電話音声認識における特定の文脈へのドメイン適応のための合成音声によるデータ拡張

東佑樹 [1] 友松祐太 [1]
[1] 株式会社 AI Shift
{azuma_yuki,tomomatsu_yuta}@cyberagent.co.jp

概要

タスク指向型の音声対話システムでは予約受付や本人確認等のタスクにおいて，予めパターンの決まっている発話内容 (会員番号，商品 ID, etc.) を認識する場面がしばしば存在する．そこで本研究では，電話音声認識のタスクにおいて，音声合成により生成した音声を用いて転移学習を行い，特定パターンの認識性能の向上を試みた．提案手法では認識したい発話群の読みを正規表現により生成し，それぞれの読みに対応する音声を合成し，転移学習用のデータとする．汎用的な音声認識モデルとの認識性能の比較実験をおこなった結果，提案手法は汎用的なモデルを上回る性能を示した．

1 はじめに

タスク指向型の音声対話システムでは，一般にユーザの発話を音声認識 (Automatic Speech Recognition; ASR) によりテキストに書き起し，そのテキストをもとにインテント抽出やシステム応答文の選択などの対話戦略を行う，という構成が取られている．そのような音声対話システムでは ASR の認識誤りが後段のタスクの精度を著しく低下させる [1, 2] ため，高い認識性能が求められる．加えて，ドメインおよびタスク毎に出現する固有名詞や表現は変化するため，それらを適切に認識するために都度モデルの学習が必要となる．近年の End-to-End 型の汎用の ASR システムは非常に高い性能を発揮するようになった [3, 4] が，ドメイン固有の単語は誤認識しやすい傾向にある．ドメイン毎に転移学習用のデータを整備することでこの問題は解決できる可能性があるが，データの作成には多大なコストを必要とする．また，End-to-End 型の ASR は従来の ASR システムに見られるような音響モデルや言語モデル等が統合されているため，ドメインごとの言語モデルの切り替えが困難となる．

本研究では，予めパターンの決まっている発話内容を認識するタスクにおいて，多様なパターンの発話を考慮した転移学習の手法を提案する．具体的には認識したい発話群の読みを正規表現により生成し，各読みに対応する音声を合成音声により作成することで，ドメインごとの音声を収集することなくモデルの転移学習を実現する．実験では，番地表現を認識対象とし，自社で運用している音声対話システムを通して録音された音声を用いて, 汎用的に使用される ASR システムとの間で認識性能の比較を行う．これにより提案手法の有効性を示す．

2 関連研究

教師あり音声データを使用せずに ASR をドメイン適応させるためにはいくつかのアプローチが提案されている．

[5] や [6] ではドメイン適応の選択肢として合成音声を利用している．これらの手法では転移学習時に Encoder 層のパラメータを freeze させることがドメイン外のデータに対する認識性能を向上させるのに有効であることが報告されている．

End-to-End 型の ASR システムにおいて，外部言語モデルを学習時に効果的に利用する手法も提案されている．代表的な手法としては [7] や [8] などが挙げられる．これらはモデル内部に暗黙的に存在しうる言語モデルをタスクに応じて効果的に調整するための試みである．

また，近年では追加の学習を行うことなく専門用語などの低頻度語の認識精度を改善する手法も提案されており [9, 10, 11]，これらはデコーディング時に認識させたいキーワードの出力確率を引き上げるアルゴリズムを組み込むことで出力結果の調整を実現している．

本研究では，音響的な特性が事前学習データと大

きく異なる電話音声をターゲットドメインとしており，その特性を転移学習時に考慮するために，合成音声を利用するアプローチを採用する．

3 提案手法及び実験設定

3.1 ASR システム

本実験では，ベースとなる ASR として Conformer [3] を採用し，電話音声へのドメイン適応を考慮した ASR システムを探求する．モデルの学習には ESPnet[12] を利用し，パラメータ等の学習設定は基本的に CSJ のレシピ[1)] を踏襲している．このレシピは Hybrid CTC/Attention Architecture[13] を採用しており，学習時に Encoder の出力から CTC Loss(L^{ctc}) と Attention Decoder を通した損失 (L^{att}) を個別に計算し，両方のスコアの重み付き和により最終的な損失関数 L を決定する．

$$L = \alpha L^{ctc} + (1-\alpha)L^{att} \tag{1}$$

また，ESPnet には Shallow Fusion [14](以下，LM) の機能が用意されており，推論時にテキストコーパスのみから学習された言語モデルの出力確率と ASR の出力確率を組み合わせることができる．本実験では LM の有無による性能の比較も行う．より詳細なモデルの説明はレシピを参照されたい．

