様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １３ 日現在

機関番号：８２６３６
研究種目：若手研究
研究期間：2021～2023
課題番号：２１Ｋ１７７７６

研究課題名（和文）Self-supervised graph-based representation for language and speaker detection

研究課題名（英文）Self-supervised graph-based representation for language and speaker detection

研究代表者
沈　鵬（Shen, Peng）

国立研究開発法人情報通信研究機構・ユニバーサルコミュニケーション研究所先進的音声翻訳研究開発推進センター・主任研究員

研究者番号：８０７７３１１８

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本プロジェクトでは、自己教師あり学習または事前学習技術を開発することにより、言語と話者の認識を改善することを目指していた。私たちは、音声信号から言語と話者の特徴をよりよく捉えるためのさまざまな方法を試験した。提案された技術は、言語と話者の認識だけでなく、音声認識タスクにも効果的であった。さらに、単一のモデルを使用して言語、話者、音声認識を対応するマルチタスク認識も検討された。研究成果はIEEE ICASSP、SLT、ASRU、Interspeechなどの国際会議で発表された。

研究成果の学術的意義や社会的意義
本プロジェクトは、音声信号の理解と表現を進化させることをその大きな目的としており、このことは重要な科学的意義を有する。言語と話者の認識におけるパフォーマンス向上のための技術は、技術的な応用を進めることに役立つ。

研究成果の概要（英文）：In this project, we focus on developing self-supervised or pre-trained techniques to enhance spoken language and speaker recognition tasks. We experimented with different methods to better capture the characteristics of languages and speakers from speech signals. Our proposed techniques include transducer-based language embeddings, pronunciation-aware character encoding, cross-modal alignment, and generative linguistic representations. These innovations aim to improve language and speaker recognition, as well as speech recognition tasks. Further, we explored multi-task recognition to advance language, speaker, and speech recognition using a single model. The results of this project have been published at top international conferences, including IEEE ICASSP, SLT, ASRU, and Interspeech.

研究分野：知覚情報処理関連

キーワード：language identification　Speech recognition　self-supervised learning　speaker recognition

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

Developing high-performance techniques for language and speaker detection is crucial for enhancing the usability of real-time multilingual speech translation systems. However, these advanced models are sensitive to data and tasks because they determine the decision boundaries between classes based mainly on the given data, with limited prior knowledge or assumptions about data distributions. Relying on label-based prior knowledge often results in the loss of inherent speech data characteristics, such as background, channel, and rhythm, which can reduce system robustness. In this project, we proposed using self-supervised learning to address the problems of channel and domain overfitting commonly seen in current deep learning techniques.

２．研究の目的

In this project, my goal was to develop a self-supervised, graph-based representation of speech signals to enhance the performance of spoken language and speaker recognition tasks. To achieve this, we needed to address two key questions:

(1) How can we build effective self-supervised learning models to extract features of language, speaker, and other acoustic information from speech signals?

(2) How can we design effective networks that utilize these representations to improve the target tasks?

３．研究の方法

The research is structured into three phases, carried out as follows:

(1) **Understanding the Representation of Speech Signals**: I explored how to represent speech signals using self-supervised learning, speech recognition-based pretraining, and generative learning to enhance our understanding of speech signal representations.

(2) **Learning to Use the Representation**: To improve performance for specific tasks, I constructed various networks or model units that incorporate phonetic, speaker, and language information.

(3) **Evaluation on Task-Specific and Multi-Task Systems**: Initially, my research focused on developing task-specific systems, such as language identification and speech recognition. Additionally, I explored multi-task systems to further leverage the proposed representations.

４．研究成果

The achievement in this project are as follows:

(1) **Improving the representation of speech signal for language identification (LID)**: We proposed a novel transducer-based language embedding approach for LID tasks by integrating an RNN transducer model into a language embedding framework. Benefiting from the advantages of the RNN transducer's linguistic representation capability, the proposed method can exploit both phonetically-aware acoustic features and explicit linguistic features for LID tasks. The research paper was accepted by Interspeech 2022. Additionally, we further investigated these techniques on the National

Institute of Information and Communications Technology (NICT) LID system, which also demonstrated robustness on cross-channel data.

