

再学習を必要としないカスタマイズ可能な End-to-end 音声認識の研究

Retraining-free Customizable End-to-end Automatic Speech Recognition

周 藤 唯¹
Yui SUDO

Shakeel MUHAMMAD²

住 田 直 亮³
Naoaki SUMIDA

要 旨

本稿では、ユーザが任意に登録したフレーズに対する認識精度を改善するため、編集可能な用語リストを用いた音声認識モデルのカスタマイズ手法を提案する。提案手法では、従来の Connectionist Temporal Classification/attention 型音声認識モデルにテキストエンコーダと Cross-attention 層を追加し、音声信号に含まれる登録用語を検出する。また、登録用語の認識精度向上のため、検出された登録単語の確率や Connectionist Temporal Classification を統合した推論アルゴリズムを提案する。日本語の社内データセットおよび英語の Librispeech コーパスを用いた評価の結果、提案手法の適用により、登録用語の文字誤り率と単語誤り率が改善した。

ABSTRACT

In this paper, the authors propose a method of customizing automatic speech recognition models using editable term lists to increase recognition accuracy for phrases registered at the discretion of the user. The proposed method adds a text encoder and cross-attention layer to the conventional Connectionist Temporal Classification/Attention-based automatic speech recognition model to detect registered terms in speech signals. To increase registered term recognition accuracy, the authors propose an inference algorithm that integrates the probability of detected registered words and Connectionist Temporal Classification. Results of their evaluation using the in-house Japanese-language dataset and the English-language Librispeech corpus showed that using the proposed method lowered the character error rate and the word error rate for registered terms.

1. まえがき

深層学習の発展に伴い、音声認識技術は発展を遂げている。特に、近年では、End-to-end 音声認識モデルと呼ばれる、音声入力から直接文字列を推定する手法が大きな性能向上を果たしている⁽¹⁾。この飛躍的な性能向上により、スマートスピーカやカーナビゲーションシステム、議事録書き起こしシステム、ホンダコミュニケーションアシスタンスシステム^{(2), (3)}など、多くのアプリケーションに音声認識が搭載されるようになってきた。

End-to-end 音声認識は、従来のような音響モデル、発音辞書、言語モデルといった複数のコンポーネントで構成されるシステムとは異なり、単一のニューラルネットワークを用いて入力された音声信号を直接文字列に変換する。ニューラルネットワークは深層学習を用いて最適化され、

従来システムにおけるコンポーネント間の誤差蓄積を緩和することができる。一般的な音声認識モデルは、入力音声符号化をエンコーダと符号化された音声特徴をテキスト列に変換するデコーダから構成されている。これまでに、Connectionist Temporal Classification (CTC)^{(4), (5)}や Recurrent Neural Network Transducer^{(6), (7)}、注意機構 (Attention)^{(8), (9)}など、さまざまなデコーダが提案されている。また、これらの手法を組み合わせた複合型のシステムも提案されており^{(10), (11)}、その中でも CTC/Attention 型 End-to-end 音声認識は、その性能の高さから最も広く使用されている手法の一つである⁽¹⁰⁾。

しかし、深層学習を用いた音声認識の性能は、学習データへの依存傾向が高いため、学習データにほとんど出現しなかった単語やフレーズを正しく認識することが難しい。例えば、専門用語や個人名は学習データにはほとんど出現

¹ 株式会社ホンダ・リサーチ・インスティテュート・ジャパン, 博士 (工学)

² 株式会社ホンダ・リサーチ・インスティテュート・ジャパン, Ph.D.

³ 株式会社ホンダ・リサーチ・インスティテュート・ジャパン

しないため、正しく認識されないことが多い。しかし、これらの単語は特定のユーザにとって重要なキーワードとなる場合があり、誤認識してしまうとユーザの満足度を著しく低下させてしまう可能性がある。このような単語を含む学習データを収集し、ニューラルネットワークを再学習することで、認識性能を改善することは可能だが、学習データの収集やアノテーション、ニューラルネットワークの学習には人的リソースや高性能な GPU サーバが必要であるため、頻繁に再学習をおこなうことはできない。したがって、再学習を必要としない音声認識モデルのカスタマイズ手法が求められている。

これらの課題に対し、任意の単語を登録可能なカスタマイズ手法が提案されている^{(12), (13)}。これらの手法では、デコーダに Cross-attention 層⁽¹⁴⁾を追加することで、再学習することなく任意の品詞を持つ登録用語を検出できる。さらに、登録用語の検出精度を向上するため、補助ブランチを導入する手法や^{(13), (15)}、Cross-attention 層に補助損失関数を導入する方法なども提案されている⁽¹⁶⁾。しかし、これらの手法では、推論時に補助ブランチや補助損失を明示的に使用していないため、登録用語に対する性能向上は限定的である。

