MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Exploring Keyword Enrollment for Japanese End-to-End Automatic Speech Recognition using Contextual Biasing

Mitsui, Yoshiki; Aihara, Ryo; Hori, Takaaki; Le Roux, Jonathan; Taguchi, Shinya

TR2024-073 June 18, 2024

Abstract

End-to-end (E2E) automatic speech recognition (ASR), which has emerged with the development of deep learning, exhibits generally higher performance than conventional modular ASR methods. However, E2E ASR has the drawback that it is difficult to enroll keywords for specific domains, which was easily realized in conventional ASR. Contextual biasing has been proposed for keyword enrollment methods for E2E ASR, but, for Japanese ASR, the performance is not sufficient when we enroll keywords which do not appear in the training data. To overcome this problem, we propose an updated keyword enrollment method where we use phonetic letter notations such as katakana or hiragana to recognize enrolled keywords, converting them back to their original notations in a postprocessing step. Additionally we propose an improved E2E ASR model training method to strengthen the connection between acoustic features obtained from input speech and phonetic letter notations by replacing some words from origial notation to phonetic letter notation. We observed higher keyword enrollment performance for keywords longer than five moras by using the proposed methods.

OTOGAKU Symposium 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

[ポスター講演] Contextual Biasingを用いた日本語End-to-End音声認識向け語彙登録の検討

三井 祥幹[1,a)]　相原 龍[1]　堀 貴明[2,†1]　ルルー ジョナトン[2]　田口 進也[1]

概要：深層学習の発展に伴い登場した end-to-end (E2E) 音声認識は，従来の階層型音声認識と比較し，総合的に高い性能を発揮する．しかし，階層型音声認識で容易に実現できていた，特定ドメイン向けの語彙登録が困難である欠点を抱えている．E2E 音声認識向けの語彙登録手法として，contextual biasing を用いる方法が提案されているが，特に日本語音声認識では，学習データに現れない表記を含む語彙を登録する場合に，十分な認識性能を得られない．これを解消するため，本稿では，語彙の登録にカタカナ・ひらがな等の表音文字による表記を利用し，音声認識結果テキストに対する後処理で，登録に用いた表記を，元の表記へと戻す改良手法を提案する．更に，表音文字による語彙の表記と，入力音声より得られる音響特徴量との結びつきを強めるため，E2E 音声認識モデルを学習させる際に，学習用テキストの一部の単語を，ランダムに表音文字表記へ置換する改良学習手法を併せて提案する．提案手法により，5 モーラ以上からなる語彙の登録タスクにおいて，元表記を利用し語彙を登録する従来手法よりも高い語彙登録性能が得られることを確認した．

Exploring Keyword Enrollment for Japanese End-to-End Automatic Speech Recognition using Contextual Biasing

Yoshiki MITSUI[1,a)]　Ryo AIHARA[1]　Takaaki HORI[2,†1]　Jonathan LE ROUX[2]　Shinya TAGUCHI[1]

Abstract: End-to-end (E2E) automatic speech recognition (ASR), which has emerged with the development of deep learning, exhibits generally higher performance than conventional modular ASR methods. However, E2E ASR has the drawback that it is difficult to enroll keywords for specific domains, which was easily realized in conventional ASR. Contextual biasing has been proposed for keyword enrollment methods for E2E ASR, but, for Japanese ASR, the performance is not sufficient when we enroll keywords which do not appear in the training data. To overcome this problem, we propose an updated keyword enrollment method where we use phonetic letter notations such as katakana or hiragana to recognize enrolled keywords, converting them back to their original notations in a postprocessing step. Additionally we propose an improved E2E ASR model training method to strengthen the connection between acoustic features obtained from input speech and phonetic letter notations by replacing some words from origial notation to phonetic letter notation. We observed higher keyword enrollment performance for keywords longer than five moras by using the proposed methods.

1. まえがき

音声認識機能は，録音データの書き起こしや，機器・装置の音声インタフェースなど，広い場面で使用されている．深層学習の発展に伴い登場した，単一の deep neural network (DNN) のみを用いて音声特徴量よりテキストの書き起こしを実現する end-to-end (E2E) 音声認識 [1], [2] は，従来の階層型音声認識エンジン [3] と比較し，総合的に高い音声認識性能を実現する．

音声認識エンジンに対しては，しばしば認識対象とするドメイン向けに，認識対象とする語彙のカスタマイズ性が

[1] 三菱電機株式会社 情報技術総合研究所
〒 247–8501 神奈川県鎌倉市大船 5–1–1
[2] Mitsubishi Electric Research Laboratories
201 Broadway, Cambridge, Massachusetts 02139, U.S.A.
[†1] 現在，Apple
Presently with Apple
[a)] Mitsui.Yoshiki@eb.MitsubishiElectric.co.jp

