6H-06

音韻特徴を用いた聴覚障害者音声のEnd-to-End音声認識

小林 彰夫 † 安 啓一 †
筑波技術大学 産業技術学部 †

1 はじめに

近年のスマートフォンやタブレットなどでは，対話的な情報検索やIoTデバイスを通じて電化製品を制御するなど，音声認識の用途が多様化している．しかし，音声認識の性能は聴覚障害者を含む発話障害者に対しては不十分な精度のままである．筆者らは，このような課題を解決するために聴覚障害者の発話を対象とした音声認識に取り組んでおり [1]，本稿では，音韻特徴を用いた音声認識手法による性能改善を報告する．

2 音韻特徴を用いた音声認識

2.1 聴覚障害者の音声コーパス

発話障害を含む非典型的な音声認識は，これまで運動機能性発話障害者を対象とすることが多く，多数の聴覚障害者を対象とした研究はあまりなかった．音声学的な観点からみた聴覚障害者の発話の特徴は，鼻音化や母音の中性化，二重母音化，子音の欠落や有声・無声の置換などとされているが [2]，障害者の発話は聴者の発話と比べてコーパスとして整備されてこなかった．筆者らは現在，音声認識を含む音声言語処理分野での利用を目的として，読み上げ文や自由発話などを含む，聴覚障害者による多様な発話を収集している．原稿執筆時点では合計50名の筑波技術大学学生の発話からなるコーパスを整備中である．

2.2 音韻特徴を用いたネットワーク

調音音声学では，調音機構の部位や運動などにより言語音を特徴づけている．また，音韻論では音素に共通する特徴をとらえるために，音素よりも小さな単位として音韻素性を用いている．例えば音素/s/と/z/を比較した場合，どちらも歯茎摩擦音（子音の調音位置および方法）であるが，有声性に関して対立していることになる．したがって，調音音声学や音韻論に基づく音韻的な特徴を併用すれば，ニューラルネットワークで学習される音素の識別能力を改善し，聴覚障害者の不明瞭な発話に対して，より頑健な音声認識が可能になると考えられる．

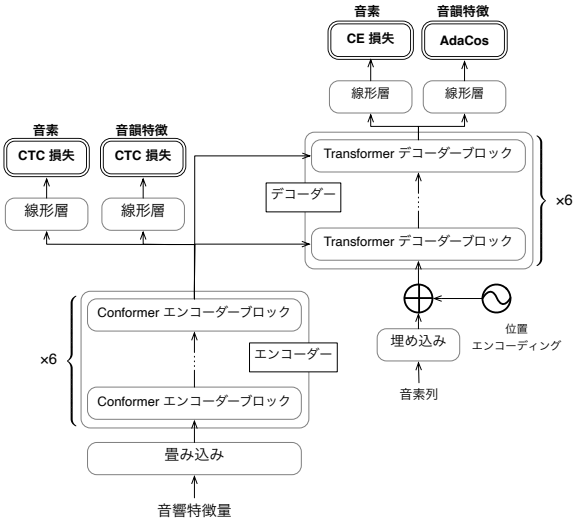

図 1: 実験で用いたネットワークと損失

本稿では，音素が所属する音韻カテゴリ（クラス）を音韻特徴と定める．音韻特徴を利用して音素認識を行うニューラルネットワークでは，音素と音韻特徴を表すクラスの両者を出力記号とする．音韻特徴は大きく音声アクティビティー，有声/無声，母音，子音，その他の5つのグループ (合計44カテゴリ) に分け，グループ内で対立関係にある特徴を表すカテゴリをネットワークから出力する．本稿の実験で用いるモデルは，図1に示すエンコーダー/デコーダー型のニューラルネットワークである．エンコーダーはConformer，デコーダーはTransformerとする．デコーダーの主たる出力は音素とし，デコーダー出力に対してクロスエントロピー損失，エンコーダー出力に対してCTC損失に基づくマルチタスク学習を行う．音韻特徴も音素同様，エンコーダー/デコーダー双方から出力するが，エンコーダーではCTC損失，デコーダーではAdaCos[3]を損失とする．音韻特徴による深層距離学習を追加することにより，デコーダーの出力となる音素の識別面は音韻特徴により補正され，音素の識別性能が改善すると期待できる．

An End-to-End Automatic Speech Recognition Using Corpus of Hearing-Impaired Speakers
†Akio KOBAYASHI, Keiichi YASU
Faculty of Industrial Technology, Tsukuba University of Technology

