合成単語データを用いた低コスト高品質な音声認識のドメイン適応

小松秀輔 [1,2,4]　大西一誉 [1,2,4]　田中康紀 [1,2]　金道鉉 [1,2]　吉野幸一郎 [3,2,4]

[1] *mocomoco* 株式会社　[2] 奈良先端科学技術大学院大学
[3] 東京科学大学　[4] 理化学研究所ガーディアンロボットプロジェクト
{komatsu,onishi,tanaka,kim}@mocomoco.ai
koichiro.yoshino@riken.jp

概要

深層学習により高い性能を発揮している汎用音声認識モデルは，ドメインに適応した出力を行うにはファインチューニングによる追加学習を必要としている．先行研究では大規模言語モデルと制御音声合成を用いたデータ拡張による音声認識モデルのファインチューニング手法が提案されているが，コストと未知語の学習の面で課題が残る．本研究では単語のみの制御音声合成を結合した音声データを用いた低コストかつ，高品質なファインチューニング手法を提案する．実験では既存手法と比較して，提案手法によるファインチューニングは 1/3 程度の学習コストでより多くの未知語を学習することができることを示した．

1　はじめに

会話における同音異義語や外来語・固有表現等のテキスト上での表記は，会話が行われるドメインや文脈に依存して異なる場合がある．そのため，音声認識モデルはそれが使用されるドメインに適応されることが望ましい．たとえば医療や法律などの専門分野では，固有の略語や特有の言い回しが多用されるため，正しく認識できなければ実運用での信頼性を損なう可能性がある．また，会社やチームなどのより詳細なドメインにおいては書き起こし結果として出力したい固有名詞や所属人物などの単語が学習済み音声認識モデルの学習データに含まれない場合もある．

深層学習や自己注意機構などの手法により，音声認識モデルは高性能な音声認識精度を実現した [1][2]．しかし，汎用モデルとしての音声認識モデルは多様な領域の音声を認識可能である反面，特定の専門分野における用語の認識精度に課題を抱える場合がある．特定のドメインに適応した音声認識モデルの研究はこれまでに多く進められてきた [3][4]．ドメイン適応の手法として従来はそのドメインにおける実際の音声コーパスやクラウドソーシングを用いたファインチューニングが行われてきた [5][6]．しかし，これらのデータを収集するには多くの費用と時間がかかり，この手法による学習は容易ではない．

そこで，低コストなドメイン適応の手法として大規模言語モデルと制御音声合成を用いたデータ拡張によるファインチューニングが提案されている [7]．ここでは学習したい単語（以下，ターゲット単語）を含む文を大規模言語モデルを用いて生成し，その文を制御音声合成により音声化して学習データを拡張する．しかし，この手法は学習データ全体に占めるターゲット単語の割合が小さく，学習が非効率であると考えられる．また，文脈を考慮した学習をするという近年の音声認識モデルの持つ性質により，学習に使用した例文の文脈を過学習し，未知の音声データでは学習した単語を書き起こすことができないことが懸念される．過学習を低減するためにはさらに文のバリエーションが必要となり学習コストも膨らむ．

本研究では単語のみの合成音声を結合した学習データを用いることで，既存手法よりも効率のよいターゲット単語の学習を提案する．これにより，従来の手法で課題となっていた文脈や隣接単語の影響を最小限に抑えつつ，専門用語や頻出単語を効率よく増強できる．実験では提案手法が (1) 既存手法と比較してより多くのターゲット単語を音声認識結果

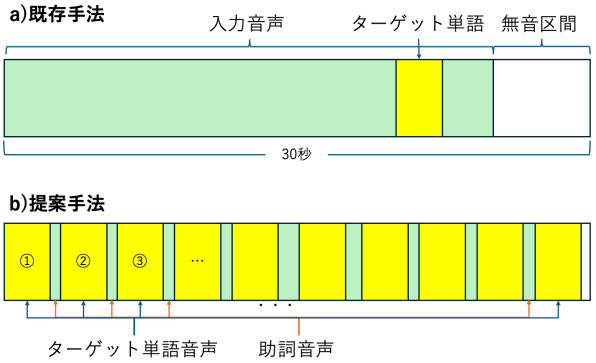

図 1　データ作成手法の概要

に反映できること，(2) 学習コストの面でも大規模言語モデルや音声合成の呼び出しコストに加え、学習に必要なステップを低減し、低コストな学習を可能にすることを示した．