求められる．例えば，医療関係者向けの音声認識エンジンでは，疾病や薬品など医療ドメインの語彙を高い精度で認識できる性能が求められる．このほか，人名や地名といった固有名詞についても，状況に応じ認識させやすくすることが求められる．従来の階層型音声認識エンジンでは，語彙の発音（読み）と表記の情報を有する「発音辞書」モジュールが存在しており，当該モジュールへ語彙を登録すれば，特定ドメイン向けのカスタマイズを実現できていた．しかし，原則として単一 DNN のみで構成されている E2E 音声認識では，語彙と発音の関係を示すモジュールが陽に存在しない．このため，特定ドメインにのみ頻出する語彙を認識できるようカスタマイズするためには，新たなデータセットを用意し，E2E 音声認識モデル全体を再学習させる必要がある．再学習には多大な演算コストが必要であり，従来型エンジンと比べ大幅に利便性が低下している．近年では，E2E 音声認識モデルに contextual biasing を組み合わせ，音声認識の出力結果テキストへスコアを付ける際に，特定の語彙へバイアスを付与し，出力されやすくする手法が提案されている [4], [5]．バイアスの付与には，階層型音声認識においても活用されている weighted finite-state transducer (WFST) が利用されている．しかし，筆者らの検証では，本手法を日本語の E2E 音声認識へ適用する場合に，登録した語彙が認識結果に現れない事象の発生を確認している．特に固有名詞など，E2E 音声認識モデルの学習データに含まれない表記に対して，そのような傾向が顕著である．この原因として，学習データに出現しない表記を含む語彙に対しては，E2E 音声認識モデルの内部において「音声」と「文字」との間の結びつきが十分に確立されていないことが考えられる．前述の課題を解消する一手段として，より適した他の表記を語彙の登録に利用し，認識結果テキスト中に現れる当該表記を，後処理で元の表記へと戻す手法が考えられる．例えば，登録したい語彙の読みを示すカタカナの表記は，カタカナ文字単独であれば学習データ中に多数出現することから，元の表記と比べ，語彙登録に適すると考えられる．本稿では，E2E 音声認識と contextual biasing を組み合わせた語彙の追加登録を実現する枠組みにおいて，本来表記の代わりに，他の表記を用いて語彙を登録する手法および，カタカナ・ひらがなを他の表記として用いる場合の E2E 音声認識モデルの改良学習手法について検討し，結果を報告する．

2. 語彙登録機能付き E2E 音声認識

2.1 E2E 音声認識

従来の階層型音声認識 [3] は，「音響モデル」「発音辞書」「言語モデル」といった複数のモジュール同士を組み合わせる構成であることから，全体での最適化が困難であり，性能向上の余地が残されていた．近年では，音声特徴量の系列データから単一 DNN のみでテキストを出力する E2E 音声認識システムが盛んに研究されており，従来型音声認識システムよりも総合的に高い性能を得られる旨が報告されている．

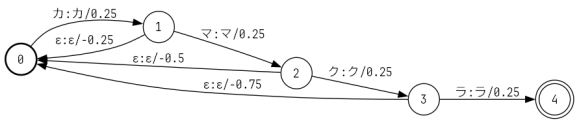

図 1 WFST の一例
Fig. 1 Example of WFST

代表的な E2E 音声認識の手法として, connectionist temporal classification (CTC) [1] を用いる手法や，attention 機構を有する encoder-decoder モデル (AED) を用いる手法 [6], [7] などが知られている．近年では，Transformer [8] 構造や，Transformer と 1 次元畳み込みを組み合わせた Conformer [9] 構造を採用する AED 方式および，CTC 方式と AED 方式を組み合わせる hybrid CTC/attention 機構を用いた音声認識方式 [2] も提案されている．また，N-gram や recurrent neural network (RNN) に基づく言語モデルを併用することで，音声認識の精度を高める手法も利用される [10]．これらの音声認識モデルは，open source software (OSS) のツールキットである ESPnet [11] 等を利用することで試用できる．

E2E 音声認識は，概して階層型音声認識よりも高い認識性能を示す一方で，発音辞書が独立したモジュールでないため，認識対象語彙の追加登録が容易でない．一例として，登録したい語彙を含むデータを収集し，音声認識 DNN モデルを再学習させることで，語彙の追加登録を実現できるものの，GPU 等の演算資源と音声データを要する．また，誤ったスペリングを修正するための言語モデルを後段に接続することで，特殊な語彙を認識できるようにする手法 [12] も提案されている．

