公開

研 究 終 了 報 告 書

「言語表現の階層構造に基づくEnd-to-End音声認識の研究」

研究期間: 2021年10月～2024年3月
研 究 者: 樋口 陽祐

1. 研究のねらい

近年、深層学習技術の進展に伴い、画像・動画情報処理や自然言語処理などの分野でニューラルネットワークを活用したモデルの性能が飛躍的に向上している。これらのモデルは大量の教師データ(入出力データのペア)を用いて訓練され、入力から出力への直接変換を行う End-to-End な処理を実現する。音声情報処理の分野では、発話音声をテキストに変換する音声認識システムの End-to-End 化が検討されている。このとき、正確なテキストを生成するには出力単語間の依存性をいかに捉えるかが鍵となるが、これを音声情報から直接抽出するのは容易でない。連続的な音声信号と離散的な言語記号では情報の性質が大きく異なり、これらの変換を効果的に学習するためには膨大な量の教師データを必要とする。

本研究課題では、End-to-End 音声認識モデルの認識精度および機能性を向上することを目的とし、音声情報から言語情報を効果的に抽出する技術の開発に取り組む。特に、言語的な事前知識を取り入れた特徴抽出プロセスを明示的に設計することで、データ駆動型のアプローチのみに頼らない新たな End-to-End 音声認識の枠組みを確立する。また、こうして得られる言語特徴表現は、音声認識精度の向上に留まらず、音声認識と連携した意味理解を要する課題にも有用であることを示す。

2. 研究成果

(1)概要

本研究課題では、End-to-End 音声認識の特徴抽出過程に組み込む言語的な知識として、大きく分けて以下の二つの方向性を検討した。

A) サブワード分割に基づいた出力単位の階層構造
B) 事前学習済み大規模言語モデルから得られる出力記号間の文脈情報

研究項目Aでは、「文字」や「単語」といった文の構成単位に着目し、出力系列を階層的に推定するモデルを開発した。モデルの内部で簡単なタスク(細かい単位での系列予測)から難しいタスク(粗い単位での系列予測)を段階的に解くことで、単語単位の特徴表現を効果的に学習し、従来モデルよりも認識精度が大きく改善することを確認した。また、発展的な成果として、粒度の異なる単位で生成される擬似ラベルを用いた教師あり学習手法を開発し、さらなる精度改善を実現した。他にも応用成果として、音声翻訳タスクや音声言語理解タスクにおいても提案の階層的な学習法が有効であることを確認した。研究項目 B では、BERT の埋め込み表現を音声処理の過程に明示的に組み込むことで、文脈情報を効果的に扱える End-to-End 音声認識モデルを考案した。BERT から得られる汎用的な言語知識を用いることで、音声認識精度が向上することを示した。また、提案モデルは音声意図理解タスクにおいても有効であることを確認した。BERT の他にも、近年の指示チューニングされた言語モデルを用いた音声認識手法も検討し、End-to-End 音声認識に有用な言語情報が得られることを確認した。

公開

上記研究成果を上げる上で、本 ACT-X 研究課題を通じて様々な研究機関との連携が実現した。カーネギーメロン大学へ研究訪問者として滞在した際は、音声処理分野において世界を牽引するチームと共に本研究課題において主要となる音声認識技術を開発し、自然言語処理分野の研究者とも密に交流することで音声言語理解に関する研究も取り組んだ。また、Google 社の音声認識チームにインターンシップとして参加した際は、本研究課題で得られた知見を積極的に共有することで、実環境下に対して頑健な音声認識モデルの開発に繋げることができた。

（2）詳細

研究項目 A「サブワード分割に基づいた出力単位の階層構造」

出力系列を階層的に推定する End-to-End 音声認識モデルを開発した。End-to-End 音声認識では、単語の推定に適した特徴表現が音声から暗黙的に抽出されることを期待しているが、これを教師データのみから学習するのは困難である。そこで、単語単位の特徴表現を効果的に獲得するために、End-to-End 音声認識の階層的条件付きモデルを提案した（図 1）。提案モデルは Transformer のエンコーダ層から構成され、中間層の出力表現を用いて Connectionist Temporal Classification（CTC）損失の計算を行った。このとき、音声入力に近い中間層からは「細かい単位」の系列が、テキスト出力に近い中間層からは「粗い単位」の系列が予測されるように、各損失に対する目的系列の単位を変化させた。また、細かい単位での系列予測によって、より粗い単位での系列予測を明示的に条件付けることで、単語レベルのスパースな単位での音声認識が効果的に学習されることを期待した。各目的系列の単位は、サブワードの語彙サイズを小さく/大きくすることで調整した。評価実験の結果、提案モデルは CTC に基づいた従来モデルよりも高い認識性能を与えることを確認した。特に教師データが限られた条件下において、提案の階層的な学習法は有効であった。本研究は音声処理のトップ国際会議の 1 つである ICASSP2022 に採択[1]された。