また，本実験では，汎用的な ASR システムとして Google Speech-to-Text[2)](以下， Google STT) を選択した．これをベースラインとして，後述の評価データを用いて電話音声の性能比較を行う．

3.2 データセット

本実験では，日本語話し言葉コーパス (Corpus of Spontaneous Japanese; CSJ)[15] を事前学習データとして採用した．転移学習時の学習データ及び検証データは Google Text-to-Speech[3)](以下， Google TTS) の API により生成された合成音声を用いた．評価データは, 自社で運用している自動音声対話サービス AI Messenger Voicebot[4)](以下，Voicebot) を通じて収集した電話音声を用いた．以下，各データの作成手順について記述する．

1） https://github.com/espnet/espnet/tree/master/egs2/csj/asr1
2） https://cloud.google.com/speech-to-text
3） https://cloud.google.com/text-to-speech
4） https://www.ai-messenger.jp/voicebot/

表 1 番地表現とそれに対応する読みの例．Google TTS にはこのうち対応する読みのテキストを入力として与える

番地表現	対応する読み
1-5	イチノゴ
32	サンジューニ, サンニ
6-105-9	ロクノヒャクゴノキュー
	ロクノイチマルゴノキュー
	ロクノイチゼロゴノキュー
	ロクノイチレイゴノキュー

表 2 学習用の合成音声の数

	train	dev	test
small	9,600	1,200	1,200
large	96,000	12,000	12,000

3.2.1 事前学習データの前処理

CSJ の音声は 16kHz でサンプリングされている．一方で，電話音声のサンプリング周波数は 8kHz である．学習時と評価時のサンプリング周波数のミスマッチを防ぐため，本実験では CSJ の音声を 8kHz にダウンサンプリングした．さらに，電話音声で用いられる μ-law アルゴリズム [16] によるコンパンディングにより，評価データである電話音声の音質に近づけた．以下，CSJ の音声を 8kHz にダウンサンプリングさせて学習させた事前学習モデルを Conformer$_{8k}$，さらに μ-law を通したものを Conformer$_{ulaw}$ と呼ぶ．

3.2.2 転移学習データの作成

転移学習用のデータは以下の通りに作成する．まず，認識させたい表現を受理する正規表現を記述する．本実験では番地表現 (数字とハイフンの組み合わせ[5)]) であるため，一つの表現に対して複数の読みの候補が存在しうる (例：0 → レイ，ゼロ，マル)．そこで，各表現に対してとりうる読みを列挙した上で，その読みに対応する音声を Google TTS によって作成する (表 1 参照)．その際，多様な音声表現を得るために話者 4 種類，音量 3 種類の組み合わせ (計 12 種類) の設定で音声を作成しデータを拡張した．最終的に生成可能な合成音声の量は膨大な数に及ぶため，今回はその中からランダムに分量を選択した 2 種類を用意した．以下，データ量の小さい方，大きい方をそれぞれ small，large と呼ぶ．それぞれのデータの数は表 2 に示す．

5） 実際の番地表現は漢字やアルファベットが含まれる場合があるが，本稿では簡単のため考慮していない．

また，3.2.3 節で述べるように，評価データはユーザ発話の前後に背景雑音のみの区間が入りうるため，条件を近づけるため合成音声の前後に無音区間を加えた音声も用意した．具体的には，合成音声の前に 0〜5 秒の中からランダムな長さの無音区間を追加し，合計の長さが 5 秒間になるように合成音声の後に無音区間を加える．以下，前述の前処理を加えていない場合，無音を追加した場合をそれぞれ TTS_{org},TTS_{pad} と表す．

