（続紙 １）

京都大学	博士（情報学）	氏名	KAK SOKY
論文題目	Incorporating Meta Information for Speech Recognition of Low-resource Language（低資源言語の音声認識のためのメタ情報の活用）		

（論文内容の要旨）

Automatic speech recognition (ASR) systems have been developed as an aid for communications, not only in human-to-machine interfaces but also in human-to-human interactions. The ASR systems have achieved impressive performance in the last decade with the advancement of deep learning techniques and computing resources. However, the performance is drastically degraded for low-resource languages because of data scarcity, especially in the current trend of end-to-end (E2E) deep neural networks (DNN) architecture, which requires a massive amount of labeled speech data for training. This study addresses the problem of improving the ASR systems for low-resource languages by incorporating meta-information or auxiliary knowledge. Here, meta-information is concerned with the speaker, domain, and language, which can be extracted together with the speech content, whereas the auxiliary knowledge is the translated information from other rich-resource languages.

We first present a large parallel speech corpus of the Extraordinary Chambers in the Courts of Cambodia (ECCC) for transcription and translation in Khmer, English, and French in Chapter 3. We address the problem of sentence segmentation in low-resource languages by conducting bilingual sentence alignment from rich-resource to low-resource language with the monotonic assumption and then enhance the alignment using the ROVER method that combines multiple machine translation (MT) outputs. We also enhance the baseline MT systems of a low-resource language by finetuning the model using the pretrained MT model of the rich-resource languages.

In Chapter 4, we address the effective use of speaker information for enhancing ASR systems in the speaker-imbalanced dataset. The proposed approach jointly trains speaker recognition (SRE) and ASR in an E2E model. With a direct connection of SRE to the ASR decoder, it allows for backpropagating the ASR loss to the SRE decoder, resulting in joint optimization. Moreover, conducting speaker clustering can compensate for minor speakers, which is beneficial for the speaker-sparse datasets. The proposed method improved the character error rate (CER) of the baseline model by 3.4% relative, with SRE improvement by 8.2% relative.

In Chapter 5, we present the effective finetuning of a large-scale pretrained model for low-resource language ASR with very low-resource settings. The finetuning process is composed of two-step adaptation: domain adaptation and language adaptation, using heterogeneous datasets which are matched with either domain or language. We incorporate meta-information such as domain and language in multi-task learning or adversarial learning for effective adaptation. Moreover, the fusion of domain or language identification to the ASR decoder is effective. The proposed method outperformed the naive adaptation in the CER relatively by 31.8%, 16.3%, and 9.3% for one-hour, 5-hour, and 10-hour target speech datasets, respectively.

In Chapter 6, we present an effective framework of incorporating the translation knowledge from rich-resource languages to improve the

transcription of a low-resource language in multi-lingual scenarios. It assumes that the content of its back-translation is the same as the transcription of the original speech. We formulate this framework as a joint process of ASR and MT with the cross-attention mechanism of the decoder module. The proposed method improved the word error rate (WER) of Khmer and Spanish relatively by 8.9% and 1.7%, respectively.

Chapter 7 concludes the thesis and a brief look at future work.

（続紙 ２）

（論文審査の結果の要旨）

音声認識技術は深層学習の進展により大きな性能向上を遂げたが、大規模な学習データを前提としており、音声言語コーパスが十分にない言語ではモデル学習が容易でない。本研究は、このような低資源言語、特にクメール語を対象として、効率的に音声認識のモデル学習及び推論を行う方法を複数提案し、その実験的評価をまとめたもので、主な成果は以下の通りである。

1. クメール・ルージュの犯罪を裁いているカンボジア裁判所特別法廷(Extraordinary Chambers in the Courts of Cambodia)の公判記録を元に、クメール語・英語・フランス語の３か国語の音声・書き起こしを対応付けたコーパスを構築し、ベースラインとなる音声認識・機械翻訳・音声翻訳システムを構成した。

2. 裁判のように話者がある程度限定され、発話量に偏りがある状況を対象として、話者認識と音声認識の統合的な学習・推論を行う枠組みを提案した。話者認識結果を利用することで音声認識精度が改善することを示した。

3. 言語とドメインのいずれかが一致している他のデータセットを効果的に活用して、多言語で自己教師付き事前学習された音声モデルをファインチューニングする方法を検討した。言語適応をドメイン認識と、ドメイン適応を言語認識と各々マルチタスク学習または敵対的学習することで行い、さらにこれらの認識結果の情報を利用することで、音声認識精度が大きく改善することを示した。対象タスクの音声が５時間しかない条件でも、単純なファインチューニングと比べて絶対値で約2%改善し、約10%の文字誤り率を実現した。

4. 国際会議・法廷のように同時翻訳がある状況を想定して、高資源言語（英語またはフランス語）の翻訳結果を参照することで、元の低資源言語（クメール語）の音声認識を改善する枠組みを提案した。これにより、単語誤り率で相対的に約9%の改善を実現した。

以上のように本論文は、低資源言語の音声認識を効率的に実現する方法を提示するとともに、クメール語の話し言葉コーパスと音声認識システムを初めて構築しており、学術上・実用上寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和5年2月16日に論文とそれに関連した内容に関する口頭試問を行った結果、合格と認めた。また、本論文のインターネットでの全文公開についても支障がないことを確認した。