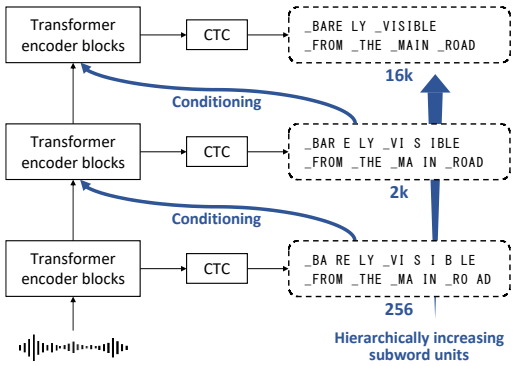

図 1 階層的条件付き End-to-End 音声認識

【発展的成果】「擬似ラベルを用いた半教師あり学習の適用」

提案の階層的条件付きモデルの認識精度をさらに向上させるために、擬似ラベルを用いた半教師あり学習手法を提案した。高精度な End-to-End 音声認識モデルを構築するには、大量の教師データを必要とする。これに対し、ラベルなし（音声のみの）データから擬似的なラベルを生成し、この音声と擬似ラベルのペアを用いて教師データを拡張することの有効性が知られている。本研究では、階層的条件付きモデルの中間層でも出力系列の生成を行い、ラベルなし音声データに対して異なる出力単位の擬似ラベルを生成した（図

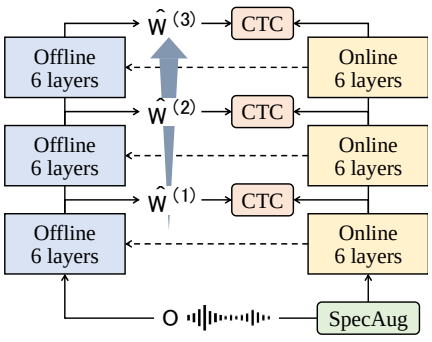

図 2 中間的な擬似ラベルを用いた半教師あり学習

2)。これら複数の擬似ラベルを用いて、階層的条件付きモデルを新たに学習することで、音声認識精度が大きく改善することを確認した。本研究は音声処理のトップ国際会議であるICASSP2023 に採択[4]された。

【応用成果】「音声言語理解、音声翻訳タスクへの拡張」

提案モデルで用いた中間的な損失が、音声言語理解や音声翻訳の End-to-End 学習においても有効であることを示した。音声言語理解や音声翻訳は、入力音声に対して音声認識を行い、得られたテキストを理解または翻訳するプロセスとして分解できる。本研究では、End-to-End 音声言語理解/音声翻訳モデルの中間層に対して、CTC に基づいた音声認識損失を適用し、その認識結果を目的タスクの損失計算に条件付ける学習法を提案した。これにより、各タスクにおける End-to-End モデルの性能が向上することを確認した。本研究はカーネギーメロン大学へ訪問研究員として滞在して際の成果であり、音声処理の国際会議であるSLT2022 および自然言語処理の国際会議である EACL2023 に採択[5,6]された。

研究項目 B「事前学習済み大規模言語モデルから得られる出力記号間の文脈情報」

事前学習済みマスク言語モデルである BERT から得られる汎用的な言語知識を用いて、End-to-End 音声認識モデルの性能を向上することを試みた。音声認識において正確なテキストを生成するには、出力単語間の依存性を捉えることが重要となるが、これを音声情報のみから抽出するのは容易でない。例えば、ある発話音声に対して「あめ」という音を認識するだけでは不十分であり、その音が「雨」と「飴」のどちらを意図しているのかについても、文脈に応じて推定する必要がある。

