３版

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　３年　６月１４日現在

機関番号：８２６３６

研究種目：研究活動スタート支援

研究期間：2019～2020

課題番号：１９Ｋ２４３７６

研究課題名（和文）Next generation multilingual End-to-End speech recognition (from G30 to G200)

研究課題名（英文）Next generation multilingual End-to-End speech recognition (from G30 to G200)

研究代表者

李　勝（Li, Sheng）

国立研究開発法人情報通信研究機構・先進的音声翻訳研究開発推進センター 先進的音声技術研究室・研究員

研究者番号：７０８４０９４０

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：市販のASR対応製品の多くは、英語、フランス語、中国語、日本語など、特定のポピュラーな言語を対象としています。しかし、ASEAN諸国の言語のように、一般的ではない言語の音声認識については、今後も研究が必要です。グローバルな国際化に伴い、地域のイベントや文化交流、お祭りなど、現実に多言語コミュニケーションが必要な場面が増えています。

提案されたプロジェクトでは、リソースの少ないデータの問題に取り組み、現在の最先端のEnd-to-Endモデリングフレームワークの下で、多くの言語を1つのモデルでモデリングすることに焦点を当てました。また、これらの問題を徹底的に調査しました。

研究成果の学術的意義や社会的意義
This research shows we can integrate linguistic knowledge into the neural network instead of adding more layers or enlarging the model size. The proposed method is universally available for broad tasks for Society 5.0 (such as multilingual speech recognition, disordered speech recognition).

研究成果の概要（英文）：As the most natural way of communication, voice interface with the support of automatic speech recognition (ASR) technology has become crucial in human-computer interaction (HCI) in various devices of today's high-digitized society. Most commercial ASR-enabled products focus on specific popular languages such as English, French, Chinese, Japanese. The speech recognition of less popular languages, such as the ASEAN languages, is still a topic worthy of continued research. Global internationalization raises many real-life situations of multilingual communication, such as regional events, cultural exchanges, festivals.

The proposed project focused on tackling the problems of the low-resource data and modeling many languages in a single model under the current state-of-the-art End-to-End modeling framework. We also made an in-depth investigation of these problems.

研究分野：知覚情報処理

キーワード：speech recognition　multilingual　articulation　End-to-End

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

This project will focus on tackling the problems of the low-resource language (e.g., ASEAN languages) and modeling languages as many as we can (hundreds of languages from all language families) in a single model under the current state-of-the-art End-to-End automatic speech recognition (ASR) framework.

２．研究の目的

Most commercial automatic speech recognition (ASR)-enabled products to focus on specific popular languages such as English, French, Chinese, Japanese. The speech recognition of less popular languages, such as the ASEAN languages, is still a topic worthy of continued research. Global internationalization raises many real-life situations of multilingual communication, such as regional events, cultural exchanges, festivals.

This project will focus on tackling the problems of the low-resource data and modeling many languages in a single model under the current state-of-the-art End-to-End modeling framework. We will also make an in-depth investigation of these problems.

３．研究の方法

Current work from industry and academia are fascinating with huge self-attentional End-to-End models, and huge multilingual dataset as shown in Figure 1. Recent works from Facebook-AI greatly improved the performance with self-supervised learning and data augmentation from GAN model.