2　データ作成手法

本章では音声認識モデルのファインチューニングに用いるデータの作成手法を説明する．本手法ではターゲットとなる単語の合成音声と助詞の合成音声を組み合わせることで，効率的かつ柔軟に学習データを構築することを主眼としている．

本手法における音声データの構造は図 1 に示す通りである．まず，音声認識モデルのドメイン適応が必要な分野や使用環境に合わせてターゲット単語を選定し，それらの単語のみを発話する合成音声を生成する．加えて，ターゲット単語間の境界を明確にするために単語間に挿入する助詞の合成音声を生成する．格助詞，係助詞，並列助詞，副助詞を選定し，合計 20 種類の合成音声を生成する．

生成した単語音声と助詞音声を組み合わせ，結合することで効率的にターゲット単語を学習できるデータセットを作成する．各ターゲット単語の合成音声に対して，助詞の合成音声を挟みながら連続的につなぎ合わせる手法をとることで，一般的な語連接らしさを担保しつつ，一つの音声でターゲット単語を可能な限り多く参照できるデータを作成することが可能となる．一度の入力によって学習可能なターゲット単語が最大となるよう，ベースモデルがが一つの入力系列として扱う音声の長さまで結合を行い，一点の音声データを作成する．出力がターゲット単語の学習時に隣接している単語や音声データ上の位置の影響を抑制するために，上記の結合手順をターゲット単語の順番をシャッフルして任意の回数繰り返す．

3　実験

実験では本手法によるファインチューニングと LLM によって生成した文によるファインチューニングを行い，音声認識精度と学習にかかるコストの観点で比較をする．

3.1　ベースモデル

実験では OpenAI が公開している Whisper(openai/whisper-large-v3-turbo)[1] をベースモデルとして使用する．Whisper は深層学習を用いた汎用音声認識モデルであり，多言語による音声認識や翻訳を可能にしている．Whisper は 30 秒を一つの入力系列として扱い，30 秒に満たない音声はパディングされる. 音声のメルスペクトログラムを入力として用い，エンコードした埋め込みを交差注意機構によって各デコーダ層に入力し，最終的な出力を行うモデルである．

3.2　ターゲット単語

実験に用いるドメインとして日本の政治・安全保障に関する用語，人名を 1490 語を収集した．これらの単語ごとにターゲット単語が含まれる文を OpenAI の GPT-4o[8] を用いて生成し，Google text-to-speech API[9] を用いて合成音声を作成した．これらの合成音声のうち，ベースモデルが正しく書き起こすことができなかった単語 (469 語) をファインチューニングのためのターゲット単語として用いる．

3.3　実験設定

実行環境　学習には Google Cloud の Nvidia L4 Tensor core GPU を用いる．

ハイパーパラメータ　学習時のハイパーパラメータは学習率 1e-5，バッチサイズ 8，最大ステップ 500 である．学習に用いるハイパーパラメータは全てのモデルに対して共通とする．

最良モデル選定　学習データからの 10%をランダムに抽出し，検証データとして用いる．加えて，過学習によるターゲット単語以外の出力の悪化を検知するために文から作成した合成音声を学習データから抽出したデータと同数，検証データに加える．この音声はドメインとは関係のない日常会話の文である．20 ステップごとにモデルの検証を行い，CER

表 1　音声認識結果

学習手法	WER(%)	CER(%)	内容語 WER(%)	機能語 WER(%)	ターゲット単語正解率
base-Whisper	4.28	3.90	7.90	0.63	0
Word-1	1.30	1.19	2.33	0.68	78.67
Word-2	1.02	0.97	1.87	0.49	85.50
Word-3	1.02	0.86	1.89	0.56	86.14
Sentence-1	1.00	0.97	1.80	0.34	81.02
Sentence-2	1.17	0.97	2.00	0.40	79.10
Sentence-3	0.91	0.70	1.69	0.21	81.02

が最も低いモデルを最良モデルとして保存する．

モデル　評価を行うモデルとして，提案手法を用いた学習データはシャッフル回数が 1〜3 まで (各 27 個，54 個，80 個) データで学習したの三つのモデルを用意する．ここで，一つの音声ファイルに含まれるターゲット単語数の平均は 17．37 となっている．