2.2 WFST

重み付き有限状態トランスデューサ (weighted finite-state transducer: WFST) は，入力された系列情報に対して，それに対応する出力系列と，入出力系列の尤もらしさを示す重み (スコア) 情報を出力する計算モデルであり，階層型音声認識で利用されている [13]．図 1 は，WFST の一例を示す．この WFST は，「カマクラ」が入力された場合に，スコアとして 1.0 を出力し，それ以外の系列が入力された場合，スコアとして 0.0 を出力する．階層型音声認識で用いられる hidden Markov model (HMM)・発音辞書・N-gram 言語モデルや，E2E 音声認識における CTC は，WFST として表現可能である．また，ある WFST から出力された系列を，異なる WFST の入力とする場合，両者を合成して 1 つの WFST として表現したり，最適化演算を加えたりできる．これを用いることで，階層型音声認識は単一の WFST として効率的に実現できる．

2.3 Contextual Biasing を用いた語彙登録

Contextual biasing とは，音声認識結果候補（仮説）のテキストに特定の語彙が含まれる場合，当該仮説のスコアにバイアスを付与することで，特定語彙を含む結果を出力させやすくする手法である．例えば，「電機メーカー」と話している音声を認識させたい場合，しばしば仮説「電気メーカー」のスコア（対数尤度）が，仮説「電機メーカー」のスコアを上回り，誤って「電気メーカー」が認識結果となる場合がある．ここで，仮説中に「電機メーカー」が現れたら，スコアにバイアスを加算するようにすることで，「電機メーカー」を認識結果とできるようになる．バイアスの加算処理には，2.2 節で述べた WFST を用いることができる．例えば，図 1 に示す WFST は，語彙「カマクラ」を含む仮説に対しバイアススコア 1.0 を加算するバイアス WFST として利用できる．

N-gram 言語モデルと，contextual biasing を併用する E2E 音声認識は，以下のように定義される目的関数 $\mathcal{L}_{\text{recog}}$ を最大とするような系列を，ビームサーチ等のアルゴリズムを利用し探索する問題として定式化できる．

$$\mathcal{L}_{\text{recog}} = \log P_{\text{E2E}}(\mathbf{y}|\mathbf{x}) + \alpha(\log P_{\text{N-gram}}(\mathbf{y}) + \beta \text{Bias}(\mathbf{y})) \quad (1)$$

ここで，α，β は，それぞれ言語モデル及び contextual biasing の重みを示すハイパーパラメタであり，$\mathbf{x}$ は入力音声から得られる特徴量，$\mathbf{y}$ は音声認識の仮説テキスト，$P_{\text{E2E}}(\mathbf{y}|\mathbf{x})$ は E2E 音声認識の尤度，$P_{\text{N-gram}}(\mathbf{y})$ は N-gram 言語モデルの尤度，$\text{Bias}(\mathbf{y})$ はバイアススコアをそれぞれ示す．バイアススコア $\text{Bias}(\mathbf{y})$ は，以下のように表現できる．

$$\text{Bias}(\mathbf{y}) = \sum_k \#\mathbf{w}_k\{\mathbf{y}\} \cdot b_{\mathbf{w}_k} \quad (2)$$

ここで，k はバイアス付与の対象とする語彙のインデクス，$\mathbf{w}_k$ は k 番目の語彙，$\#\mathbf{w}_k\{\mathbf{y}\}$ は仮説 $\mathbf{y}$ に含まれる語彙 $\mathbf{w}_k$ の個数，$b_{\mathbf{w}_k}$ は語彙 $\mathbf{w}_k$ に加算するバイアス，をそれぞれ示す．

なお，語彙ごとのバイアス $\mathbf{w}_k$ を決定する方法はいくつか考えられるが，学習データに出現しやすい語については小さな値となり，反対に出現頻度が低い語については大きな値となる性質を有することが好ましい．これを満たす一例として，N-gram 言語モデルの符号を反転させ得られる値

$$b_{\mathbf{w}_k} = -\log P_{\text{N-gram}}(\mathbf{w}_k) \quad (3)$$

を利用できる．

近年では，WFST を利用せず，E2E 音声認識に登録対象語彙より生成された embedding の入力部を設けることで語彙登録を実現する手法 [4] も提案されている．

3. 提案手法

3.1 カタカナ・ひらがな表記を用いた語彙登録

従来の語彙登録手法 (contextual biasing) の課題点として，

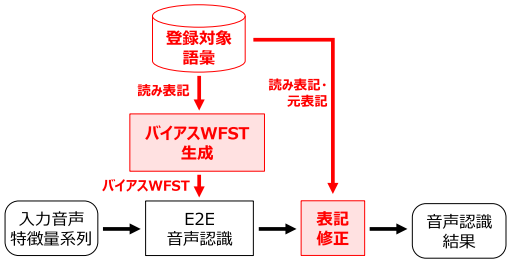

図 2 提案手法を示す概念図
Fig. 2 Conceptual diagram of proposed method

E2E 音声認識モデルの学習データに含まれる表記の語彙に対しては有効に機能する一方で，そうでない語彙に対しては効果が不十分，といったものがある．これは，学習データに出現しない語彙を含む仮説のスコアが低すぎるため，contextual biasing を利用し，そのような仮説にバイアスを付与したとしても，ビームサーチ時の候補から当該仮説が外れてしまうためである．E2E 音声認識モデルを学習させる際に，学習データ中の固有名詞を予め収集した同発音の異なる語彙へとランダムに置換することで，学習データに存在しない表記を減らし，上述の課題を緩和する手法も提案されているが [14]，登録対象としたい語彙をモデルの学習時に全て列挙しておくことは困難である．