3.2.3 評価データの作成

評価データは 21 名の日本語話者によって，Voicebot を通して収集した音声 (合計 131 発話) を利用した．具体的には，bot の発話が終了した時点から次に bot が再度発話を開始するまでの区間をユーザ発話区間とみなし，当該区間を切り出した．そのため，評価データには音声の前後に余分に切り出された背景雑音が含まれうる．この背景雑音区間の有無による性能の影響を検証するため，雑音区間を人手で削除したものもあわせて作成した．以下，人手で無音区間を切り出した音声を $test_{cut}$，切り出す前の音声を $test_{org}$ と表す．

4 実験結果と考察

学習した各モデルの認識性能の評価を表 3 に示す．TTS_{org} の条件下において，$test_{cut}$ に対して最も低い文字誤り率 (Character Error Rate; CER) となった学習設定では，ベースラインよりも高い精度を示したが，$test_{org}$ に対してはむしろ低い精度となった．また，他の条件を揃えた場合 LM を利用しない方が一貫してスコアが改善するという結果になった．そこで，最良の条件時 (事前学習モデル：Conformer_{ulaw}，学習データ量：large, LM：なし) の転移学習データを前述の TTS_{pad} に変更したところ，$test_{cut}$，$test_{org}$ 双方でさらに性能が改善し，いずれもベースラインを上回る結果となった．

表 3 の実験結果は転移学習時のパラメータの更新箇所を全層に対して行った結果となるが，Encoder を freeze させて実験を行ったところ，性能が著しく悪化するという結果になった (表 4 参照)．先行研究 [5, 6] では転移学習時に Decoder のパラメータを更新することが有効な学習戦略であったのに対し，本実験ではそれに反する結果となった．先行研究 [5] では Encoder-Decoder ネットワーク [17] を，[6] では RNN-Transducer [18] をベースにした構造を ASR モデルに採用しているが，本実験では ESPnet の CSJ レシピと同様に Hybrid CTC/Attention Architecture[13] を採用している．このモデルは学習時に Encoder からの出力から CTC Loss と Attention Decoder を通した損失を個別に計算し，両方のスコアの重み付き和により最終的な損失関数を決定する．CTC Loss は Encoder からの出力を線形層に通した後の値をもとに計算されるため，Encoder のパラメーターを freeze してしまうと更新するパラメータ数が極端に少なくなってしまい，学習がうまく進められなくなったと考えられる．

表 3 実験結果

モデル	転移学習データ	LM	CER	
			$test_{cut}$	$test_{org}$
Google STT (Baseline)			11.7	10.7
Conformer_{8k}	small, TTS_{org}	あり	12.3	23.6
Conformer_{8k}	small, TTS_{org}	なし	8.9	21.6
Conformer_{8k}	large, TTS_{org}	あり	**9.9**	30.1
Conformer_{8k}	large, TTS_{org}	なし	**7.2**	25.0
Conformer_{ulaw}	large, TTS_{org}	あり	**9.7**	27.0
Conformer_{ulaw}	large, TTS_{org}	なし	**6.3**	22.6
Conformer_{ulaw}	large, TTS_{pad}	なし	**4.9**	**8.7**

TTS_{pad} の条件下では，$test_{cut}$，$test_{org}$ の双方で性能が向上した．TTS_{org} の条件下では，事前学習時，転移学習時双方において，人間の発話の前後に背景雑音区間はほとんど含まれない．そのため，入力音声長と正解データのテキスト長がある程度対応付けられる学習条件になっており，背景雑音に対し hallucination [19, 20] が発生した可能性が示唆され，TTS_{pad} の条件によりその影響が低減されたと考えることができる．

表 4 更新パラメータの箇所による挙動の変化．なお，更新パラメータの箇所以外の学習条件は同一 (転移学習データは small, TTS_{org}，LM あり) とする．

モデル	更新箇所	CER($test_{cut}$)
Conformer	Decoder のみ	26.2
Conformer	全層	12.3

5 まとめ

本研究ではターゲットとなるドメインの学習データが得られない状況で，転移学習のためのデータを音声合成により作成する手法を電話音声認識タスクにおいて行った．比較実験の結果，最も良い性能を発揮した学習設定において，汎用的な音声認識モデルの精度を上回ることを示した．ただし，前後の無音区間を切り取っていない音声に対する性能の改善