比較に用いる既存手法として，LLM によって生成した文の合成音声データによってファインチューニング行ったモデルを用意する．条件を揃えるために比較用モデルについても異なるプロンプトによる文の生成を用いたデータ拡張を 1〜3 種類の三つのモデルを用意する．

各手法でのファインチューニングではエンコーダの重みを凍結し，デコーダのみファインチューニングを行う．

テストデータ　テストデータはターゲット単語選定時にベースモデルの Whiper がターゲット単語を書き起こすことができなかった音声とその書き起こしペアを用いる．

4　結果

4.1　音声認識精度

各モデルの音声認識結果について，WER，CER を算出する．加えて，提案手法によって作成された音声はその文構造が特殊であることを踏まえ，機能語，内容語ごとの WER を算出する．さらに，ドメイン適応がどの程度進んでいるかを測定するためにターゲット単語正解率を算出する．これは音声認識結果に元となるターゲット単語が含まれていた場合に正解とする．

モデル毎の各数値は表 1 に示す．提案手法は Word-1 モデルの内容語 WER を除く全ての項目においてベースモデルよりも改善している．また，Word-2 モデルと Word-3 モデルのターゲット単語正解率は他のモデルを大きく上回っている．

最も多くの項目において最良の数値を示しているのは Sentence-3 モデルである．しかし，他のモデルと比べてターゲット単語正解率は大きく向上していない．このことから，Sentence-3 モデルにおけるこれらのスコア向上は，文に含まれるターゲット単語以外の単語の出現傾向を学習したことに起因すると考えられる．

また，全てのモデルを比較したときに提案手法によって学習したモデルの機能語 WER は既存手法によって学習したどのモデルよりも高い．ただし，その値はいずれも 1%以下と非常に軽微であり，また base-Whisper と同等かそれより改善しているため，音声認識を活用したアプリケーションに与える影響は少ない．

4.2　収束速度

表 2　学習収束までのステップ数

学習手法	ステップ数
Sentence-1	280
Sentence-2	300
Sentence-3	300
Word-1	60
Word-2	80
Word-3	80

学習が収束するまでの平均ステップ数は表 2 に示す通りである．提案手法は既存手法で学習した全てのモデルと比較して，1/3 以下のステップで収束している．これにより，Whisper のドメイン適応に要する計算資源は提案手法を用いることにより削減することができることがわかる．

また，データ拡張の回数が増加することで学習の収束が多少遅くなるが，顕著な差は見られない．これは学習する音声データが異なったとしてもそれに

表 3　追加のデータ拡張による音声認識結果

学習手法	WER(%)	CER(%)	内容語 WER(%)	機能語 WER(%)	ターゲット単語正解率
Sentence-3	0.91	0.70	1.69	0.21	81.02
Word-3	1.02	0.86	1.89	0.56	86.14
Word-5	0.98	0.84	1.72	0.95	89.55
Word-10	0.81	0.77	1.45	0.64	94.02

含まれるターゲット単語を参照した回数は変化しないことが要因であると考えられる．データ拡張回数が増加した時の収束速度の推移については 5.1 で説明する．

5　考察・今後の展望

5.1　データ拡張回数

本研究で扱った各手法によって全てのターゲット単語が正しく認識されるようになったわけではない．データ拡張によって反映されるターゲット単語は増加するものの線形に増加しているわけではないことから，全ての単語を完全に認識できるようにすることは難しい．実際に実験で行なったデータ拡張の回数に加えて，5 回，10 回のデータ拡張の回数を増やして学習を行なった．各モデルの音声認識結果における各スコアは 3 に示された通りである．また，ターゲット単語正解率の推移を表 2 に描画した．

実験で現れた傾向と同様にデータ拡張の回数を増やすことで WER，CER，ターゲット単語正解率などのスコアが改善している．加えて，Word-5 モデルでは上昇している機能語 WER が Word-10 モデルでは改善している．しかし，収束速度の推移 (図 2 を見るとはデータ拡張回数を増やすことで収束に必要なステップは大きく増加した．以上のことから，低コストかつ高性能な学習のために最適なデータ拡張の回数をターゲット単語の数や性質に応じて検証する必要があると考えられる．