2. 研究のねらい

本稿では、ユーザが編集可能な用語リストを用いて、再学習することなく End-to-end 音声認識モデルをカスタマイズする手法を提案する。提案手法は、従来のカスタマイズ手法⁽¹⁶⁾と同様に、CTC/Attention 型音声認識モデル (Fig. 1)⁽¹⁰⁾に対し、テキストエンコーダと Cross-attention 層を追加することで、音声信号に含まれる登録用語を検出する (Fig. 2)。ただし、従来のカスタマイズ手法⁽¹⁶⁾とは異なり、登録用語の認識精度をさらに向上させるため、検出された登録用語の確率を増加させ、CTC デコーダによる出力確率を統合する推論アルゴリズムを提案する。本研究のねらいは以下の通りである。

- ・ 推論時に検出された登録用語の確率を増加させることで、登録用語の認識性能を改善する。
- ・ 推論時に CTC デコーダを用いた出力確率を統合することで、全体的な認識性能を改善する。
- ・ 日本語のデータセットにおける性能を詳細に分析する。

本稿は、以前の研究⁽¹⁶⁾を拡張したものである。具体的には、先行研究⁽¹⁷⁾に対して、CTC デコーダを用いた出力確率の統合および、日本語データセットにおける詳細な性能分析を新たにおこなっている。

3. 従来手法 (CTC/Attention モデル)

Fig. 1 に従来の音声認識モデルである CTC/Attention

モデル⁽¹⁰⁾を示す。CTC/Attention モデルは、音響エンコーダ、Attention デコーダ、および CTC デコーダを持つ。Attention デコーダと CTC デコーダは音響エンコーダを共有しており、それぞれ、長さ L の音声特徴列 $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_L]$ が与えられると、長さ S の文字列 $y = [y_1, \dots, y_S]$ の事後確率 $P_{\text{att}}(y | \mathbf{X})$, $P_{\text{ctc}}(y | \mathbf{X})$ を推定する。これらのデコーダおよび音響エンコーダは共同で学習され、推論時には二つのデコーダの出力確率が統合される。このアプローチにより、学習の安定化と推論時の性能向上が期待される。以下の節で各コンポーネントについて詳述する。

3.1. 音響エンコーダ

音響エンコーダには Conformer を使用する⁽⁷⁾。Conformer を用いた音響エンコーダは、音声特徴列 $\mathbf{X}$ を次のように、長さ T の隠れ状態ベクトル $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_T] \in \mathbb{R}^{T \times d}$ に変換する。

$$\mathbf{H} = \text{AudioEnc}(\mathbf{X}) \quad (1)$$

ここで、 d は隠れ状態ベクトルの次元を表す。

3.2. Attention デコーダ

Attention デコーダには、Transformer デコーダを使用する⁽¹⁴⁾。式 (1) に示した音響エンコーダによって抽出された隠れ状態ベクトル $\mathbf{H}$ および、1 フレーム前までの文字列 $y_{0:s-1} = [y_0, \dots, y_{s-1}]$ が与えられると、Attention デコーダは s 番目の文字 y_s の事後確率 $P(y_s | y_{0:s-1}, \mathbf{X})$ を以下のように推定する。

$$P_{\text{att}}(y_s | y_{0:s-1}, \mathbf{X}) = \text{AttDecoder}(y_{0:s-1}, \mathbf{H}) \quad (2)$$

式 (2) を $s = 1, \dots, S$ まで繰り返すことで、文字列 y の事後確率が以下のように定式化される。

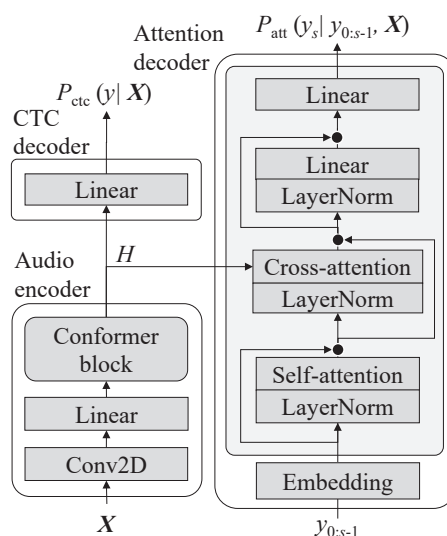


Fig. 1 Overall architecture of CTC/Attention model

$$P_{\text{att}}(y|X) = \prod_{s=1}^S P_{\text{att}}(y_s | y_{0:s-1}, X) \quad (3)$$