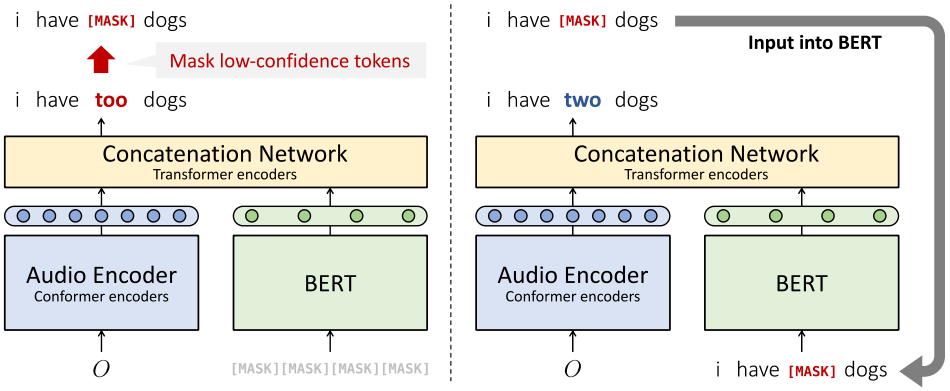

図 3 BERT の埋め込み表現を用いた End-to-End 音声認識

本研究では、大規模言語モデルである BERT から得られる汎用的な言語知識を音声処理の過程に明示的に組み込むことで、文脈情報を効果的に扱える音声認識手法を提案した。具体的には、CTC による End-to-End 音声認識モデルの学習および推論時に、BERT の埋め込み表現を条件付けることで、出力記号間の依存性を陽に考慮する系列推定を実現した(図 3)。様々な言語や発話スタイル、学習データ量を用いた音声認識実験において提案モデルを評価した結果、従来手法よりも高い認識精度が得られることを確認した。また、提案モデルから抽出される BERT の埋め込み表現と音声情報を統合した特徴量は、意図理解を伴う音声言語理解タスクにおいても有効であることを示した。本研究の成果は、カーネギーメロン大学へ訪問研究員として滞在した際に得られたものであり、自然言語処理の国際会議である EMNLP2023

に Findings に採択[2]された。

【発展的成果】「BERT と音声認識モデルにおけるテキスト形式の差異の解消」

BERT と音声認識モデルで扱われるテキスト形式の違い（語彙の違い、句読点の有無や大文字・小文字の区別など）を解消するための手法を検討した。提案モデルは、自身の出力結果を BERT の入力として用いるため（図 3 右）、BERT の語彙に合わせてモデルを学習する必要があった。しかし、BERT の語彙は主に Wikipedia などの Web 上のテキストデータから構築され、これは音声認識の学習に必ずしも適していない。例えば、音声認識では話し言葉を対象とする場合があり、BERT の語彙をそのまま用いることで、ドメインの不一致が生じうる。この課題に対し、新たなデコーダネットワークを提案モデルに導入し、音声認識に適した語彙を用いて系列推定が行えるようにした。実験による評価の結果、音声認識に適したデコーダネットワークを用いることで認識精度がさらに向上することを確認した。本研究の成果は音声処理のトップ国際会議である ICASSP2023 に採択[3]された。

【発展的成果】「指示チューニングされた言語モデルの利用」

生成型言語モデルから得られる言語知識を用いて、End-to-End 音声認識モデルの性能を向上することを試みた。ChatGPT といった最新の生成型言語モデルは、自然言語による指示を含むプロンプトでファインチューニングすることで、様々な自然言語処理タスクに対して高い汎化性能を示している。本研究では、生成型言語モデルに音声認識仮説の文法誤り訂正タスクを解かせ、そこから得られる特徴表現を用いて End-to-End 音声認識モデルにおける系列生成を学習した。複数の音声認識用データセットを用いた評価実験の結果、生成型言語モデルを用いることで高い認識精度が達成できることを確認した。本研究の成果はプレプリントとして公開[7]しており、査読付きの国際学会に投稿予定である。

3. 今後の展開

本研究を通して、言語情報を効果的に扱える End-to-End 音声認識の基盤となる技術を開発してきた。構築したモデルの評価実験にはノイズが少なく整備された公開データセットを用いたが、音声認識が日常生活で使われる環境はノイズが多く、話者や発音の特性も多様であることが想定される。このような実環境下における提案モデルの有効性を評価すると共に、認識精度を担保するための要素技術を導入・開発していく必要がある。そのために、事前学習済みの音響モデルの利用、大規模で多様な実データの構築、さらに音声強調システムとの統合などの方向性を検討する予定である。他にも、音声対話といったインタラクティブな用途を想定する場合、オンラインストリーミング音声認識システムを実装する必要がある。ストリーミング音声認識は、入力音声を逐次的に認識する技術であり、発話から認識までにかかる時間（遅延）を削減することが求められる。この課題に関しては、本研究で開発した技術がストリーミング音声認識の事前学習において有用であることが初期検討で明らかになっており、この方向性を引き続き探求していく予定である。これらの研究は今後 3 年以内で実現することを目指しながら、音声言語理解への応用も逐次検討していく予定である。

4. 自己評価

研究目的の達成状況

当初の計画である、音声情報から言語情報を効果的に抽出する End-to-End 音声認識技術の

公開

開発を大きく進めることができた。また、開発した技術を基盤として意味理解を伴うタスクへの応用にも取り組み、音声言語理解の実現に向けた成果も出すことができた。一方で、社会実装を考えると、実環境下における頑健性の評価やオンラインストリーミング認識システムの実装など、まだ多くの課題が残る。本研究の成果は、これらを実現するための基礎基盤を整理したものでもあり、今後の関連分野の発展に寄与し、実応用においても有益な知見をもたらすと考えている。以上のことから、研究目標をおおむね達成できたと言える。