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

表 1: 損失の組み合わせ

	エンコーダー		デコーダー	
	CTC (音素)	CTC (音韻)	CE (音素)	AdaCos (音韻)
ベースライン	✓		✓	
提案法-1	✓		✓	✓
提案法-2	✓	✓	✓	✓

表 2: 音素誤り率 (%)

	聴者	障害者	
		NH	HI
ベースライン	7.3	38.8	34.1
提案法-1	7.1	37.5	33.7
提案法-2	6.0	36.9	31.4

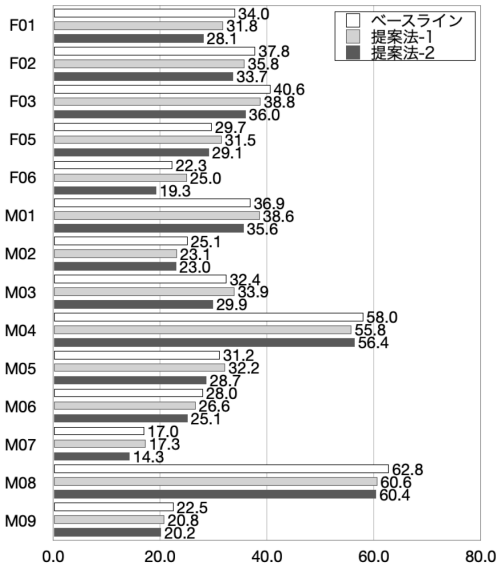

図 2: 話者別の音素誤り率 (%)

3 実験

評価対象として筑波技術大学の学生 14 名 (男子 9 名, 女子 5 名) による音素バランス文 (計 699 文) を用いた. また，比較のため，JNAS から聴者 120 名による音素バランス文 5,996 文を用いて評価する．学習データは聴者音声として ASJ，JNAS，ATR 多話者音声データベースから 77.9k 文 (計 123.7 時間) を用いた．また，学習用の聴覚障害者音声は，上の 14 名によるバランス文 3,525 文 (計 6.3 時間) を用いた.

ネットワークは図 1 の Conformer-Transformer とし，音素，音韻特徴に対する損失関数の組み合わせを変更して実験を行った (表 2)．JNAS を用いて学習したモデルを NH モデル，聴覚障害者の音声を用いて追加学習したモデルを HI モデルとする.

3.1 実験結果

表 1 に音素誤り率を示す。ベースラインに対してデコーダー出力に音韻特徴による AdaCos 損失を加えた (提案法-1) 場合，NH モデルによる障害者音声の誤り率はベースラインに比べて 3.4%の誤り削減率となった．一方，障害者で追加学習した HI モデルではベースラインに比べて 1.2%の誤り削減となった．エンコーダー出力に対しても音韻特徴による損失 (提案法-2) を加えると，NH モデルによる障害者音声の誤り率は 4.9%の誤り削減，HI モデルでは 7.9%の誤り削減率となった．エンコーダーおよびデコーダーの両者に音韻特徴による損失を加えた方が誤り削減効果が高く，音韻特徴は音素に基づく損失に対して相補的に機能すると考えられる．図 2 に話者別の音素誤り率を示す．ベースラインと比べると，提案法-1 は一部の話者で誤りを削減できなかったが，提案法-2 はすべての話者で誤りを削減した.

4 おわりに

聴覚障害者の end-to-end 音声認識に，音韻特徴を導入して認識性能の改善を行った．聴覚障害者の音声は，聴者に比べて遥かに少量規模のデータであることから，頑健な音声認識のためには，自己教師あり学習等による膨大な聴者音声からの事前学習モデルやデータ拡大法などを導入する必要がある.

謝辞

本研究の一部は JSPS 科研費 20H01716，21H00901 の助成を受けたものである.

参考文献

[1] A. Kobayashi et al., “Corpus Design and Automatic Speech Recognition for Deaf and Hard-of-Hearing People,” in *Proc. IEEE GCCE*, 2021, pp. 17–18.

[2] Osberger M. et al., “Speech Production Characteristics of the Hearing Impaired”, *Speech and Language*, 1982, vol.8, pp. 221-283.

[3] Zhang X. et al., “Adacos: Adaptively Scaling Cosine Logits for Effectively Learning Deep Face Representations”, in *Proc. CVPR*, 2019, pp. 10823-10832.

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.