上記に対し、クロスエントロピー損失を導入する。

$$L_{\text{att}} = -\log P_{\text{att}}(y|X) \quad (4)$$

3.3. CTC デコーダ

CTC デコーダは、一層の線形層から構成される^{(4), (5)}。Fig. 1 に示すように、CTC デコーダは音響エンコーダが出力する隠れ状態ベクトル H を使用して、以下に示すように文字列 y の事後確率 $P_{\text{ctc}}(y|X)$ を推定する。

$$P_{\text{ctc}}(y|X) = \text{CTC}(H) \quad (5)$$

上記の出力に対し、以下に示す CTC 損失を導入する。

$$L_{\text{ctc}} = -\log P_{\text{ctc}}(y|X) \quad (6)$$

3.4. 共同学習と共同デコード

CTC/Attention モデルは、式 (4), (6) で定義した損失関数に対し、以下に示す重み λ を用いた共同学習により、損失が最小になるよう最適化される。

$$L = (1 - \lambda) L_{\text{att}} + \lambda L_{\text{ctc}} \quad (7)$$

また、推論時の性能向上のため、Attention デコーダと CTC デコーダの出力確率を以下に示す重み μ を用いて統合する。

$$\alpha = (1 - \mu) \alpha_{\text{att}} + \mu \alpha_{\text{ctc}}, \quad (8)$$

$$\alpha_{\text{att}} = \log P_{\text{att}}(y|X), \quad (9)$$

$$\alpha_{\text{ctc}} = \log P_{\text{ctc}}(y|X), \quad (10)$$

ここで、 α , α_{att} , α_{ctc} は、それぞれ、統合確率、Attention デコーダ、CTC デコーダの出力確率を表す。

従来の CTC/Attention モデルは、用語リストなどの外部入力を想定していないため、学習後に音声認識モデルをカスタマイズすることができず、学習データに出現しなかった単語を正確に認識することが難しい。

4. 提案モデルのネットワーク構造

Fig. 2 に提案モデルの全体構造を示す。提案手法は、従来の CTC/Attention モデルに加えて、用語リストを入力するためのテキストエンコーダを持ち、Attention デコーダに Cross-attention を追加している (Fig. 2 緑色部分)。テキストエンコーダは、ユーザが任意に編集可能な用語リストを符号化し、Cross-attention 層は入力音声の中の登録用語を検出する。

拡張された Attention デコーダは、音声特徴列 X と用語リスト $B = \{b_0, b_1, \dots, b_N\}$ が与えられると、長さ S の文字列 y の事後確率 $P_{\text{ext}}(y|X, B)$ および検出された登録用語のインデックス $n = [n_1, \dots, n_S]$ の事後確率 $P_{\text{idx}}(n|X, B)$ を推定する (Fig. 2)。以下の節で追加された各コンポーネントについて述べる。

4.1. テキストエンコーダ

テキストエンコーダには Transformer エンコーダを使用する⁽¹²⁾。Transformer エンコーダは、埋め込み層、Transformer ブロック、平均プーリング層、および編集可能な用語リスト $B = \{b_0, b_1, \dots, b_N\}$ から構成される。 b_N は N 番目の登録用語の文字列を表す (例えば、[朝, 礼] など)。 b_0 は音声信号の中に登録用語が存在しない場合に使用されるダミー単語である。用語リスト B は、埋め込み層および Transformer ブロックに入力され、以下に示すように文字ごとの特徴列 $G \in \mathbb{R}^{(N+1) \times l_{\text{max}} \times d}$ に変換される。

$$G = \text{Transformer}(\text{Embedding}(B)) \quad (11)$$

l_{max} は、用語リストに登録されている単語の最大長を表す (例えば、[朝, 礼] が最大長を持つ場合、 $l_{\text{max}} = 2$)。その後、平均プーリング層を使用して単語全体の特徴を表す特徴量 $V = [v_0, \dots, v_N] \in \mathbb{R}^{(N+1) \times d}$ に変換される。

$$V = \text{MeanPool}(G) \quad (12)$$

4.2. 拡張 Attention デコーダ

拡張された Attention デコーダは、式 (1), (12) に示す音響エンコーダおよびテキストエンコーダによって抽出された隠れ状態ベクトル H , V を使用して、文字列の事後確率 $P_{\text{ext}}(y|X, B)$ および検出された登録用語のインデックス確率 $P_{\text{idx}}(n|X, B)$ を推定する⁽¹²⁾。