研究の進め方（研究実施体制及び研究費執行状況）

代表者主導で研究を集中して遂行することができた。また、国内外の著名な研究機関との共同研究を通じて、音声処理と自然言語処理における最新の動向を効率的に情報交換することで、当初計画していた以上の成果も得ることができた。一方で、ACT-X 繋がりの研究者との連携が不十分だったことは反省すべき点の一つである。領域会議やサイトビジットでは最小限の議論に留めてしまい、他分野の研究者との交流や自身の分野からの意見提供が上手く行えなかった。この反省を踏まえ、研究者としての意識を今後改善していきたい。研究費については、AI 橋渡しクラウド（ABCI）上の計算資源の購入に大半を充てた。これにより、大学の環境では実施が困難である大規模な実験を行うことが可能となり、ACT-X の支援のおかげで多くの成果を生み出すことができた。

研究成果の科学技術及び社会・経済への波及効果

本研究課題で得られた成果の多くは、音声処理分野および自然言語処理分野の主要国際会議で発表を行った。他の研究機関からも、本研究の成果を基にした手法が既に多数発表されており、分野内での注目も高まっていると考えている。

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：12件 （国際発表8件、国内発表4件）

1. Yosuke Higuchi, Keita Karube, Tetsuji Ogawa, Tetsunori Kobayashi. Hierarchical Conditional End-to-End ASR with CTC and Multi-Granular Subword Units. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7797-7801, 2022.
概要 End-to-End 音声認識において、単語単位の認識を行うための特徴表現を獲得することを検討し、階層的条件付きモデルを提案した。モデルの内部で簡単なタスク（細かい単位での系列予測）から難しいタスク（粗い単位での系列予測）を段階的に解くことで、単語単位の特徴表現を効果的に学習し、従来モデルよりも認識精度が大きく改善することを確認した。
2. Yosuke Higuchi, Brian Yan, Siddhant Arora, Tetsuji Ogawa, Tetsunori Kobayashi, Shinji Watanabe. BERT Meets CTC: New Formulation of End-to-End Speech Recognition with Pre-trained Masked Language Model. In Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 5486-5503, 2022.

公開

概要　事前学習済みのマスク言語モデルから得られる汎用的な言語知識を、End-to-End 音声認識に活用することを試みた。音声と BERT の埋め込み表現を統合する手法として、BERT を用いて CTC に出力記号間の依存性を条件付ける BERT-CTC を提案した。複数の音声認識ベンチマークを用いた実験の結果、提案モデルは BERT の知識を用いることで、既存モデルよりも高い認識性能を与えることが明らかとなった。

3. Yosuke Higuchi, Tetsuji Ogawa, Tetsunori Kobayashi, Shinji Watanabe. BECTRA: Transducer-based End-to-End ASR with BERT-Enhanced Encoder. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023.

概要　BERT-CTC [2] は BERT と共通の語彙を用いて構築する必要がある。しかし、BERT で使われる語彙は必ずしも音声認識の学習に適していない。例えば、音声認識では話し言葉を対象とする場合があり、BERT の語彙をそのまま用いることで、ドメインの不一致が生じうる。この語彙に関する制約を解決するために、音声認識に適した語彙を用いて系列生成を行うネットワークを BERT-CTC に追加した。結果として、BERT-CTC の性能が効果的に改善することを確認した。

（2）特許出願

該当なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

4. Yosuke Higuchi, Tetsuji Ogawa, Tetsunori Kobayashi, Shinji Watanabe. InterMPL: Momentum Pseudo-Labeling with Intermediate CTC Loss. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.
5. Yifan Peng*, Siddhant Arora*, Yosuke Higuchi, Yushi Ueda, Sujay Kumar, Karthik Ganesan, Siddharth Dalmia, Xuankai Chang, Shinji Watanabe. A Study on the Integration of Pre-Trained SSL, ASR, LM and SLU Models for Spoken Language Understanding. In Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT). pp. 406-413, 2022.
6. Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, Alan W Black, Shinji Watanabe. CTC Alignments Improve Autoregressive Translation. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL). pp. 1623-1639, 2023.
7. Yosuke Higuchi, Tetsuji Ogawa, Tetsunori Kobayashi. Harnessing the Zero-Shot Power of Instruction-Tuned Large Language Model in End-to-End Speech Recognition. arXiv preprint arXiv:2309.10524. 2023.