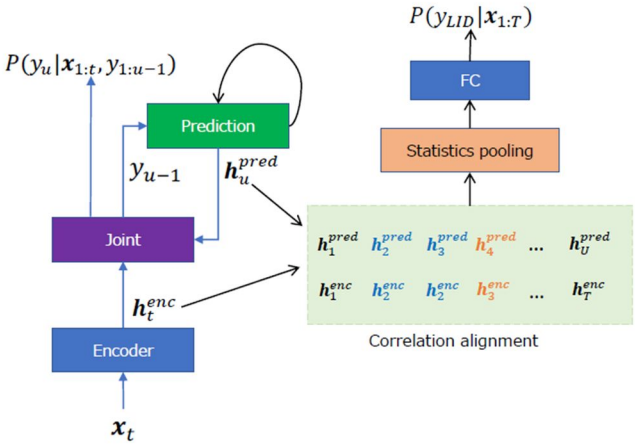

Fig.1 RNN-Transducer-based LID system

(2) **Pronunciation-aware unique character encoding for Mandarin ASR**: I proposed to use a novel pronunciation-aware unique character encoding for building end-to-end RNN-T-based Mandarin ASR systems. The proposed encoding is a combination of pronunciation-based syllable and character index (CI). By introducing the CI, the RNN-T model can overcome the homophone problem while utilizing the pronunciation information for extracting modeling units. With the proposed encoding, the model outputs can be converted into the final recognition result through a one-to-one mapping. This paper was accepted by IEEE SLT 2022.

(3) **Speaker mask transformer**: I proposed a novel speaker mask branch to detection the speech segments of individual speakers. With the proposed model, we can perform both ASR and speaker diarization tasks simultaneously using a single model. In this project, I conducted investigations and utilized models trained with self-supervised techniques or pre-trained techniques for LID, speaker recognition, and ASR tasks.

(4) **Generative linguistic representation**: With the success of ChatGPT, I began to investigate the generative model and tried to use the knowledge from this model to improve the performance of LID. Such investigations are important to understand the behavior of large language models. Our work was published by IEEE ASRU 2023.

(5) **Utilizing generative and discriminative model for speaker verification**: We proposed a hybrid learning framework, i.e., coupling a joint Bayesian generative model structure and parameters with a neural discriminative learning framework to improve the recognition performance (The related results were published in the IEEE/ACM TASLP(journal) and APASIPA(international conference)).

(6) **Cross-domain and transfer learning**: I focused on improving cross-domain ASR tasks. We tried to use pre-trained large models, such as BERT, and proposed using optimal transport techniques to better utilize the knowledge transferred from the large models. Our works were published by IEEE ICASSP 2022, 2024, and IEEE ASRU 2023.

Through this research, we classified how to better utilize the knowledge inside the pre-trained models and proposed several techniques, such as RNN-T-based LID and optimal

transport-based ASR to improve the performance of these tasks. Especially, our proposed techniques were successfully used to build the NICT LID system, which showed very robust performance.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 X. Lu, P. Shen, Y. Tsao, H. Kawai	４．巻 29
２．論文標題 Coupling a generative model with a discriminative learning framework for speaker verification	５．発行年 2021年
３．雑誌名 IEEE/ACM Trans. on Audio, Speech and Language Processing	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TASLP.2021.3129360	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　7件）

１．発表者名 P Shen, X Lu, H Kawai
２．発表標題 Pronunciation-aware unique character encoding for RNN Transducer-based Mandarin speech recognition
３．学会等名 IEEE SLT2022（国際学会）
４．発表年 2022年

１．発表者名 P Shen, X Lu, H Kawai
２．発表標題 Transducer-based language embedding for spoken language identification
３．学会等名 Interspeech2022（国際学会）
４．発表年 2022年

１．発表者名 X. Lu, P. Shen, Y. Tsao, H. Kawai
２．発表標題 Siamese Neural Network with Joint Bayesian Model Structure for Speaker Verification
３．学会等名 APASIPA ASC（国際学会）
４．発表年 2021年

１．発表者名
X. Lu, P. Shen, Y. Tsao, H. Kawai

２．発表標題
Unsupervised neural adaptation model based on optimal transport for spoken language identification

３．学会等名
IEEE ICASSP2022（国際学会）

４．発表年
2022年

１．発表者名
X. Lu, P. Shen, Y. Tsao, H. Kawai

２．発表標題
Cross-modal alignment with optimal transport for CTC-based ASR

３．学会等名
IEEE ASRU2023（国際学会）

４．発表年
2023年

１．発表者名
P Shen, X Lu, H Kawai

２．発表標題
Generative linguistic representation for spoken language identification

３．学会等名
IEEE ASRU2023（国際学会）

４．発表年
2023年

１．発表者名
X. Lu, P. Shen, Y. Tsao, H. Kawai

２．発表標題
Hierarchical cross-modality knowledge transfer with Sinkhorn attention for CTC-based ASR

３．学会等名
IEEE ICASSP2024（国際学会）

４．発表年
2024年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関