この課題を解決するための策として，より学習データに高頻度で現れる他の表記を利用し，語彙登録に用いる方法が考えられる．図 2 は，提案手法を示す概念図である．表音文字であるカタカナ及びひらがなは，日本語の文章で多数登場するほか，固有名詞や熟字訓など難読語彙の読み方を示す方法としても利用されており，登録対象語彙の本来の表記に代わって，語彙の登録に用いることが適切であると考えられる．当該表記で語彙を登録した後，音声認識結果に対する後処理において当該表記を元表記へと置換し戻すことで，学習データに含まれない語彙の認識が困難である課題を緩和できると期待できる．

3.2 音声認識モデルの学習方法改良

3.1 節で述べた，カタカナ・ひらがな表記による語彙登録は，学習データに出現しない語彙に対して有効である可能性が高い．しかし，学習データ中に当該表記の出現頻度が低い場合には，必ずしも十分な効果が得られないことがある．この課題を解決するため，E2E 音声認識のモデルに対して，かな表記を用いた語彙登録を有効に機能させるためのファインチューニングを追加で実施する．具体的には，ベースとなる E2E 音声認識のモデルを学習させる際に，一部の語彙を，元の漢字かな交じり表記から，カタカナ・ひらがなによる表記へと改めることで，学習データ内に含まれるカタカナ・ひらがな表記を増やす．図 3 は，当該前処理方法を示す概念図である．また，図 4 は，音声認識モデルの学習方法およびファインチューニング方法を示す概念

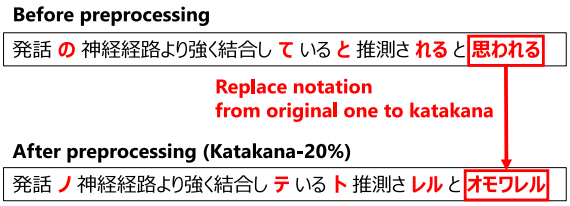

図 3 学習用テキストに対する前処理を示す概念図
Fig. 3 Conceptual diagram of preprocessing for training text

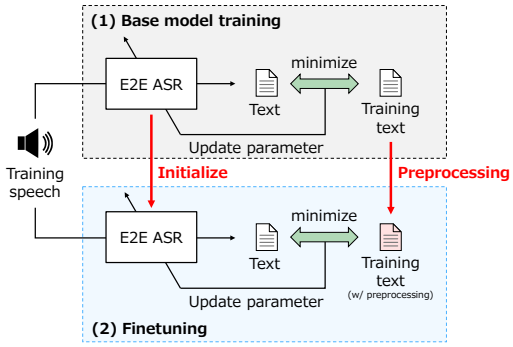

図 4 E2E 音声認識モデルの学習方法
Fig. 4 Training procedure of E2E ASR model

図である．

4. 評価実験

4.1 実験設定

E2E 音声認識モデルの学習には，OSS の音声認識ツールキットである ESPnet を利用した．学習データとして，日本語話し言葉コーパス (corpus of spontaneous Japanese: CSJ) に収録されている，音声データおよび書き起こしテキスト (約 550 時間) を使用した．音声認識モデルの学習手法として，CTC 方式と AED 方式を併用する hybrid CTC/attention [2] 構造を採用した．DNN 構造として，エンコーダに 12 層の Conformer [9] を，デコーダーに 8 層の Transformer を採用した．音声認識モデルの学習時は，CTC 方式および AED 方式の誤差重みづけをそれぞれ 0.3・0.7 に設定した．また，音声認識モデルによる推論の際は，AED 方式による誤差は使用せず，CTC 方式による誤差，N-gram 言語モデル，及び登録語彙バイアスの重み付き和である式 (1) を小さくするような系列を，ビームサーチを用いて探索した．ここで，ビーム幅は 40，重みパラメタ α は 3.0 に設定し，重みパラメタ β は 1.0 から 3.0 まで 0.25 刻みで変化させ，挙動の変化を確認した．音声認識モデルの学習では，データ拡張として SpecAugment [15] 及び speed perturbation（0.9 倍・1.0 倍・1.1 倍）を実施した．

