構音障害者音声認識における認識モデルの比較評価*

◎高島遼一，有木康雄，滝口哲也 (神戸大)

1　はじめに

構音障害とは，音声器官の障害や運動機能障害などによって正しい発音が困難となる症状である。運動機能障害によって引き起こされる構音障害は，発話だけでなく手足も不自由なケースも多いため，手話のようなコミュニケーションの代替手段が取れず，音声コミュニケーションに頼らざるを得ない患者も多い。また聴覚障害に起因する構音障害者のような，手足に運動障害の無い患者であっても，自分の声を理解してもらいたいというニーズは高い。このような背景から，構音障害者音声を対象とした高精度な音声認識システムの実現が求められている。

近年では，deep neural network (DNN) による音声認識性能の大幅向上や，構音障害者音声のオープンデータセットの登場などにより，DNN ベースの構音障害者音声認識の研究がより活発になっている。多くの従来研究では，主に学習データ不足の問題について取り組んでおり，構音障害者音声を疑似的に生成する data augumentation のアプローチ [1, 2, 3] や，健常者音声認識モデルを構音障害者音声に適応するモデル適応のアプローチ [4, 5, 6] などが提案されている。

上記に挙げた従来研究の多くは，ベースラインとする音声認識モデルを 1 種類に限定して評価を行っている。例えば文献 [4] や [5] では listen, attend and spell (LAS)[7] と呼ばれる end-to-end モデルを使用し，また文献 [6] や [8] では，lattice-free maximum mutual information (LFMMI)[9] と呼ばれる DNN-hidden Markov model (DNN-HMM) を使用し，その上で各種提案手法を評価している。しかし，これらのモデルの中でどのモデルがよりベースラインとして適しているかについてはまだ明らかになっていない。文献 [10] では，複数の音響モデルや言語モデルについて構音障害者音声認識性能の比較検証を行っているが，音響モデルについては従来の Gaussian mixture model (GMM)-HMM と，一般的な cross entropy (CE) 基準の DNN-HMM との比較に留まっており，end-to-end モデルや LFMMI モデルのような state-of-the-art の音響モデルの比較は行われていない。さらに健常者音声認識の分野では，上記以外にも connectionist temporal classification (CTC) や，hybrid CTC/attention モデル [11]，transformer といった高性能な音響モデルが様々に提案されているため，これらのモデルの構音障害者音声認識における有効性を確認することは重要であると言える。

本研究では，構音障害者音声認識の研究において，どのモデルをベースラインとして用いるべきかを明らかにすることを目的として，複数の DNN ベースの音声認識モデルの比較評価を行う。DNN ベースの音響モデルは，DNN-HMM と end-to-end モデルに大別される。一般に end-to-end モデルは DNN-HMM と比べて大量の学習データを必要とする。そのため，学習データ量の乏しい構音障害者音声に対しては，データ量の観点からは DNN-HMM が適していると考えられる。一方で，DNN-HMM は学習時にフレーム単位のラベル（アライメント）情報を必要とするが，無音や spoken noise が多く含まれる構音障害者音声は，健常者音声と比べて安定したアライメント情報を得ることが困難であるため，音声の特性の観点からはアライメント情報を必要としない end-to-end モデルが適していると考えられる。

実験では，日本語を母語とする構音障害者 4 名の音声を用いて評価を行う。DNN-HMM モデルとして，CE 基準，state-level minimum Bayes risk (sMBR) 基準 [12]，LFMMI 基準それぞれで学習したモデルを使用し，end-to-end モデルとして attention-sequence-to-sequence (attention-s2s) モデル，CTC，またその複合モデルである hybrid CTC/attention モデルを使用する。さらに end-to-end モデルについては recurrent neural network 言語モデル (RNN-LM) の有無についても比較を行う。

2　認識モデルの比較評価

2.1　評価データ

評価データとして，日本語を母語とするアテトーゼ型脳性麻痺による構音障害者 4 名（全て男性）の音声を収録した。音声は，ATR 日本語データベースのテキスト 503 文を読み上げたものである。ただし，症状による理由から，503 文全てを読み上げていない被験者も存在しており，4 名全ての話者で合計 1,933 発話の音声を収録している。収録音声のうち，1,549 発話を学習データ，384 発話を評価データに分割して使用した。このとき，学習データと評価データのどちらにも 4 名全ての音声を含め，かつ発話内容は学習データと評価データで被らないように分割した。す

