氏　　名	王　宇
博士の専攻分野の名称	博士（工学）
学 位 記 番 号	医工農博甲第１３８号
学位授与年月日	令和６年３月２２日
学位授与の要件	学位規則第４条第１項該当
専　　攻　　名	工学専攻　　システム統合工学コース
学 位 論 文 題 目	A Study on Real-Time Automatic Speech Recognition System on Edge Devices (エッジ端末におけるリアルタイム音声認識システムに関する研究)
論 文 審 査 委 員	主査　教　授　西﨑　博光 教　授　鈴木　良弥 教　授　大渕　竜太郎 教　授　福本　文代 教　授　豊浦　正広 教　授　伊藤　一帆

学位論文内容の要旨

ASR technology has long been an important field within artificial intelligence. Early ASR systems used the grammatical rules of human natural language to convert speech to text. With a large amount of speech and text data being open-sourced, statistics-based ASR approaches have gradually become mainstream. Traditional statistical ASR systems first refine the speech signal into acoustic features, then use Gaussian mixture models and hidden Markov models to model the distributional and transfer probabilities of acoustic features, and use N-Gram language models to obtain the contextual probability of human natural language. Finally, these three probabilities will be combined to obtain the final recognition result.

ASR systems are typically deployed either on a cloud server or an edge device. A mainstream ASR system comprises two primary components: a DNN model and an ASR decoder. DNN models require substantial computational resources during the forward stage, while ASR decoders generally consume significant memory. Consequently, ASR systems are often deployed on cloud servers, benefiting from advanced computational

chips and ample, cost-effective memory capacity. On the cloud side, the main emphasis of ASR systems is on identifying DNN models with enhanced accuracy. Nevertheless, considering the occasional unavailability of communication networks and the need to protect users' private information from being uploaded to server databases, deploying ASR systems on edge devices becomes essential. Because this necessitates robust hardware computational power and memory capacity to efficiently run the ASR system, this is particularly challenging when deploying ASR systems on low-end devices. Therefore, the research on cloud-side ASR systems primarily focuses on developing model structures that are optimized for various devices.

Taking into account the advantages of both the device-side and cloud-side ASR system he workflow of a complete, deployable ASR system, which is currently popular, unfolds as follows:

1. Capture the speech signal with a microphone on the device side.
2. For more complex tasks, such as speech conversion and online navigation, the speech signal is uploaded via the communication network to a remote server. Subsequently, the cloud ASR system processes the signal to obtain recognition results, which are then relayed back to the device.
3. For simpler tasks, such as activation via a wake-up word and speech commands, or when network communication is unavailable, or user data is not allowed to be uploaded, the recognition result is obtained directly using the device-side ASR system.

This study explores optimization approaches for both cloud-side and device-side ASR systems, respectively. For cloud-side ASR system, I propose a toolkit for developing real-time ASR systems in a cloud environment. As previously mentioned, an ASR system comprises a DNN model and an ASR decoder. Kaldi, one of the most popular ASR toolkits, offers integrated functions for building DNN models and decoders. However, since Kaldi is developed in C++, it poses challenges in debugging model structures and training strategies. Conversely, the flexibility of the Python language has led to its widespread use in deep learning frameworks like PyTorch and TensorFlow, which have fostered a proliferation of advanced neural network models. Consequently, ASR researchers are keen on finding convenient methods to integrate DNN models trained with Python-based deep learning frameworks into decoders built using Kaldi's C++ framework. I propose a comprehensive toolkit that encompasses all essential

functions required to establish a complete real-time ASR pipeline. This includes recording speech signals, extracting acoustic features, transmitting over networks, processing with DNN models, and decoding. The toolkit is designed for ease of integration, allowing developers to seamlessly incorporate their custom models and decoding graphs. Additionally, it supports feature mixing and the integration of denoising models.

For the device-side ASR system, I propose a lightweight ASR system, which includes a novel DNN model structure and an enhanced decoding algorithm. Recently, as edge devices increasingly integrate AI applications, manufacturers have been providing their own inference frameworks for DNN models to maximize chip computational power. However, many advanced and high-precision DNN model structures are incompatible with these frameworks. Additionally, ASR decoders typically used in cloud-based systems consume substantial computational memory, posing challenges for deployment on edge devices.

While some related works have introduced lightweight DNN model structures for device-side usage, they often overlook the compatibility with inference frameworks of lower-end edge devices, rendering them unsuitable. Furthermore, there has been limited focus on optimizing decoding algorithms in prior research.

To address these gaps, my approach encompasses several innovations. Firstly, I introduce a model structure exclusively based on convolutional neural networks, ensuring compatibility with most edge device software development kits. I also streamline the process by directly utilizing speech signals, thereby reducing CPU usage. Lastly, I have refined the prefix beam search decoding algorithm by implementing a unique pruning method using a lexicon trie and introducing a novel language model that leverages initial letters. These enhancements collectively enable the construction of a high-accuracy ASR decoder that requires less computational memory.

My experimental results demonstrate that this ASR system excels in model size, recognition accuracy, and ease of deployment, characterized by low CPU usage and a high real-time processing rate.

To summarize, my research on ASR systems for edge devices introduces new solutions for both cloud and device-side systems. On the cloud side, I have developed a toolkit to construct a real-time ASR pipeline. For the device side, I have designed a novel model

structure and an ASR decoder. These advancements contribute to creating an ASR system that is lightweight, easily deployable, highly accurate, and possesses a rapid real-time processing rate. While this study has achieved significant progress, there is still potential for fine-tuning the DNN model's structure on the device side. In future work, I aim to further investigate deploying more complex models on low-end devices.