ベースとなる音声認識モデルは，CSJ コーパスを利用して 50 エポック学習させた後，バリデーションデータの誤差が小さい順に上位 10 個のモデルに対して model averaging を適用することで作成した．提案手法におけるファインチューニングの際は，書き起こしテキストに出現する単語を，ランダムに一定の割合で，カタカナ又はひらがな表記へと置換した．また，前述のベース音声認識モデルを 10 エポック学習させた後，バリデーションデータの誤差が小さい 3 個のモデルを抽出し，重みの平均を取る処理を実施した．評価では，ファインチューニングの方法が異なる 3 種類のモデルを利用した．**モデル 1** はファインチューニング実施前のモデルであり，**モデル 2** は，学習データのテキストに対し，単語 5%をカタカナへ置換する前処理を適用してからファインチューニングを実施したモデル，**モデル 3** は，単語 5%をひらがなへ置換する前処理を適用してからファインチューニングを実施したモデルである．

音声認識やバイアス算出に利用する N-gram 言語モデルは，ファインチューニング時に利用するテキストデータを用いて学習した．WFST を用いたデコーディング処理には，OSS ツールキットである OpenFST [16] を使用した．

上述の 3 種類のモデルを利用し，CSJ コーパスの評価セット (eval1/eval2/eval3) を認識させた場合の文字誤り率 (character error rate: CER) を表 1 に示す．表より，ファインチューニングの前後において，語彙登録を利用しない通常の音声認識性能に大きな変化がないことを読み取れる．

表 1 CSJ コーパス評価セットにおける音声認識性能 (CER) [%]
Table 1 ASR performance (CER) in CSJ eval sets [%]

Model ID	Dataset for finetuning	Eval1↓	Eval2↓	Eval3↓
1	–	**4.4**	**3.2**	**3.5**
2	5% Katakana	4.6	3.3	3.6
3	5% Hiragana	4.5	3.3	3.6

4.2 評価 1：カタカナ語彙の登録性能評価

まず，3.2 節で述べた学習方法の改良によって，かな表記による語彙登録の性能が改善するか否かの評価を実施した．評価データは，新聞記事読み上げ音声コーパス (JNAS) から，以下の手順で作成した．まず，CSJ コーパス・JNAS コーパスの書き起こし文を，MeCab を用いて分かち書きした．次に，JNAS コーパスの書き起こし文のうち，CSJ コーパスに含まれないカタカナの名詞を含む音声データのみを抽出した (2717 件)．最後に，CSJ コーパスに含まれない，カタカナ 5 文字以上の名詞リストを，語彙登録の対象とした (586 語)．登録対象の語彙として，例えば「ソリブジン」「ベンヤヒア」「マックファクハー」「ツムアインホルン」などが抽出された．今後，本評価データを**評価セット 1** と呼ぶ．語彙を登録しない**条件 1A** と，語彙を登録する**条件 1B** について，結果を比較した．以後，例えばモデル 2 を用いて条件 1A の評価を実施した場合には，条件 1A-2 のように表記する．

評価指標として，音声認識結果の文字誤り率を示す **CER**，登録対象の語彙が正しく出現している割合を示す **KW–cor**，

表 2 評価セット 1 における，バイアス重み β と語彙登録性能の関係．音声認識モデルとして，モデル 2 (5%カタカナ化データでファインチューニング済) を利用している．

Table 2 Relationship between keyword bias parameter β and keyword enrollment performance on evaluation dataset 1. We use 'Model 2', which was trained with preprocessed text dataset by replacing 5% words from original notation to katakana.

Keyword enrollment	β	CER↓ [%]	KW–cor↑ [%]	KW–ins↓ [%]	KW–del↓ [%]	KW–F1↑
-	(0.00)	15.3	35.4	**1.7**	64.6	0.351
✓	1.00	14.0	71.8	2.0	28.2	0.711
✓	1.25	13.7	76.5	3.0	23.5	0.754
✓	1.50	13.4	83.4	7.1	16.6	0.806
✓	1.75	**13.2**	89.1	12.0	10.9	**0.840**
✓	2.00	13.7	92.3	24.4	7.7	0.822
✓	2.25	15.2	93.1	52.2	**6.9**	0.739
✓	2.50	18.8	**93.0**	115.7	7.0	0.589
✓	2.75	26.4	92.2	234.0	7.8	0.425
✓	3.00	40.2	90.3	430.6	9.7	0.286

登録対象の語彙が出現すべきでない箇所に誤って出現した割合を示す **KW–ins**，登録対象の語彙が出現すべき箇所に出現しない割合を示す **KW–del**，登録対象の語彙について，再現率と適合率を計算し，更に調和平均を求めた **KW–F1** の 5 種類を用いた．なお，KW–cor 指標は

$$\text{KW–cor} = \frac{\sum_{l,k} \min\left(n_{l,k,\text{recog}}, n_{l,k,\text{gt}}\right)}{\sum_{l,k} n_{l,k,\text{gt}}} \tag{4}$$