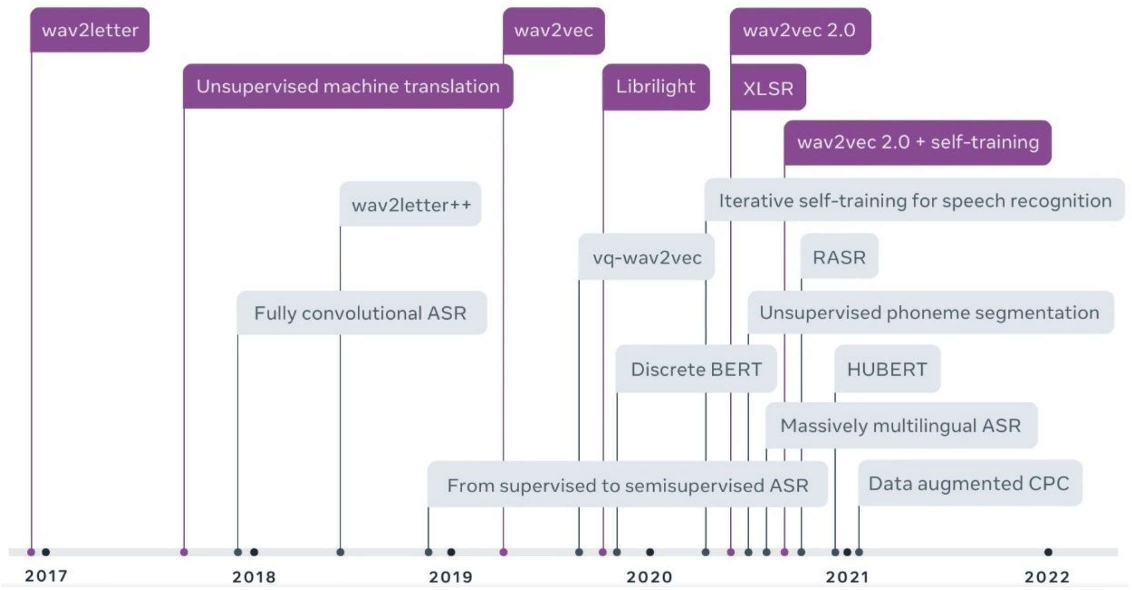
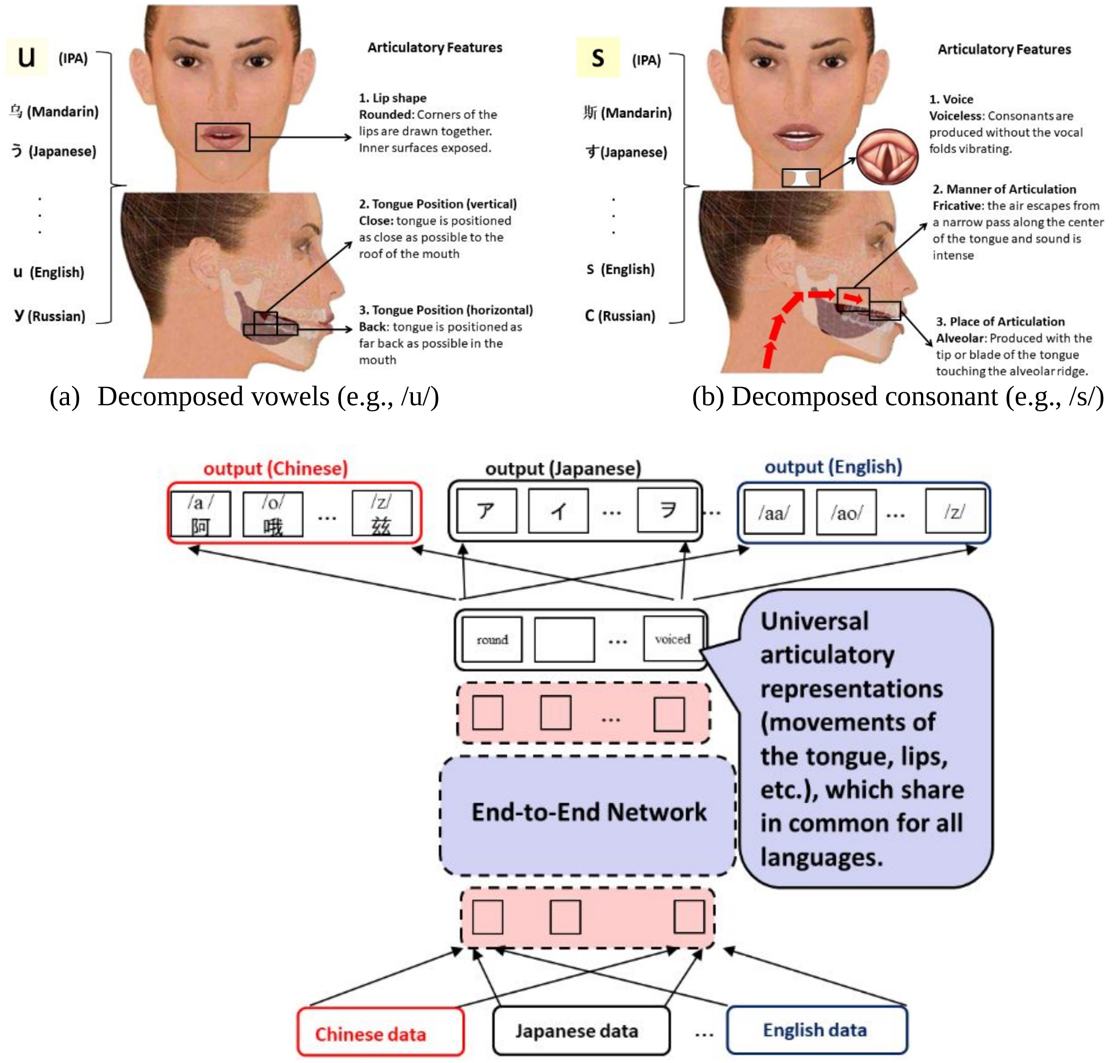