幅は比較的軽微にとどまった．今後は，多様な背景雑音が含まれる音声に対する頑健性の向上について検討を続ける．また，本研究では正規表現が受理する発話のみを対象にドメイン適応しており，多様な話し方や言い間違い，言い淀みなどによる発話の揺れを考慮をしておらず，それらの音声への対処方法は十分ではない．この点に関しても検討していきたい．

謝辞

本論文の作成にあたりご協力いただきました，株式会社 AI Shift の杉山雅和氏，戸田隆道氏，二宮大空氏, 株式会社サイバーエージェントの邊土名朝飛氏に厚く御礼申し上げます．また，有益なご助言をいただきました，吉本暁文氏, 郡山知樹氏をはじめとする株式会社サイバーエージェントの AI Lab の方々，名古屋工業大学の李晃伸氏，上乃聖氏にこの場を借りて感謝申し上げます．

参考文献

[1] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. Investigation of language understanding impact for reinforcement learning based dialogue systems, 2017. https://arxiv.org/abs/1703.07055.

[2] Bing Liu and Ian Lane. End-to-end learning of task-oriented dialogs. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop**, pp. 67–73, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics.

[3] Anmol Gulati, Chung-Cheng Chiu, James Qin, Jiahui Yu, Niki Parmar, Ruoming Pang, Shibo Wang, Wei Han, Yonghui Wu, Yu Zhang, and Zhengdong Zhang. Conformer: Convolution-augmented transformer for speech recognition. In **Interspeech**.

[4] Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. In **Interspeech**, 2020.

[5] Mimura Masato, Ueno Sei, Inaguma Hirofumi, Sakai Shinsuke, and Kawahara Tatsuya. Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition. In **2018 IEEE Spoken Language Technology Workshop (SLT)**, pp. 477–484, 2018.

[6] Jinyu Li, Rui Zhao, Zhong Meng, Yanqing Liu, Wenning Wei, Sarangarajan Parthasarathy, Vadim Mazalov, Zhenghao Wang, Lei He, Sheng Zhao, and Yifan Gong. Developing rnn-t models surpassing high-performance hybrid models with customization capability. In **Interspeech**, 2020.

[7] Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. Cold fusion: Training seq2seq models together with language models. In **Interspeech**, 2018.

[8] Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, Dong Yu, and Lei Xie. Component fusion: Learning replaceable language model component for end-to-end speech recognition system. In **ICASSP**, 2019.

[9] Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, Christian Fuegen, Ozlem Kalinli, Yatharth Saraf, and Michael L. Seltzer. Contextualized streaming end-to-end

speech recognition with trie-based deep biasing and shallow fusion. In **Interspeech**, 2021.

[10] Mirek Novak and Pavlos Papadopoulos. Rnn-t lattice enhancement by grafting of pruned paths. In **Interspeech**, 2022.

[11] Jung Namkyu, Kim Geonmin, and Chung Joon Son. Spell my name: Keyword boosted speech recognition. In **ICASSP**, 2022.

[12] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. Espnet: End-to-end speech processing toolkit. In **Interspeech**, pp. 2207–2211, 2018.

[13] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 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.

[14] Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. In **Interspeech**, 2017.

[15] Kikuo Maekawa. Corpus of spontaneous japanese : its design and evaluation. **Proceedings of The ISCA IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)**, pp. 7–12, 2003.

[16] Mark A. Castellano, Todd Hiers, and Rebecca Ma. Tms320c6000 μ-law and a-law companding with software or the mcbsp. Technical report, Texas Instruments Inc, 2000.

[17] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philémon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In **ICASSP**, pp. 4945–4949, 2016.

[18] Alex Graves, Abdel rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In **ICASSP**, 2013.

[19] K. Sagae, M. Lehr, E. Prud'hommeaux, P. Xu, N. Glenn, D. Karakos, S. Khudanpur, B. Roark, M. Saraçlar, I. Shafran, D. Bikel, C. Callison-Burch, Y. Cao, K. Hall, E. Hasler, P. Koehn, A. Lopez, M. Post, and D. Riley. Hallucinated n-best lists for discriminative language modeling. In **ICASSP**, 2012.

[20] Krithika Ramesh, Ashiqur R. KhudaBukhsh, and Sumeet Kumar. 'beach' to 'bitch' : Inadvertent unsafe transcription of kids' content on youtube. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 36, No. 11, pp. 12108–12118, Jun. 2022.