のように計算される．ここで，l は評価データのインデクス，k は登録する語彙のインデクスであり，$n_{l,k,\text{gt}}$ は l 番目の正解テキスト中に出現する k 番目の語彙の数，$n_{l,k,\text{recog}}$ は l 番目の認識結果テキスト中に出現する k 番目の語彙の数である．また，KW–ins 指標は

$$\text{KW–ins} = \frac{\sum_{l,k} \max\left(0, n_{l,k,\text{recog}} - n_{l,k,\text{gt}}\right)}{\sum_{l,k} n_{l,k,\text{gt}}} \tag{5}$$

のように計算される．さらに，KW–del = 1 – KW–cor が成り立つ．

表 2 は，カタカナ 5%置換データでファインチューニングされたモデル 2 を使用した場合における，語彙登録性能とバイアス重み β の関係を示す表である．表より，バイアス重み β の設定によって，CER 指標は 13.2%から 40.2%の範囲で，KW–F1 指標は 0.286 から 0.840 の範囲で，それぞれ変動することが読み取れる．

本結果より，登録語彙の認識性能は，バイアス重み β の設定に強く依存するため，適切な値を設定することが重要と考えられる．

表 3 は，4.1 節で言及した 3 種類のモデルのうち，ベースとなるモデル 1 と，5%カタカナ置換データでファインチューニング済のモデル 2 を利用し，語彙登録の性能を評価した結果である．なお，重みパラメタ β として，表 2 において最良と判断された 1.75 を採用する．語彙登録を実施しない条件 1A-1 と，語彙登録を実施する条件 1B-1 とを

Ground truth　: 女性は**チェンセージュ**さん
Condition 1A-1: えー女性は**チェ聖子**さん
Condition 1B-1: えー女性は**チェント女**さん
Condition 1B-2: えー女性は**チェンセージュ**さん

図 5 評価 1 における音声認識結果のサンプル
Fig. 5 Examples of ASR output in Evaluation 1

比較すると，CER は 1.2 ポイント，KW–F1 指標は 0.34 の改善がみられる．更に，ファインチューニングを実施しないモデルを使用する条件 1B-1 と，5%カタカナ化データでファインチューニングしたモデルを使用する条件 1B-2 とを比較すると，CER は 1.3 ポイント，KW–F1 指標は 0.14 の改善がみられる．図 5 に，語彙登録の有無による認識結果の変化例を示す．ここで，「チェンセージュ」は語彙登録の対象である．条件 1A-1 および条件 1B-1 では「チェンセージュ」を正しく認識できていないが，条件 1B-2 では，「チェンセージュ」を正しく認識できている．

条件 1A-1 と条件 1B-1/1B-2 の結果を比較し，CER・KW–F1 指標とも後者の方が良いことから，語彙登録の有効性を確認できる．また，条件 1B-1 と条件 1B-2 を比較した場合に，両指標とも後者の方が良いことから，3.2 節に示す E2E 音声認識モデルのファインチューニングを実施することで，カタカナ表記による語彙の登録がより有効に機能するようになった，と考えられる．

4.3 評価 2：漢字かな交じり語彙の登録性能評価

次に，3.1 節で述べた，かな表記を用いた語彙登録の後，出力テキスト中のかな表記を元の表記に置換する手法の有効性を確認するための評価を実施した．評価データは，新聞記事読み上げ音声コーパス (JNAS) から，以下の手順により作成した．まず，CSJ コーパス・JNAS コーパスの書き起こしテキストを，MeCab を用いて分かち書きした．次に，JNAS コーパスの書き起こし文のうち，CSJ コーパスに含まれない 5 モーラ以上の名詞を含む音声データのみを抽出した (793 件)．最後に，CSJ コーパスに含まれない，5 モーラ以上の名詞リストを，語彙登録の対象とする (187 語彙)．今後，本評価データを**評価セット 2** と呼ぶ．

評価指標として，4.2 節で利用したものと同一のものを使用した．また，以下の 5 条件について比較を実施した．

- **条件 2A**: 語彙登録を実施しない場合
- **条件 2B**: 元の表記で，語彙を登録する場合
- **条件 2C**: カタカナ表記で，語彙を登録する場合
- **条件 2D**: ひらがな表記で，語彙を登録する場合
- **条件 2E**: 条件 2B と条件 2C/2D の結果を比較し，KW-F1 指標が良くなるよう，語彙ごとに登録用表記を選ぶ場合

4.2 節と同様，例えば条件 2A のもとでモデル 3 を使用して

表 3　評価セット 1 における語彙登録性能
Table 3　Keyword enrollment performance on evaluation dataset 1