5.2　統語構造的解釈と他言語への応用

提案手法ではターゲット単語の間に助詞を挟むことでターゲット単語間の境界を明確にしつつ，語連接の自然さを保証した．実験で扱ったターゲット単語が全て名詞であることと使用した助詞の特性から提案手法で作成した音声は長い名詞句であると捉えることができる．このことから，動詞や形容詞についても名詞句内に含まれうる形で結合を行えば，名詞のターゲット単語と同様に提案手法で学習をおこなることができると考えられる．

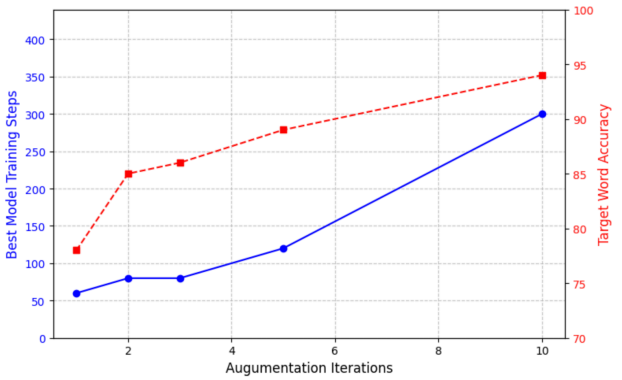

図 2　データ拡張回毎の収束速度とターゲット単語正解率

また，本研究では日本語でのドメイン適応を扱ったが，他言語についても名詞句を構成する形で単語を連結することで同様に学習を行うことができると考えられる．例えば，英語では結合部分に “in a” や “of the” などの前置詞+冠詞を挿入することで長い名詞句を構成することができる．ただし，本手法が日本語において有効であったのは，語順が与える意味への影響が小さいという特性によるものであるという解釈も可能である．他言語での本手法の適応については検証を要する．

6　おわりに

本研究では，低コストかつ学習効果が高い音声認識ファインチューニングのためのデータ作成手法を提案した．実験では LLM によって生成した文の合成音声でのファインチューニングと比較し 1/3 程度の学習ステップ数でより多くの単語を学習することができることを示した．

本研究ではターゲット単語の対象となるドメインやテストデータとして用いた音声の多様性が限られている点や実音声での検証を行っていない．本手法によるドメイン適応を行ったモデルが実用可能かを引き続き検証する必要がある．

参考文献

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In **International conference on machine learning**, pp. 28492–28518. PMLR, 2023.

[2] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. **Advances in neural information processing systems**, Vol. 28, , 2015.

[3] Juan Zuluaga-Gomez, Amrutha Prasad, Iuliia Nigmatulina, Seyyed Saeed Sarfjoo, Petr Motlicek, Matthias Kleinert, Hartmut Helmke, Oliver Ohneiser, and Qingran Zhan. How does pre-trained wav2vec 2.0 perform on domain-shifted asr? an extensive benchmark on air traffic control communications. In **2022 IEEE Spoken Language Technology Workshop (SLT)**, pp. 205–212. IEEE, 2023.

[4] Aviv Shamsian, Aviv Navon, Neta Glazer, Gill Hetz, and Joseph Keshet. Keyword-guided adaptation of automatic speech recognition. **arXiv preprint arXiv:2406.02649**, 2024.

[5] Erik Edwards, Wael Salloum, Greg P Finley, James Fone, Greg Cardiff, Mark Miller, and David Suendermann-Oeft. Medical speech recognition: reaching parity with humans. In **Speech and Computer: 19th International Conference, SPECOM 2017, Hatfield, UK, September 12-16, 2017, Proceedings 19**, pp. 512–524. Springer, 2017.

[6] Wael Salloum, Erik Edwards, Shabnam Ghaffarzadegan, David Suendermann-Oeft, and Mark Miller. Crowdsourced continuous improvement of medical speech recognition. In **Workshops at the Thirty-First AAAI Conference on Artificial Intelligence**, 2017.

[7] Hsuan Su, Ting-Yao Hu, Hema Swetha Koppula, Raviteja Vemulapalli, Jen-Hao Rick Chang, Karren Yang, Gautam Varma Mantena, and Oncel Tuzel. Corpus synthesis for zero-shot asr domain adaptation using large language models. In **ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 12326–12330. IEEE, 2024.

[8] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. **arXiv preprint arXiv:2410.21276**, 2024.

[9] Google Cloud. Text-to-Speech API Documentation, 2024. [Online; accessed 26-Dec-2024].