*A comparative evaluation of acoustic models on the task of dysarthric speech recognition. by Ryoichi Takashima, Yasuo Ariki, and Tetsuya Takiguchi (Kobe University)

なわち，話者クローズド，テキストオープンの条件で実験を行っている。

一般に日本語の連続音声認識性能を評価する場合は，単語単位の認識結果に対して word error rate を評価することが多い。一方，本実験は主に音響モデルを重視した比較を行うため，未知語や言語モデルの影響を受けない評価が望ましい。そこで音素単位の認識を行い，phone error rate (PER) での評価を行った。本実験と同様に音素単位認識を行う TIMIT コーパス用の Kaldi レシピに倣い，ラベルを 39 種類の音素で定義し，DNN-HMM 評価時における言語モデルとして，音素 bi-gram を使用した。

2.2 モデル

DNN-HMM の学習・評価，および end-to-end モデルの学習・評価はそれぞれ Kaldi[13] と ESPnet[14] を用いて行った。DNN-HMM を評価する際は 40 次元のメルフィルタバンク特徴量を使用し，end-to-end モデルを評価する際は 80 次元のメルフィルタバンク特徴に加えて 3 次元のピッチ特徴を使用した。以下に各評価モデルと，その学習条件を記載する。

CE-DNN-HMM: クロスエントロピー基準で学習した DNN-HMM である。前後 5 フレームの特徴量を結合したものを入力とし，中間層は 5 層の全結合層により構成した。各層のノード数は 1,024 とし，活性化関数は sigmoid 関数を用いた。

sMBR-DNN-HMM: 学習済みの CE-DNN-HMM を初期モデルとして，sMBR 学習 [12] を実施した。学習率の初期値は 1e-5 とし，エポック数は 6 とした。

LFMMI-TDNN: LFMMI 基準 [9] で学習した DNN-HMM である。モデル構造は先行研究 [6] と同じく，全結合層と time-delay neural network (TDNN) 層で構成される，10 層のネットワークを使用した。各中間層のノード数は 625，活性化関数は ReLU とし，batch normalization を適用した。

CTC および attention-s2s: 本実験では，CTC と attention-s2s をそれぞれ単独での評価に加えて，両者を結合した hybrid CTC/attention モデルも評価した。hybrid CTC/attention モデルについては，マルチタスク学習時の重みとデコード時の重みを同じパラメータ α に統一した。α の値が大きいほど CTC を重視するモデルとなり，$\alpha = 1.0$ の場合は CTC 単独での評価，$\alpha = 0.0$ の場合は attention-s2s 単独での評価に相当する。CTC および attention-s2s のエンコーダ部分については，5 層の pyramid 型 bidirectional gated recurrent unit を使用した。[1]attention-s2s のデコーダ部分については，1 層の unidirectional long-short term memory (LSTM) を用いた。

[1]フェアな比較のため，DNN-HMM ベースの 3 モデルについても，bidirectional LSTM を用いた評価を行ったが，性能は低下した。

Table 1　PERs [%] on non-dysarthric dataset.

Model	PER [%]
CE-DNN-HMM	5.8
sMBR-DNN-HMM	5.6
LFMMI-TDNN	4.7
CTC	5.4
attention-s2s	6.1

2.3 評価結果

2.3.1 同規模の健常者音声を用いた予備実験

構音障害者音声を用いた評価の前に，学習データ量とモデルの性能の関係を明らかにするため，同等規模の健常者音声を用いた予備実験を行った。健常者音声は ATR データセットより男性 4 名の音声を各話者 503 発話使用し，そのうち 1,612 発話を学習データ，400 発話を評価データとした。学習・評価データの分割方法は 2.1 節に記載の方法と同じく，話者クローズド，テキストオープンの条件である。

健常者音声を用いた各モデルの評価結果を Table 1 に示す。end-to-end モデルよりも DNN-HMM ベースのモデルの方が全体的に低い PER を示した。また end-to-end モデルの中でも，attention-s2s より CTC の方が低い PER を示した。一般に，end-to-end モデルは DNN-HMM に比べて大量の学習データが必要であり，また CTC よりも attention-s2s の方が複雑なモデル構造をしていることから，これらの結果は end-to-end モデルを十分に学習するためには，学習データ数が不足していることを表していると言える。一方，DNN-HMM の中では，最もモデル規模が大きい LFMMI-TDNN が最も良い性能を示していることから，DNN-HMM においては，本実験のデータ量でも十分に性能を引き出せていると考えられる。以上のことから，本実験条件において，データ量の観点からは DNN-HMM の方が有利であることが分かる。

2.3.2 構音障害者音声を用いた比較評価結果

構音障害者音声を用いた各モデルの評価結果を Table 2 に示す。attention-s2s 以外のモデルに注目した場合，CTC，LFMMI-TDNN，sMBR-DNN-HMM，CE-DNN-HMM の順に低い PER を示した。特に CTC は健常者音声を用いた予備実験では LFMMI-TDNN よりも PER が高かったにも関わらず，構音障害者音声を用いた実験においては，LFMMI-TDNN

Table 2　PERs [%] on dysarthric dataset.