ID	Model ID	Keyword enrollment	β	CER↓ [%]	KW–cor↑ [%]	KW–ins↓ [%]	KW–del↓ [%]	KW–F1↑
1A-1	1	-	(0.00)	15.7	36.1	**1.6**	63.9	0.358
1A-2	2	-	(0.00)	15.3	35.4	1.7	64.6	0.351
1B-1	1	✓	1.75	14.5	72.1	6.8	27.9	0.697
1B-2	2	✓	1.75	**13.2**	**89.1**	12.0	**10.9**	**0.840**

Ground truth　: 壮太郎は兄夫婦の強引な勧めで里子と見合いする
Condition 2A-1: 総太郎は兄夫婦の強引な勧めで佐藤子と見合いする
Condition 2B-1: 宝太郎は兄夫婦の強引な勧めで佐藤子と見合いする
Condition 2C-2: 壮太郎は兄夫婦の強引な勧めで佐藤子と見合いする

図 6　評価 2 における音声認識結果のサンプル
Fig. 6　Examples of ASR output in Evaluation 2

評価する場合を条件 2A-3 と表記する．

表 4 に，評価 2 の結果を示す．ここで，重みパラメタ β を 2.00 とした場合の結果を報告している．ファインチューニング実施前のモデル 1 を使用する場合，語彙登録を実施しない条件 2A-1 と，原表記で語彙を登録する条件 2B-1 を比較すると，CER は 3.1 ポイント，KW–F1 指標は約 0.50 の改善がみられる．しかし，原表記で語彙を登録する条件 2B-1 と，カタカナ表記で語彙を登録する条件 2C-1，及びひらがな表記で語彙を登録する条件 2D-1 を比較すると，いずれの場合も，CER・KW–F1 指標とも悪化している．この一方で，モデル 1 へ原表記の語彙を登録する条件 2B-1 と，ファインチューニング後のモデル 2 へカタカナ表記の語彙を登録する条件 2C-2 を比較すると，CER は 1.5 ポイント，KW–F1 指標は約 0.14 の改善がみられる．条件 2B-1 と，モデル 3 へひらがな表記の語彙を登録する条件 2D-3 を比較した場合も同様に，CER は 0.9 ポイント，KW–F1 指標は約 0.19 の改善がみられる．条件 2B-2 と条件 2C-2 の結果をもとに，語彙の登録に原表記・カタカナ表記のいずれを用いるか決定する条件 2E-2 では，常にカタカナ表記を用いる条件 2C-2 と比較し，CER が 0.4 ポイント，KW–F1 指標が約 0.12 改善する．さらに，条件 2B-3 と条件 2D-3 の結果をもとに，語彙の登録に原表記・ひらがな表記のいずれを用いるか決定する条件 2E-3 では，常にひらがな表記を用いる条件 2D-3 と比較し，CER が 0.7 ポイント，KW–F1 指標が約 0.09 改善する．図 6 に，音声認識結果の結果サンプルを示す．語彙登録を利用しない条件 2A-1 や，原表記で語彙を登録する条件 2B-1 では，「壮太郎」を正しく認識できていない一方で，5%カタカナ化データで音声認識モデルをファインチューニングし，カタカナ表記で語彙を登録する条件 2C-2 では，「壮太郎」を正しく認識できている．

条件 2B-1 と，条件 2C-1 及び条件 2D-1 の比較より，通常の漢字かな交じりの語彙を登録する場合において，表音文字を用いた表記による語彙の登録は，有効に機能しないことを読み取れる．この一方で，条件 2B-1 と，条件 2C-2 及び条件 2D-3 の比較より，3.2 節で示す音声認識モデルのファインチューニングの実施によって，表音文字表記による語彙の登録が有効に機能している．これらの結果から，3.1 節で述べたカタカナ・ひらがな表記による語彙登録の手法は，3.2 節のファインチューニングと組み合わせることで，有効に機能する，と言える．さらに，条件 2C-2 と条件 2E-2，条件 2D-3 と条件 2E-3 の結果を比較することで，語彙ごとに原表記・表音文字表記のいずれが良いかが異なっており，適切な表記を選ぶ手法があれば，語彙登録の性能を更に高められる可能性が示唆される．具体的な表記の決定方法として，本評価のように実験的に決定する方法のほか，登録対象の語彙を入力すると，いずれの表記を用いるのが良いか出力する機械学習モデルを用いる方法などが考えられ，今後検討の余地がある．