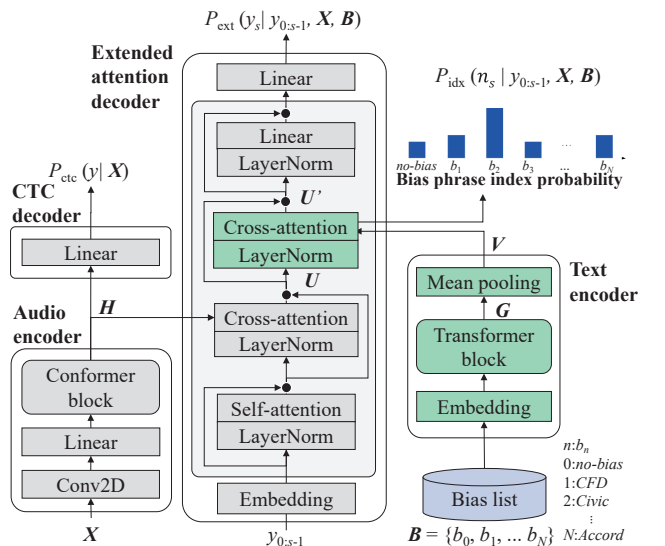


Fig. 2 Overall architecture of proposed method

4.2.1. 文字列確率

式 (1), (12) に示した音響エンコーダおよびテキストエンコーダによって抽出された隠れ状態ベクトル $\mathbf{H}$, $\mathbf{V}$ および, 1 フレーム前までの文字列 $y_{0:s-1}$ が与えられると, デコーダは s 番目の文字 y_s を以下のように推定する.

$$P_{\text{ext}}(y_s | y_{0:s-1}, \mathbf{X}, \mathbf{B}) = \text{ExtDecoder}(y_{0:s-1}, \mathbf{H}, \mathbf{V}) \quad (13)$$

式 (13) を $s = 1, \dots, S$ まで繰り返すことで, 文字列 y の事後確率は以下のように定式化される.

$$P_{\text{ext}}(y | \mathbf{X}, \mathbf{B}) = \prod_{s=1}^S P_{\text{ext}}(y_s | y_{0:s-1}, \mathbf{X}, \mathbf{B}) \quad (14)$$

上記に対し, クロスエントロピー損失を導入する.

$$L_{\text{ext}} = -\log P_{\text{ext}}(y | \mathbf{X}, \mathbf{B}) \quad (15)$$

4.2.2. 登録用語のインデックス確率

Fig. 2 における Cross-attention 層の入力, 出力をそれぞれ $\mathbf{U}$ および $\mathbf{U}'$ とすると, Cross-attention 層は以下のように定義される.

$$\mathbf{U}' = \text{Softmax}\left(\frac{\text{LN}(\mathbf{U})\mathbf{V}^T}{\sqrt{d}}\right)\mathbf{V} + \mathbf{U} \quad (16)$$

ここで, 式 (16) の Softmax 項は, 入力特徴列 $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_s]$ と用語リストの特徴列 $\mathbf{V} = [\mathbf{v}_0, \dots, \mathbf{v}_N]$ の内積であるため, それらの類似度を意味する. すなわち, 以下に示すように, 入力音声に含まれている登録用語のインデックス列 $n = [n_1, \dots, n_s]$ の確率と解釈することができる.

$$P_{\text{idx}}(n | \mathbf{X}, \mathbf{B}) = \text{Softmax}\left(\frac{\text{LN}(\mathbf{U})\mathbf{V}^T}{\sqrt{d}}\right) \quad (17)$$

この確率が最も高いインデックスが音声に含まれている登録用語であるといえる. 例えば, 用語リストに二つの単語 (1:朝礼, 2:PR) が登録されている状態で, $y = [\text{朝礼でPRする}]$ という音声が発話された場合, 登録用語のインデックス列は $n = [1, 1, 0, 2, 2, 0, 0]$ となる. なお, 登録用語ではない文字に関しては, $n = 0$ (no-phrase) が適用される. 上記の出力に対し, クロスエントロピーを用いた登録用語インデックス損失を導入する.

$$L_{\text{idx}} = -\log P_{\text{idx}}(n | \mathbf{X}, \mathbf{B}) \quad (18)$$

5. モデル学習および推論方法

5.1. 学習方法

提案手法は, 式 (6), (15), (17) で定義した損失関数に対し, 以下に示す重み λ および γ を用いたマルチタスク学習により損失が最小になるよう最適化される.

$$L = (1 - \lambda) L_{\text{ext}} + \lambda L_{\text{etc}} + \gamma L_{\text{idx}} \quad (19)$$

また, 学習中の用語リストに関しては, 推論時には任意の単語を用語リストに登録できるのに対し, 学習時にはバッチごとに Ground truth からランダムに用語リストが作成される. このランダムな選択は, モデルが幅広い単語やフレーズに対応できるようにするためである.

5.2. 推論アルゴリズム

認識性能をさらに向上させるため, 本稿では新たな推論アルゴリズムを提案する. 具体的には, 拡張された Attention デコーダが推定した文字列 y の事後確率 $P_{\text{ext}}(y | \mathbf{X}, \mathbf{B})$ に対して, (1) 登録用語のインデックス確率 $P_{\text{idx}}(n | \mathbf{X}, \mathbf{B})$ の統合 (Fig. 3 赤色部分), (2) CTC デコーダによる出力確率の統合 (Fig. 3 青色部分) をおこなう. 提案したアルゴリズムを Fig. 3 に示す.