Model	PER [%]
CE-DNN-HMM	48.4
sMBR-DNN-HMM	44.8
LFMMI-TDNN	35.6
CTC	26.8
attention-s2s	86.5

よりCTCの方が相対的に24.7%低いPERを示した。CE-DNN-HMMはフレーム単位での学習をしているのに対し，他のモデルはシーケンス単位での学習を行っている。sMBR-DNN-HMMはCE-DNN-HMMを初期モデルとしているため，その性能の影響を強く受ける。LFMMI-TDNNはスクラッチからシーケンス単位の学習を行えるが，学習の前処理などで，フレーム単位のラベル情報を必要とする。CTCは，フレーム単位のラベル情報を一切必要とせず，スクラッチからシーケンス単位での学習を行う。以上の性質から，フレーム単位のラベルを必要としないモデルほど，構音障害者音声の認識に適していると考えられる。これは，構音障害者音声は発話中にポーズやspoken noiseが多く存在するため，フレーム単位のラベル（アライメント）の推定が困難であるためと推察される。

CTCが最も良い性能を示している一方で，同じend-to-endモデルであるattention-s2sは著しく高いPERを示した。Fig. 1に，1エポック目と最終エポックにおける学習サンプルのattention matrixを示す。左側の図は，健常者音声を用いた予備実験時のattention-s2sの結果を示している。音声認識では，未来の音素を出力するにつれて，未来の音声フレームを参照するため，出力されるattention matrixは左側の図のように，対角成分に高い値が出ることが望ましい。しかしながら，構音障害者での評価結果である中央の図では，対角成分に高い値がでておらず，attention matrixが正しく推定されていないことが分かる。文献[11]では，学習データに雑音が多い場合，attention matrixの推定が不安定になることが指摘されており，構音障害者音声認識の場合においても，発話が不安定なため，attention matrixの推定が困難であったと考えられる。

文献[11]では，attention matrixの推定を安定化させることを目的として，hybrid CTC/attentionモデルの枠組みの中でCTCとattention-s2sをマルチタスク学習することを提案している。本実験においても，マルチタスク学習の重みであるαを変えながら，

Table 3　PERs [%] of hybrid CTC/attention models

α	w/o RNN-LM	w/ RNN-LM
1.0 (CTC)	26.8	26.1
0.8	27.4	27.3
0.6	26.3	26.1
0.5	25.7	25.7
0.4	28.2	28.2
0.2	30.5	30.6
0.05	55.4	57.6
0.0 (attention-s2s)	86.5	86.4

hybrid CTC/attentionモデルの評価を行った。実験結果をTable 3の2列目（w/o RNN-LM）に示す。表より，重みが0に近い場合に性能が大きく劣化するが，重みが0.2以上であれば性能が改善し，DNN-HMMを上回る性能が得られた。Fig. 1の右側の図は，重みが0.5の場合における attention matrix を示している。健常者音声を用いた場合ほど鮮明ではないものの，対角成分に高い値が得られていることが分かる。このことから，CTCとのマルチタスク学習により，attention matrixの推定を安定化することが，構音障害者音声認識において重要であると言える。

Table 3の2列目において，αが0.5の場合において，hybrid CTC/attentionモデルはCTCよりも良い性能を示している。ここで，αは前述のマルチタスク学習における重みを決めるパラメータであると同時に，デコーディングの際のCTC出力とattention-s2s出力の重みを決めるパラメータでもある。一般に，入力のコンテキスト情報のみを考慮しているCTCに比べて，attention-s2sはさらに出力のコンテキスト情報も考慮しているため，attention-s2sはCTCよりも性能が高いとされている。そこで，音響モデルの学習に用いた1,612発話を用いてRNN-LMを学習し，デコーディングの際に併用することで性能の変化を調査した。RNN-LMを併用した際の評価結果をTable 3の3列目（w/ RNN-LM）に示す。CTC単体を用いた場合 ($\alpha = 1.0$) では，RNN-LMの併用により若干の性能向上を示したのに対して，attention-s2sを用いているモデル ($\alpha < 1.0$) はほとんど性能に差が出なかった。これは，音響モデルと同じ学習データを使ってRNN-LMを学習させているため，既にコンテキスト情報を考慮して学習しているattention-s2sでは効果が無かったためであると考えられる。CTC単体とRNN-LMを併用した場合は，αが0.5のhybrid CTC/attentionモデルと同等の性能が得られている

