医療論文から合成した音声による音声認識モデルの医療ドメイン適応

北出 祐[*1], 辻川 剛範[*1], 郭 翎[*2], 岡部 浩司[*2], 石井 亮[*3],
山本 仁[*2], 久保 雅洋[*1], 中川 敦寛[*4], 香取 幸夫[*3]
[*1] 日本電気株式会社 バイオメトリクス研究所, [*2] 日本電気株式会社 データサイエンスラボラトリー, [*3] 東北大学病院 耳鼻咽喉・頭頸部外科, [*4] 東北大学病院 産学連携室

Adaptation of Speech Recognition Model to Medical Domain using Synthesized Speech from Medical Papers

Tasuku Kitade[*1], Masanori Tsujikawa[*1], Guo Ling[*2], Koji Okabe[*2], Ryo Ishii[*3], Hitoshi Yamamoto[*2], Masahiro Kubo[*1], Atsuhiro Nakagawa [*4], Yukio Katori [*3]
[*1] NEC Corporation, Biometrics Research Laboratories
[*2] NEC Corporation, Data Science Laboratories
[*3] Tohoku University Hospital, Department of Otolaryngology-Head and Neck Surgery
[*4] Tohoku University Hospital, Experience Design and Alliance Section

抄録: 特定のドメインの音声を認識するために，汎用音声認識モデルに特定のドメインのデータを追加学習してそのドメインに適応させるのが一般的である．しかし，医療分野においては個人情報等の問題から医療現場のデータを大量に入手することは非常に困難であるため，一般に公開されている医学系学術論文を収集し，そのテキストから合成音声を作成した．これらを追加データとして次の 2 つの方法で音声認識モデルのドメイン適応を試みた．(1)約 30 万文のテキストおよびその合成音声を用いて汎用音声認識モデルのエンコーダーとデコーダーを追加学習する．(2)同データを用いて Encoder-freezing 学習を行い，汎用音声認識モデルのデコーダーだけを更新する．医療現場音声(18 ファイル，約 3.5 時間)を用いて，3 モデル(汎用モデル，ドメイン適応モデル(1)，(2))を比較評価したところ，汎用モデルに比べて(1)は医療用語の認識率は改善したものの(59.1%→68.5%)，単語認識率はほとんど変わらなかった(87.3%→87.4%)．一方，(2)は医療用語の認識率，単語認識率ともに改善した(各 71.0%，88.6%)．

キーワード 音声認識モデル 医療ドメイン適応 合成音声 Encoder-freezing 学習

1. はじめに

近年，Transformer により音声認識精度は大幅に改善されたが，特定の業界・ドメイン(話題)に対しては限定的で，認識したい内容・話し方等に近いデータ(音声及びその書き起こし)を用いてドメイン適応する方法[1]が一般的である．

しかし，医療分野において，例えば現場の医療従事者の音声を認識させたいときに，個人情報等の問題から，実データを入手することは非常に困難である．また，医療分野は他分野で使われない専門用語が多く含まれ，これらを認識するためには膨大なデータを必要とする．

そこで我々は合成音声を用いる手法[2]を医療ドメインに適用した．一般に公開されている医療系学術論文データのテキストおよびその合成音声を用いて追加学習した音声認識モデルを構築，評価した．追加学習する際，音声認識モデルのエンコーダー・デコーダー両方を更新する方法と，デコーダーのみを更新する方法を検討した．

2. 医療ドメイン適応

1) 学習テキストの収集

通常，音声認識モデルの学習には認識対象音声に近い内容の音声(入力)及びその書き起こし(正解)を用いるが，現場の音声を入手することが困難であるため，医療系学術論文より医療分野で用いられる専門用語や文脈を学習する．2000 年 1 月から 2022 年 6 月までの医学系学術論文 18 種類 33,238 本から本文を抽出し，303,046 文の文章を得て，これを学習テキストとした．

2) 合成音声の生成

3.1 節で収集した学習テキストから合成音声を生成した．合成音声の生成に必要な読みを取得するため，日本語形態素解析器 Mecab[3]を用いた．システム辞書に含まれない医療用語をユーザー辞書し，形態素解析を実行した．ESPnet2 TTS[4]を用いて1文につき1種類の合成音声を生成，303,046 個の音声ファイルを作成した．

3) 音声認識モデルの追加学習

音声認識モデルは大きくエンコーダーとデコーダーに分かれる．その内部はニューラルネットワークで構成されており，学習データを与えたときにモデルが最適な出力が得られるように各ネットワークの層が最適化(更新)される．しかし，入力データが合成音声であるため，入力音声から計算される音響特徴量が人間の声とは異なり，音声認識モデルへの悪影響が懸念されたため，デコーダーのみを更新する Encoder-freezing 学習[2]も検討した．

3. 評価実験

1) 実験環境

東北大学病院耳鼻咽喉・頭頸部外科の外来で収録した音声 18 ファイル(約 3.5 時間)から医師の音声を切り出して評価した．汎用音声認識モデル(1)，追加学習によりエンコーダー・デコーダーを更新したモデル(2)，Encoder-freezing 学習を行ったモデル(3)の 3 種類で，全単語および医療用語の認識率を用いて比較評価した．全単語数は 28,997，医療用語の数は 861 である．

2) 実験結果

結果を Table.1 に示す．医療用語を含む文章を学習した(1)(2)で医療用語の認識率は大幅に改善した．一方，単語全体の認識率は，医療用語の認識率とは異なり，(3)では 1.3%改善したものの，(2)は(1)とほぼ認識率が変わらなかった．

4. 考察

(1)のモデルの学習データに含まれていない，もしくはほとんど学習されていない医療用語を追加学習したことにより，これらの単語が認識されるようになった．一方で，(2)は医療用語以外の発話部分も医療用語が出力されるなど，過適応されている箇所が散見された．エンコーダーの更新も行ったことが一因と考えられる．(3)は，(2)に比べ過適応が少なく，また医療用語も多く認識されており，Encoder-freezing 学習が有効であることが分かった．

Table 1. ドメイン適応の効果

	医療用語認識率	単語認識率
(1)	59.1%	87.3%
(2)	68.5%	87.4%
(3)	71.0%	88.6%

5. まとめ

医療現場での音声を正しく認識するため，医療用語の認識強化を目的に，音声認識モデルの医療ドメイン適応を行った．通常用いられる，大量の実音声の取得が困難であることから，医学系論文のテキスト及びその合成音声を用いて，追加学習を行った．その際，デコーダー部分のみを更新する Encoder-freezing 学習を採用することで，単語認識率，医療用語の認識率ともに改善することが分かった．今後は，学習データを増やして音声認識モデルを改良し，医療用語の認識精度改善により医療現場で使える音声認識モデルの構築を目指す．

謝 辞

本研究を行うにあたり，音声収録にご協力下さった東北大学病院耳鼻咽喉・頭頸部外科の先生方，サポート・アドバイス等いただきましたバイオデザイン部門のスタッフの方々に，厚く御礼申し上げます．

参 考 文 献

[1] J. Huang et al.: Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition, arXiv preprint arXiv:2005.04290, 2020.

[2] Zheng, X. et al.: Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems, In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5674–5678. 2021.

[3] https://taku910.github.io/mecab/

[4] https://github.com/espnet/espnet