- (1) まず, 拡張された Attention デコーダは式 (4) に基づいて, s 番目の各文字の確率 α_{ext} を算出する (5 行目). 次に, 式 (8) および argmax 関数を用いて, s 番目の登録用語インデックス n_s を推定する (6 行目). 登録用語が検出された場合 ($n_s \neq 0$), 登録用語に含まれる文字の出力確率を重み α_{bonus} に基づいて増加する (9 行目). 例えば, 検出された登録用語が $b_n = [\text{朝}, \text{礼}]$ の場合, $[\text{朝}], [\text{礼}]$ という文字の出力確率を増加する.
- (2) 続いて, CTC Prefix scoring⁽¹⁰⁾ を用いて CTC デコーダの出力確率 α_{ctc} を算出する. ただし, CTC デコーダ

Algorithm 1 Decoding algorithm

```

1:  $\text{hyps} = \{\langle \text{sos} \rangle; 1.0\}$ ;  $\text{end\_hyps} = \{\}$ ;
2: for  $s = 1$  to  $\text{maxlen}$  do
3:    $\text{ext\_hyps} = \{\}$ 
4:   for  $l \in \text{hyps}$  do
5:      $\alpha_{\text{ext}} = \log P_{\text{ext}}(y_s | y_{1:s-1}, \mathbf{X}, \mathbf{B})$ 
6:      $n_s \leftarrow \text{Argmax}(P(n_s | y_{1:s-1}, \mathbf{X}, \mathbf{B}))$ 
7:     if  $n_s \neq 0$  then
8:       for  $\text{token} \in b_n$  do
9:          $\alpha_{\text{ext}}[\text{token}] \leftarrow \alpha_{\text{ext}}[\text{token}] + \alpha_{\text{bonus}}$ 
10:      end for
11:    end if
12:    for  $y_s \in \text{Top-k}_{\text{beam}}(\alpha_{\text{ext}})$  do
13:      Add  $\tilde{l} = l \oplus y_s$  and the score  $\alpha_{\text{ext}}$  to  $\text{ext\_hyps}$ 
14:       $\alpha_{\text{ctc}} = \text{CTCPrefixScore}(\tilde{l}, \mathbf{H})$ 
15:      if  $\hat{n}_s = 0$  then (“no-bias”)
16:         $\text{ext\_hyps}[\tilde{l}] = \mu \alpha_{\text{ctc}} + (1 - \mu) \alpha_{\text{ext}}$ 
17:      else
18:         $\text{ext\_hyps}[\tilde{l}] = \alpha_{\text{ext}}$ 
19:      end if
20:    if  $y_s = \langle \text{eos} \rangle$  or  $\text{len}(\tilde{l}) = \text{maxlen}$  then
21:      Add  $\tilde{l}$  to  $\text{end\_hyps}$ 
22:    end if
23:  end for
24: end for
25:  $\text{hyps} = \text{Top-k}(\text{ext\_hyps})$ 
26: end for
27: return  $\text{Top-1}(\text{end\_hyps})$ 

```

Fig. 3 Joint decoding algorithm

には用語リスト B が入力されないため (Fig. 2), 登録用語に対する確率算出はおこなわない. すなわち, 登録用語が検出されなかった場合のみ ($n_s = 0$), 重み μ を用いて出力確率を統合する (16 行目).

6. 評価実験

6.1. 実験条件

入力される音声特徴量 X は, 512 サンプルの窓長を持つ 80 次元のメルフィルタバンクである. メルフィルタバンクとは, 人の聴覚特性を模したメル尺度を用いた時間-周波数特性を表す. 音響エンコーダは, 二つの畳み込み層, 256 次元の線形射影層, 12 個の Conformer ブロックから構成される. テキストエンコーダは, 埋め込み層, 6 個の Transformer ブロックから構成され, 最終層に平均プーリング層を持つ. 拡張された Attention デコーダは, 埋め込み層, 6 個の Transformer ブロック, 線形層から構成される. CTC デコーダは 1 層の線形層で構成される. 式 (19) に示す学習時の重み λ , γ は, それぞれ 0.3, 1.0 に設定した. また, 5.2. 節で述べた推論時のパラメータ α_{bonus} , μ は, それぞれ 1.0, 0.3 に設定した.

評価には, 社内データセット (日本語) と Librispeech-100 コーパス (英語)⁽¹⁸⁾ を用いた. 社内データセットは, 会議・朝礼シナリオを含む 93 時間の社内音声データ, 日本語話し言葉コーパス⁽¹⁹⁾, 株式会社国際電気通信基礎技術研究所が開発したデータベース⁽²⁰⁾を含む. 提案手法は, 文字誤り率 (CER), 登録用語の CER (B-CER), およびそれ以外の CER (U-CER) で評価した. 提案手法の目標は, U-CER および全体の CER の増加を最小限に抑えつつ, B-CER を改善することである. Librispeech-100 に関しては, 単語誤り率 (WER), U-WER, B-WER を使用した.

