf{T}} \tanh(\boldsymbol{W} q_{l-1} + \boldsymbol{V} h_t + \boldsymbol{U} f_{l,t} + \boldsymbol{b}) \tag{7}$$

$$a_{l,t} = \text{Softmax}(\{e_{l,t}\}_{t=1}^{T}) \tag{8}$$

$$\{\boldsymbol{f}_t\}_{t=1}^{T} = \boldsymbol{F} * [a_{l-1,1}, ..., a_{l-1,T}]^{\mathsf{T}} \tag{9}$$

$\boldsymbol{w}$,$\boldsymbol{W}$,$\boldsymbol{V}$,$\boldsymbol{U}$,$\boldsymbol{b}$,$\boldsymbol{F}$ は学習によって得られるモデルのパラメータである．

2.3 Hybrid CTC/attention モデル

Hybrid CTC/attention モデル [6] は，CTC と Attention モデルを組み合わせた手法である．Fig. 2 に示すように，Encoder は共通のネットワークを用いており，Decoder で CTC と Attention モデルの 2 つに分かれる．訓練時は，CTC と Attention モデルの損失関数を組み合わせてマルチタスク学習を行い，その損失関数は各モデルの損失の重み付き線形和で表される．

$$L_{MTL} = \alpha L_{ctc} + (1 - \alpha) L_{att} \tag{10}$$

CTC の損失関数は学習時に入力と出力間の単調なアライメントを明示的に学習する一方で，Attention 機構は単調なアライメントを明示しない．CTC の損失関数を組み合わせることは Attention 機構に単調なアライメントのみを許し，収束性

能の向上に繋がるという利点がある．また，推論時にも CTC と Attention モデルの両方の出力確率を組み合わせ，ビームサーチで推論を行う．

$$\log p_{hyb}(\boldsymbol{y}|\boldsymbol{h}) = \lambda \log(p_{ctc}(\boldsymbol{y}|\boldsymbol{h})) + (1-\lambda)\log(p_{att}(\boldsymbol{y}|\boldsymbol{h})) \quad (11)$$

3　評価実験

3.1　実験条件

提案手法は音素認識タスクで評価を実施した．モデルの学習および評価は，End-to-End 音声認識ツールキット ESPnet [7] を用いて行った．構音障害者音声は，アテトーゼ型構音障害者男性 2 名 (DYS1/DYS2) について，ATR 研究用日本語音声データベース [8] に含まれる音素バランス文 503 文を収録した．ただし，1 人目の話者 (DYS1) については症状により全ての文を読み上げておらず，503 文のうち 429 文を収録した．発話データのうち，50 文を評価データ，50 文を開発データ，残りを訓練データとして使用した．入力音響特徴量として，80 次元のメルフィルタバンク特徴にピッチ特徴を加えた計 83 次元の特徴量を用いた．出力次元数は，音素 39 種類に未知文字<unk>・始端記号<sos>・終端記号<eos>を加えた 42 次元とした．

音声認識モデルは，Hybrid CTC/attention モデルを使用した．共有の Encoder は 320 次元の隠れ層を持つ 4 層からなる pBLSTM とした．Attention 機構を用いた Decoder は 320 次元の隠れ層を持つ 1 層からなる単方向 LSTM と，その後の 42 次元のノードを持つ softmax の出力層から構成される．CTC と Attention モデルのマルチタスク学習における損失関数の重み α は 0.5 とし，デコード時の CTC 重み λ は 0.5 とした．最適化には Adadelta を用い，学習率は 1e-8，エポック数は 100 とした．また，上記のモデルを不特定健常者音声を用いて事前学習を行った．学習に際しての健常者音声は，日本語話し言葉コーパス (CSJ) の約 240 時間の音声を使用した．

3.2　実験結果

ベースラインとして，Table 1 に CSJ の不特定健常者音声データを使用して学習したモデルでの音素誤り率 (phoneme error rate; PER) を示す．構音障害者の認識精度が著しく悪く，これは構音障害者の発話スタイルが健常者のものとは大きく異なるためであると考えられる．

Table 1　PER [%] of a model trained on physically unimpaired person speech data

Speaker	PER [%]
DYS1	79.8
DYS2	95.3

次に，特定障害者音声を用いてモデルをランダム初期値から学習させた場合 (Scratch) と，Table 1 の実験で用いた不特定健常者モデルを初期モデルとしてモデル適応により学習させた場合 (CSJ adaptation) の比較を Table 2 に示す．不特定健常者話者モデルのパラメータを初期値として学習した場合に認識精度の向上が見られた．このことから多量の健常者音声を用いて事前学習を行うことで，より優れた特徴量空間を学習できていると考えられる．