5. むすび

本稿では，日本語 E2E 音声認識モデルに対し，contextual biasing に基づく語彙登録を有効に機能させるための改良手法について検討した．特に，学習データに含まれない表記を含む語彙を登録する場合，カタカナやひらがななど表音文字による表記で語彙を登録し，認識結果テキストに対する後処理で，表音文字から元の表記へと戻す手法を提案した．更に，カタカナ表記やひらがな表記による語彙登録をより有効に機能させるため，E2E 音声認識モデルの学習データに対し，表音文字による表記を増やす前処理を実施した上でファインチューニングする改良手法を提案した．実験的な評価を通じ，表音文字表記を用いた語彙の登録及び，モデル学習方法改良がいずれも有効に機能していることを確認した．今後の課題として，本稿では評価の対象としていない，3 モーラ程度の短い単語における語彙登録性能の改善や，重みパラメタ β の適切な決定方法などが挙げられる．

参考文献

[1] Graves, A. and Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks, *Proc. International Conference on Machine Learning*, pp. 1764–1772 (2014).
[2] Watanabe, S., Hori, T., Kim, S., Hershey, J. R. and Hayashi, T.: 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).
[3] Seide, F., Li, G. and Yu, D.: Conversational speech transcription using context-dependent deep neural networks, *Proc.*

表 4　評価セット 2 における語彙登録性能

Table 4　Keyword enrollment performance on evaluation dataset 2

ID	Model ID	Keyword enrollment	β	CER↓ [%]	KW–cor↑ [%]	KW–ins↓ [%]	KW–del↓ [%]	KW–F1↑
2A-1	1	-	(0.00)	17.3	18.1	**0.0**	81.9	0.181
2A-2	2	-	(0.00)	17.8	17.4	0.2	82.6	0.174
2A-3	3	-	(0.00)	18.2	17.1	0.2	82.9	0.170
2B-1	1	Original	2.00	14.2	71.0	6.5	29.0	0.687
2B-2	2	Original	2.00	13.7	81.8	13.4	18.2	0.766
2B-3	3	Original	2.00	14.2	84.2	10.5	15.8	0.800
2C-1	1	Katakana	2.00	16.6	26.3	0.2	73.7	0.263
2C-2	2	Katakana	2.00	**12.7**	84.4	3.1	15.6	0.831
2D-1	1	Hiragana	2.00	16.1	32.3	**0.0**	67.7	0.323
2D-3	3	Hiragana	2.00	13.3	**89.0**	2.6	**11.0**	**0.878**
2E-2	2	Better notation (Orig/Kata)	2.00	12.3	93.0	3.2	7.0	0.915
2E-3	3	Better notation (Orig/Hira)	2.00	12.6	98.0	2.4	2.0	0.968

Interspeech, pp. 437–440 (2011).

[4] Pundak, G., Sainath, T. N., Prabhavalkar, R., Kannan, A. and Zhao, D.: Deep context: End-to-end contextual speech recognition, *Proc. IEEE Spoken Language Technology Workshop*, pp. 418–425 (2018).

[5] Kojima, A.: A study of biasing technical terms in medical speech recognition using weighted finite-state transducer, *Acoustical Science and Technology*, Vol. 43, No. 1, pp. 66–68 (2022).

[6] Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y.: Attention-based models for speech recognition, *Proc. Advances in Neural Information Processing Systems*, Vol. 28, pp. 1–9 (2015).

[7] Chan, W., Jaitly, N., Le, Q. and Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4960–4964 (2016).

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is all you need, *Proc. Advances in Neural Information Processing Systems*, Vol. 30, pp. 1–11 (2017).

[9] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. et al.: Conformer: Convolution-augmented Transformer for speech recognition, *Proc. Interspeech*, pp. 5036–5040 (2020).

[10] Karita, S., Soplin, N. E. Y., Watanabe, S., Delcroix, M., Ogawa, A. and Nakatani, T.: Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration, *Proc. Interspeech*, pp. 1408–1412 (2019).

[11] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A. and Ochiai, T.: ESPnet: End-to-end speech processing toolkit, *Proc. Interspeech*, pp. 2207–2211 (2018).

[12] Guo, J., Sainath, T. N. and Weiss, R. J.: A spelling correction model for end-to-end speech recognition, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5651–5655 (2019).

[13] Mohri, M., Pereira, F. and Riley, M.: Weighted finite-state transducers in speech recognition, *Computer Speech & Language*, Vol. 16, No. 1, pp. 69–88 (2002).

[14] Zhao, D., Sainath, T. N., Rybach, D., Rondon, P., Bhatia, D., Li, B. and Pang, R.: Shallow-fusion end-to-end contextual biasing, *Proc. Interspeech*, pp. 1418–1422 (2019).

[15] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D. and Le, Q. V.: SpecAugment: A simple data augmentation method for automatic speech recognition, *Proc. Interspeech*, pp. 2613–2617 (2019).

[16] Riley, M., Allauzen, C. and Jansche, M.: OpenFst: An open-source, weighted finite-state transducer library and its applications to speech and language, *Proc. Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 9–10 (2009).