6.2. 日本語データセットでの検証

Table 1 に, $N = 203$ 個の単語やフレーズを用語リストに登録した場合の評価結果を示す. ベースラインはカスタマイズ性を持たない従来の CTC/Attention モデル⁽¹⁰⁾である (A1). また, 従来のカスタマイズ手法^{(15), (21)}とも比較した (B1, 2). これらの手法は, 提案手法と異なり, 推論時に検出された登録用語の確率を明示的に使用しない手法である. なお, 提案手法に関しては, 5.2. 節で述べた推論アルゴリズム (登録用語のインデックス確率統合, CTC デコーダによる確率統合)を除いた場合の影響も検証した (C2, 3).

まず, 提案手法とベースライン, 従来のカスタマイズ手法を比較する (A1, B1, B2 対 C1). C1 は, ベースラインと比較して, U-CER (登録用語“以外”の CER) が 0.06 ポイント増加したものの, B-CER (登録用語の CER) を 11.71 ポイント改善した. その結果, 全体の CER が 1.40 ポイント改善した. また, 提案手法は従来のカスタマイズ手法^{(15), (21)}

Table 1 Experimental results on in-house dataset

ID	Model	CER [%]	U-CER [%]	B-CER [%]
A1	Baseline	9.85	8.17	21.86
B1	NEA-ASR [21]	9.75	8.11	21.90
B2	Intermediate DB [15]	9.28	8.23	16.93
C1	Proposed method	8.45	8.23	10.15
C2	C1 w/o index probability	8.65	8.19	12.11
C3	C2 w/o CTC integration	9.78	9.16	14.54

Table 2 Performance analysis on in-house dataset

Class	Accuracy [%]	Examples
Katakana	86	ジャイロ, リファイン
Alphabet	84	CFD, PR
Kanji	64	接種, 風洞
English	29	SKETCH, CASCADE

よりも全体の CER および B-CER の両方で改善を示した.

次に, 登録用語のインデックス確率統合, CTC デコーダによる確率統合を除いた場合の影響について議論する. C2 および C3 に示すとおり, 登録用語のインデックス確率統合を除くことで B-CER は 1.96 ポイント増加し, CTC デコーダによる確率統合を除くと U-CER が 0.97 ポイント増加した. これにより, それぞれが登録用語およびそれ以外の単語に対する認識性能を改善する効果があることが確認された.

6.2.1. 性能分析

続いて, 登録用語をカタカナ, アルファベット, 漢字, 英語で表記された単語に分類し, 認識精度を分析した. 複数の文字種が使用されている単語に関しては, 最も多く使われている文字の分類とした. 例えば, 「CFD 計算」の場合, アルファベットが最も多く使用されているためアルファベットに分類した. なお, アルファベットとは, 「PR (ピーアール)」のようにアルファベット読みで発音される単語を示し, 英語は, 「SKETCH (スケッチ)」のようにそれ以外で発音される単語を示す.

Table 2 および Fig. 4 に, 登録用語の分類ごとの認識精度, ならびに, 典型的な推論結果例を示す. 推論結果例 1 から 3 は提案手法により登録用語が正しく認識できた例を示し, 推論結果例 4 は正しく認識できなかった例を示す. Table 2 より, カタカナやアルファベットのように, 各文字と発音が一意に決定する単語に関しては 80% 以上の認識精度を達成したのに対し, 英語に対する認識精度は 29% と明らかに低かった. 特に「SKETCH (スケッチ)」や「CASCADE (カスケード)」といった単語のように, 一般にはカタカナで表記されることの多い英語の認識精度が著しく低かった.

これは, 学習中に登録される単語がランダムに選択されたことに起因すると考えられる. 提案手法 (C1) におけるテ

キストエンコーダおよび拡張された Attention デコーダに追加された Cross-attention 層は、入力音声から登録単語の発音を検出するよう学習が進む。そのため、一般的ではない表記と発音の組み合わせを持つ単語に対しては十分に学習が進まなかったことが考えられる。

6.2.2. スケーラビリティ検証

実際のユースケースでは、ユーザが明示的に登録用語を削除しない限り、用語リストに登録される単語数は増加する傾向にある。Fig. 5 に、B-CER (登録単語の CER) に対する用語リストサイズの影響を示す。本実験では、用語リストに単語を登録しない場合 ($N = 0$)、そのユースケースに適切な単語を登録した場合 ($N = 203$)、不必要な単語を追加した場合 ($N = 500$ から 2000) の B-CER を比較した。