論文審査結果の要旨

この博士論文では，音声認識技術を小型エッジ端末機器上で動作させ，製品に展開できる形にすることを目指して，クラウドとエッジ端末上で動作する音声認識システムの課題解決を図っている。クラウド側では，既存の音声認識ツールキットである，"Kaldi ツールキット"を Python コードや既存のディープラーニングフレームワークで動かせるようにラップした ExKaldi-RT というツールキットを新しく開発し，最新のディープラーニングモデルを容易に Kaldi デコーダーと統合できるソフトウェアを開発した。これにより高精度なリアルタイム音声認識システムの構築が可能となった。一方エッジ端末側では，組み込みデバイスの互換性を確保しつつ高精度を実現するために，畳み込みニューラルネットワークに基づく軽量な end-to-end 音声認識モデルと改良されたビームサーチデコーダーを提案した。これにより，従来の方法と比べて大幅に記憶容量を削減しつつ，高い音声認識性能を維持できることが確認できている。

博士論文発表では，大きく 3 つの主要なトピックに分けて成果が報告されている。

まず，最初のトピックはクラウド上で動作する音声認識システムの開発についてである。Python ベースで既存の音声認識エンジンである Kaldi ツールキットをラップし，リアルタイム音声認識パイプライン開発を支援するツールキットである ExKaldi-RT を開発した。もともとの Kaldi は C++言語，Perl，そしてシェルスクリプトで開発されているため柔軟性に欠けており，TensorFlow や PyTorch といった既存のディープラーニングフレームワークで訓練したモデルとの統合が難しいという大きな課題が存在していた。ExKaldi-RT はこうした課題を解決し，音声の録音から，特徴量抽出，ニューラルネットワーク（DNN）への転送，DNN の推論，デコーディングといったリアルタイム音声認識に必要な機能をフレキシブルに提供できる便利なツールである。ExKaldi-RT を用いることで，ディープラーニングフレームワークで訓練した DNN ベースの音響モデルを直感的に統合ができ，Kaldi と同等の精度とリアルタイム性能を達成できることが確認されている。加えて，異なる音響特徴量のフュージョンやノイズ低減 DNN モデルの統合などの新たな音声認識の工夫を簡単に実装することができ，音声認識システムの精度とロバスト性を大幅に向上できるこ

とも実証されている。以上より，ExKaldi-RT はリアルタイム高精度音声認識システム開発に有用なツールキットであるという結論付けが行われている。

２つ目の研究トピックは，リソース制約のあるエッジデバイス上で動作する軽量かつ高精度な音声認識システムの実現を目指して，畳み込みニューラルネットワークをベースとした end-to-end 音声認識モデルと，ビーム探索アルゴリズムの最適化を提案している。方法として，音声波形を直接入力することで CPU 使用率を抑えつつ，辞書トライ木と単語の先頭文字のみを使った言語モデルを取り入れることで，ビーム探索の精度向上を図りつつメモリの使用量を大幅に削減する。実験の結果，提案手法は従来の音声認識方法と比較して大幅にパラメータ数を削減しつつ，評価セット上では同等の認識精度を達成できることが確認されている。さらに組み込みデバイスへのデプロイにおいても，低 CPU 使用率とリアルタイム処理能力が実証されている。以上より，畳み込みニューラルネットワークといった小さなネットワークとビーム探索アルゴリズムの最適化によって，エッジデバイス上での高精度音声認識の実現可能性を示した。

最後の研究トピックは，エッジデバイス上で動作可能な軽量かつノイズに頑健な音声区間検出（VAD）モデルの構築である。提案手法では，音声認識モデルの特徴抽出部との共有を図ることでパラメータ数を抑え，ノイズデータ拡張によりノイズ頑健性の向上を図っている。モデル訓練時には，セルフラベリングにより擬似ラベルを作成し，これを教師データとして CNN ベースの VAD モデルを訓練し，頑健性の評価を行っている。その結果として，音声認識システム全体へのパラメータ増加を抑えたにもかかわらず，高い精度で達成できることが実証され，強いノイズ下でも高い頑健性を確保できていることが確認されている。

本博士論文の成果をまとめる。1 つ目は，クラウド向けの ExKaldi-RT と呼ばれるツールキットを開発し，簡易に音声認識システムを構築でき，かつ高い拡張性を実現したリアルタイム音声認識パイプラインを提供し，既存のディープラーニングモデルとの柔軟な統合を可能にした。2 つ目は，エッジデバイス向けに，小規模なネットワークを用いた軽量かつ高性能な音声認識モデルとメモリ効率の良いビーム探索デコーダを提案し，比較的小さな組み込みデバイスでも高い精度を維持したまま動作させることを実性した。最後は，組み込み端末上で動作する VAD モデルを設計し、雑音下でもエッジ端末の音声認識システム全体のパフォーマンスを高めることに成功している。このように本博士論文の研究では，音声認識技術の研究開発から実適用までを視野に，クラウドとエッジのそれぞれに適した手法を提案することで音声認識の発展に寄与しているものと考えられる。

博士論文公聴会および最終審査会においては，審査委員からエッジデバイスとクラウド

サーバの切り替えについて，音声認識実験の雑音条件について，開発した手法の新規性などについて質問がなされた。一部の質問については回答が保留され，論文の内容についてもエッジデバイスの条件についての情報の不足，執筆内容の疑義などが指摘されたが，最終審査での質問や論文の指摘事項についても修正を行い，博士論文の再提出が行われるなどして適切に対処された。以上の研究内容については，査読付き国際会議論文，雑誌論文に採録されていることからも，本博士論文は，本学の博士学位を授与するに相応であると判断し，論文審査及び最終審査を合格とすることとした。