Figure 1: Facebook AI's development timeline of several years of work in speech recognition models, data sets, and training techniques. (cite from https://ai.facebook.com/blog/wav2vec-unsupervised-speech-recognition-without-supervision)

No one can deny the contribution from these works. However, these industry works are quite misleading to common researchers. Someday we can no longer rely on adding more layers or enlarging the model and data size. Our research showed there is another way as alternative: **integrating linguistic knowledge to the neural network.**

In our proposed method, we propose **universal articulatory representation for all of the languages.** We decompose the phones of different languages into "**atom units**": articulatory units as shown in Figure 2 (a, b). Then we can universally model all of the languages together using End-to-End framework. When recognition, we first get articulatory representations, then recover the texts of multilingual languages as shown in Figure 1 (c).


u (IPA)
乌 (Mandarin)
う (Japanese)
u (English)
y (Russian)
Articulatory Features
1. Lip shape
Rounded: Corners of the lips are drawn together. Inner surfaces exposed.
2. Tongue Position (vertical)
Close: tongue is positioned as close as possible to the roof of the mouth
3. Tongue Position (horizontal)
Back: tongue is positioned as far back as possible in the mouth
s (IPA)
斯 (Mandarin)
す(Japanese)
s (English)
C (Russian)
Articulatory Features
1. Voice
Voiceless: Consonants are produced without the vocal folds vibrating.
2. Manner of Articulation
Fricative: the air escapes from a narrow pass along the center of the tongue and sound is intense
3. Place of Articulation
Alveolar: Produced with the tip or blade of the tongue touching the alveolar ridge.
output (Chinese)
/a/ 阿
/o/ 哦
/z/ 兹
output (Japanese)
ア
イ
ヲ
output (English)
/aa/
/ao/
/z/
round
voiced
Universal articulatory representations (movements of the tongue, lips, etc.), which share in common for all languages.
End-to-End Network
Chinese data
Japanese data
English data

(a) Decomposed vowels (e.g., /u/)

(b) Decomposed consonant (e.g., /s/)

(c) End-to-End multilingual modeling using universal articulatory representation

Figure 2: Universal articulatory representation for all of the languages (a, b) and proposed End-to-End multilingual modeling (c)

This method can effectively represent all of the languages. It has several advantages:

(1) No matter how many languages mix, they can be represented using a universal symbolic set around 20. This technology can make a very compressed neural network output layer for memory limit devices.

(2) Our proposed universal articulatory representation can share knowledge between different languages more effectively. The low-resource languages can also benefit from the other rich sourced languages.

(3) I am so happy to find that the proposed method is universally available for various wide tasks (multilingual mispronunciation detection, multilingual speech recognition, disordered speech recognition, language identification, and speaker diarization). This is our contribution to the community.

４．研究成果

In FY2019, I focus on algorithm optimization. I am so happy to find that the proposed method is universally available for various wide tasks (multilingual mispronunciation detection, multilingual speech recognition, disordered speech recognition, language identification, and speaker diarization). Achievements are as follows:

(1) Two international papers were accepted in ICASSP2020, one joint first author and one corresponding author.
(2) One co-authored with an equal contribution in ICME2020.
(3) One first author, and one co-author in Speaker Odyssey2020.
(4) There are also three domestic presentations were reported in ASJ2020.
(5) The patents and book chapters are also included.