用語リストに単語を登録しない場合 ($N = 0$)、B-CER には改善が見られなかったのに対し、適切な単語を登録した場合 ($N = 203$)、B-CER は改善した。対象とするユースケースに不要な単語が増加すると、徐々に B-CER は増加した ($N = 500$ から 2000)。ただし、用語リストに単語を登録しない場合 ($N = 0$) と比較して十分な B-CER の改善効果は維持できていたことから、提案手法は用語リストのサイズ増加に対してロバストであることが確認された。

- Reference: 一番のネックは風洞だって事だ
Baseline: 一番の**ネッ**カーフードだって事だ
Proposed (C1): 一番の**ネッ**クは風洞だって事だ
- Reference: それとは別の**職域**接種
Baseline: それとは別の**職域**摂取
Proposed (C1): それとは別の**職域**接種
- Reference: **被引例情報**のときに**第一引例**って
Baseline: **非陰例情報**のときに**第一因例**って
Proposed (C1): **非引例情報**のときに**第一引例**って
- Reference: **SKETCH** モードを開いて
Baseline: **スケッチ**モードを開いて
Proposed (C1): **スケッチ**モードを開いて

Boldface: Registered phrase
Red: Incorrectly recognized characters
Blue: Correctly recognized characters

Fig. 4 Typical inference results

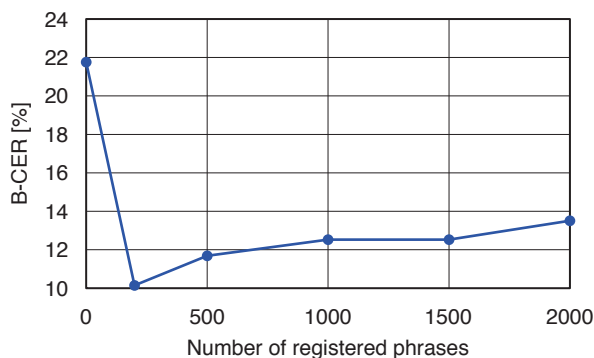


Fig. 5 Robustness to term list size

Table 3 Experimental results on Librispeech-100 h

ID	Model	WER [%]	U-WER [%]	B-WER [%]
A1	Baseline	8.59	5.87	30.71
C1	Proposed method	5.76	4.10	19.30
C2	C1 w/o index probability	6.49	4.83	19.98
C3	C2 w/o CTC integration	7.53	5.27	25.92

6.3. 英語データセットでの検証

次に、日本語とは異なる性質を持つ英語データセット (Librispeech-100) に対し、 $N = 100$ 単語を用語リストに登録した場合の評価をおこなった。日本語が表意文字を使用した言語であるのに対し、英語は表音文字を使用した言語であるため、このような性質の異なる言語に対して評価をおこなうことは重要である。Table 3 に、提案手法および登録用語のインデックス確率統合、CTC による確率統合の効果を示す。英語の評価結果についても、日本語の場合と同様の傾向が見られた。提案手法は、ベースラインと比較して、B-CER を改善し、その結果、全体の WER を改善した (A1 対 C1)。また、登録用語のインデックス確率統合および CTC デコーダによる確率統合を除くと、B-WER および U-WER が低下したことから、これらの手法がそれぞれの認識性能を向上させる効果があることが確認された (C1 対 C2, 3)。

また、Fig. 6 に式 (8) に基づく Cross-attention 層により推定された登録用語のインデックス確率を示す。横軸に入力された音声発話、縦軸に登録された単語を示している。この音声発話の中には、legality, legislature という登録単語が含まれていた。Fig. 6 より、提案手法は入力された音声から正確に登録単語を検出できることが確認できる。

以上より、提案手法は性質の異なる言語 (日本語、英語) に対しても有効であることが確認された。

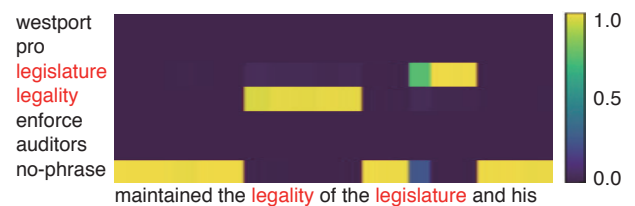


Fig. 6 Registered phrase detection

7. まとめ

本稿では、ユーザが任意に登録したフレーズに対する認識精度を改善するため、編集可能な用語リストを用いた音声認識モデルのカスタマイズ手法を提案した。提案手法では、音声信号に含まれる登録用語を検出するために、従来

の CTC/Attention 型音声認識モデルにテキストエンコーダと Cross-attention 層を追加した。また、登録用語の認識精度をさらに向上させるため、検出された登録用語の確率を増加させ、CTC デコーダによる確率統合を用いた推論アルゴリズムを提案した。社内データセット (日本語) および Librispeech コーパス (英語) を用いた評価の結果、提案手法は登録用語の文字誤り率と単語誤り率を改善した。