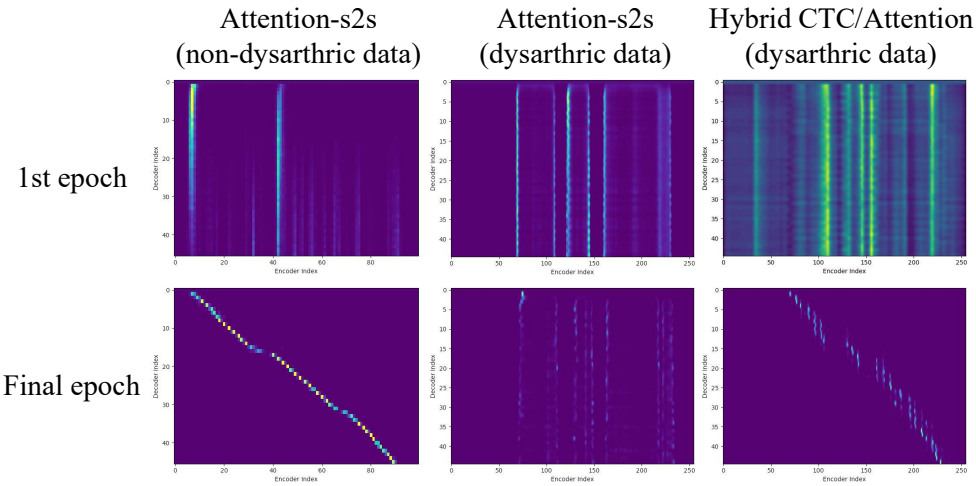

Fig. 1　Attention matrices estimated from training samples. The horizontal and vertical axes show the indexes of encoder and decoder, respectively.

ことから，CTC と hybrid CTC/attention モデルの差は，コンテキスト情報を考慮しているかの違いによって生じているものであり，構音障害者音声の性質との関連性は低いと推察される。

3　おわりに

各モデルの比較評価結果から，end-to-end モデルはデータ量が多少不足している場合であっても構音障害者音声認識にとって有意であり，また attention-s2s を用いる際は CTC とのマルチタスク学習により attention matrix の推定を安定化することが重要であることが明らかとなった。

謝辞　本研究の一部は，JSPS 科研費 19K24343 および 20K19862 の助成を受けたものである。

参考文献

[1] Yishan Jiao, et al., “Simulating dysarthric speech for training data augmentation in clinical speech applications,” in *ICASSP*, 2018, pp. 6009–6013.

[2] Bhavik Vachhani, et al., “Data augmentation using healthy speech for dysarthric speech recognition,” in *Interspeech*, 2018, pp. 471–475.

[3] F. Xiong, et al., “Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition,” in *ICASSP*, May 2019, pp. 5836–5840.

[4] J. Shor, et al., “Personalizing ASR for Dysarthric and Accented Speech with Limited Data,” in *Interspeech*, 2019, pp. 784–788.

[5] Y. Takashima, et al., “Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speech recognition,” *IEEE Access*, vol. 7, pp. 164 320–164 326, 2019.

[6] R. Takashima, et al., “Two-step acoustic model adaptation for dysarthric speech recognition,” in *ICASSP*, May 2020, pp. 6104–6108.

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

[8] H. Enno, et al., “Dysarthric speech recognition with lattice-free MMI,” in *ICASSP*, May 2020, pp. 6109–6113.

[9] D Povey, et al., “Purely sequence-trained neural networks for ASR based on lattice-free mmi,” in *Interspeech*, 2016, pp. 2751–2755.

[10] Z. Yue, et al., ”Exploring Appropriate Acoustic and Language Modelling Choices for Continuous Dysarthric Speech Recognition,” in *ICASSP*, May 2020, pp. 6094-6098.

[11] 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.

[12] K. Vesely, et al., “Sequence-discriminative training of deep neural networks,” in *Interspeech*, 2013, 2345-2349.

[13] D. Povey, et al., “The Kaldi speech recognition toolkit,” in *ASRU*, 2011.

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