Table 2　PER [%] of two aproaches to train the speaker-dependent dysarthria model

Speaker	Scratch	CSJ adaptation
DYS1	28.8	20.3
DYS2	29.4	19.3

モデル適応に使用する発話数を変えることで認識性能との関係を評価した．また比較として，ATR データベースより健常者 1 名 (MHT) についても同様にモデル適応の実験を行った．Fig. 3 は学習データ量に対する認識性能の変化を示している．学習データに含まれる発話数を増加させるにつれて性能改善の比率が小さくなり，健常者と同等の性能を達成できていないことが分かる．

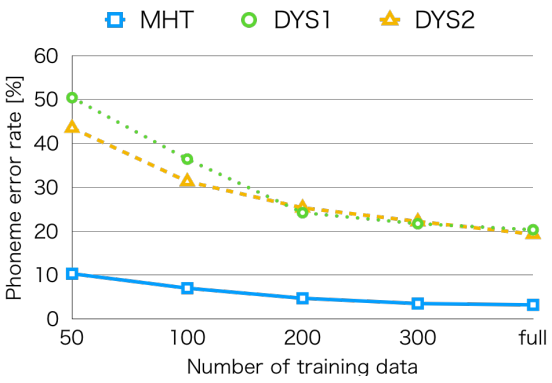

Fig. 3　The correlation between error rates and the number of traning data.　The number of 'full' in DYS1 is 329, and the others are 403.

Table 3　Example of recognition result for DYS2

Original text	老若男女が、火を囲んで飲み、手をつないで歌う。 (ATR-b02)
Ground truth	r o u ny a k u n a N ny o g a h i o k a k o N d e n o m i t e o ts u n a i d e u t a u
Predicted	r o u n a k u N n a N n o g a h i o k a k o N d e n o m i t e o ts u n a i d e u t a u

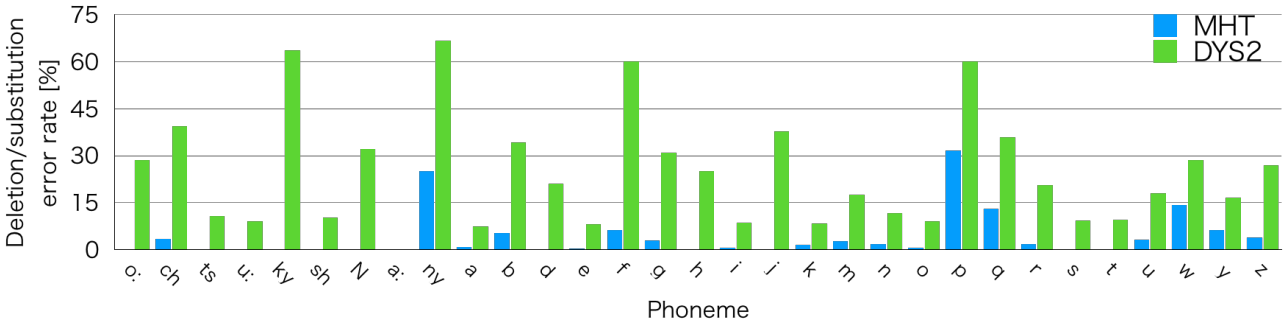

Fig. 4　Deletion/substitution error rates per phoneme

最後に，構音障害者の発話認識結果の例を示す．Table 3 は話者 DYS2 についてモデル適応を行った特定話者モデルにおける認識結果の一例である．認識結果を見ると，音素 \ny が正しく発話できていないことが分かる．Fig. 4 は話者 MHT と話者 DYS2 における音素別の誤り率を示す．ここでは評価データに 5 個以上含まれる音素について，置換と削除のどちらかの誤りをした割合を示している．構音障害者は特定の音素について認識率が著しく低下しており，特に破裂音の認識が難しい傾向が見て取れる．得られた音素誤りの傾向を調査することで，その話者にとって発音が難しい音素の分析に繋がることが期待される．

4　おわりに

本稿では，End-to-End 音声認識モデルによる構音障害者の音声認識について検討した．日本人健常者の不特定話者モデルを用いて構音障害者音声の特定話者に適応させることで，認識性能が向上することを示した．構音障害者は特定の音素の発音が困難であり，健常者の発話辞書の通りに発話ができないケースも多いと考えられる．そのため今後は，出力系列を文字や単語にすることで，発話辞書フリーなモデルの構築を検討していく．

参考文献

[1] K. Wang *et al.*, “Empirical evaluation of speaker adaptation on DNN based acoustic model,” in *Interspeech*, pp. 2429–2433, 2018.

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

[3] D. Bahdanau *et al.*, “End-to-end attention-based large vocabulary speech recognition,” in *ICASSP*, pp. 4945–4949, 2016.

[4] W. Chan *et al.*, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *ICASSP*, pp. 4960-4964, 2016.

[5] J. Chorowski *et al.*, “Attention-based models for speech recognition,” in *NIPS*, pp. 577-585, 2015.

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

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

[8] 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