参考文献

- (1) Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R., Watanabe, S.: End-to-End Speech Recognition: A Survey, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 32, p. 325-351, (2023), doi: 10.1109/TASLP.2023.3328283
- (2) <https://global.honda.jp/stories/003.html>, (参照 2024/9/23)
- (3) <https://www.jp.honda-ri.com/activity/honda-ca-system/>, (参照 2024/9/23)
- (4) Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *ICML '06: Proceedings of the 23rd international conference on Machine learning*, p. 369-376, (2006), doi: 10.1145/1143844.1143891
- (5) Peng, Y., Sudo, Y., Shakeel, M., Watanabe, S.: OWSM-CTC: An Open Encoder-Only Speech Recognition Foundation Model for Speech Recognition, Translation, and Language Identification, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Vol. 1, p. 10192-10209, (2024), doi: 10.18653/v1/2024.acl-long.549
- (6) Graves, A.: Sequence Transduction with Recurrent Neural Networks, *International Conference of Machine Learning (ICML) 2012 Workshop on Representation Learning*, arXiv preprint, (2012), doi: 10.48550/arXiv.1211.3711
- (7) Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R.: Conformer: Convolution-augmented Transformer for Speech Recognition, *Proc. Interspeech 2020*, p. 5036-5040, (2020), doi: 10.21437/Interspeech.2020-3015
- (8) Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-Based Models for Speech Recognition, *Advances in Neural Information Processing Systems*, Vol. 28, (2015)
- (9) Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust Speech Recognition via Large-Scale Weak Supervision, *Proceedings of the 40th International Conference on Machine Learning*, p. 28492-28518, (2023)
- (10) Watanabe, S., Hori, T., Kim, S., Hershey, J. R., Hayashi, T.: Hybrid CTC/Attention Architecture for End-to-End Speech Recognition, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 11, Issue 8, p. 1240-1253, (2017), doi: 10.1109/JSTSP.2017.2763455
- (11) Sudo, Y., Shakeel, M., Yan, B., Shi, J., Watanabe, S.: 4D ASR: Joint modeling of CTC, Attention, Transducer, and Mask-Predict decoders, *Proc. INTERSPEECH 2023*, p. 3312-3316, (2023), doi: 10.21437/Interspeech.2023-1303
- (12) Pundak, G., Sainath, T. N., Prabhavalkar, R., Kannan, A., Zhao, D.: Deep Context: End-to-end Contextual Speech Recognition, 2018 IEEE Spoken Language Technology Workshop (SLT), (2018), doi: 10.1109/SLT.2018.8639034
- (13) Huang, K., Zhang, A., Yang, Z., Guo, P., Mu, B., Xu, T., Xie, L.: Contextualized End-to-End Speech Recognition with Contextual Phrase Prediction Network, *Proc. INTERSPEECH 2023*, p. 4933-4937, (2023), doi: 10.21437/Interspeech.2023-767
- (14) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I.: Attention is all you need, *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 6000-6010, (2017)
- (15) Shakeel, M., Sudo, Y., Peng, Y., Watanabe, S.: Contextualized End-to-end Automatic Speech Recognition with Intermediate Biasing Loss, *Proc. Interspeech 2024*, p. 3909-3913, (2024), doi: 10.21437/Interspeech.2024-1257
- (16) Huber, C., Hussain, J., Stüker, S., Waibel, A.: Instant One-Shot Word-Learning for Context-Specific Neural Sequence-to-Sequence Speech Recognition, 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), (2021), doi: 10.1109/ASRU51503.2021.9687898
- (17) Sudo, Y., Shakeel, M., Fukumoto, Y., Peng, Y., Watanabe, S.: Contextualized Automatic Speech Recognition With Attention-Based Bias Phrase Boosted Beam Search, *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 10896-10900, (2024), doi: 10.1109/ICASSP48485.2024.10447782
- (18) Panayotov, V., Chen, G., Povey, D., Khudanpur, S.:

- Librispeech: An ASR corpus based on public domain audio books, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 5206-5210, (2015), doi: 10.1109/ICASSP.2015.7178964
- (19) Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation, Proc. ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition, (2003), https://www.isca-archive.org/sspr_2003/maekawa03_sspr.pdf, (参照 2024/9/23)
- (20) Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., Shikano, K.: ATR Japanese speech database as a tool of speech recognition and synthesis, Speech Communication, Vol. 9, Issue 4, p. 357-363, (1990), doi: 10.1016/0167-6393(90)90011-W
- (21) Sudo, Y., Hata, K., Nakadai, K.: Retraining-free Customized ASR for Enharmonic Words Based on a Named-Entity-Aware Model and Phoneme Similarity Estimation, Proc. INTERSPEECH 2023, p. 491-495, (2023), doi: 10.21437/Interspeech.2023-1320

■ 著 者 ■



周 藤 唯



Shakeel MUHAMMAD



住田 直 亮