In FY2020, I focus on accent speech recognition (English and Chinese), cross-language family speech recognition. Multilingual speech recognition technologies have also been applied to language identification, speaker recognition, disordered speech recognition, and more complex tasks, such as speech translation and adversarial attack. Achievements are as follows:

(1) Multilingual modeling technology has been applied to speaker modeling (1 domestic presentation: IEICE-SP), low-resource transfer learning (1 Interspeech SLIMT2020), and speech translation (NLP2021 presentation), language identification (1 journal paper of IEEE-TASLP), and disordered speech recognition (1 Interspeech2020 with grant honor, 1 O-COCOSDA).

(2) I also find the acoustic modeling unit selection technology can enhance single-language speech recognition with multi-unit (1 invited full paper on 1 Interspeech SLIMT2020, 1 ICASSP2021) and code-switched speech synthesis (1 Interspeech SLIMT2020, 1 ICONIP paper).

(3) Following researches also benefit from the multilingual modeling technologies: speech separation (1 Interspeech2020 with grant honor), adversarial attack (1 IEEE-SLT demo paper), voice-privacy (1 invited report on Interspeech SLIMT2020, 1 Interspeech challenge, 1 ACM-CCS demo), voice activity detection (1 ICASSP2021), Mandarin tone modeling (1 ICASSP2021).

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 P. Shen, X. Lu, S. Li, H. Kawai.	４．巻 28
２．論文標題 Knowledge Distillation-based Representation Learning for Short-Utterance Spoken Language Identification	５．発行年 2020年
３．雑誌名 IEEE/ACM Trans. Audio, Speech and Language Process.	６．最初と最後の頁 2674 - 2683
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TASLP.2020.3023627	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24件（うち招待講演　4件／うち国際学会　19件）

１．発表者名 S. Li, X. Lu, R. Dabre, P. Shen and H. Kawai
２．発表標題 Joint Training End-to-End Speech Recognition Systems with Speaker Attributes.
３．学会等名 ISCA-Odyssey (The Speaker and Language Recognition Workshop)（国際学会）
４．発表年 2020年

１．発表者名 P. Shen, X. Lu, K. Sugiura, S. Li and H. Kawai.
２．発表標題 Compensation on x-vector for short utterance spoken language identification.
３．学会等名 ISCA-Odyssey (The Speaker and Language Recognition Workshop)（国際学会）
４．発表年 2020年

１．発表者名 Y. Han, S. Li, Y. Cao, Q. Ma and M. Yoshikawa.
２．発表標題 Voice-Indistinguishability: Protecting Voiceprint in Privacy Preserving Speech Data Release.
３．学会等名 IEEE-ICME（国際学会）
４．発表年 2020年

1．発表者名
Y. Lin, L. Wang, J. Dang, S. Li, and C. Ding.

2．発表標題
End-To-End Articulatory Modeling for Dysarthria Articulatory Attribute Detection.

3．学会等名
IEEE-ICASSP（国際学会）

4．発表年
2020年

1．発表者名
H. Shi, L. Wang, M. Ge, S. Li, and J. Dang.

2．発表標題
Spectrograms Fusion with Minimum Difference Masks Estimation for Monaural Speech Dereverberation.

3．学会等名
IEEE-ICASSP（国際学会）

4．発表年
2020年

1．発表者名
S. Li, C. Ding, X. Lu, P. Shen and H. Kawai,

2．発表標題
End-to-End Articulatory Attribute Modeling for Low-resource Multilingual Speech Recognition,

3．学会等名
Acoustical Society of Japan, spring, 2020.

4．発表年
2020年

1．発表者名
S. Li, X. Lu, R. Dabre, P. Shen and H. Kawai,

2．発表標題
Joint Training End-to-End Systems for Speech and Speaker Recognition with Speaker Attributes,

3．学会等名
Acoustical Society of Japan, spring, 2020.

4．発表年
2020年

１．発表者名
P. Shen, X. Lu, K. Sugiura, S. Li, H. Kawai,

２．発表標題
Improvement of x-vector for short utterance spoken language identification,

３．学会等名
Acoustical Society of Japan, spring, 2020.

４．発表年
2020年

１．発表者名
N. Li, L. Wang, M. Unoki, S. Li, R. Wang, M. Ge, J. Dang,

２．発表標題
Robust voice activity detection using a masked auditory encoder based convolutional neural network.

３．学会等名
IEEE-ICASSP, 2021（国際学会）

４．発表年
2021年

１．発表者名
S. Chen, X. Hu, S. Li, X. Xu,

２．発表標題
An investigation of using hybrid modeling units for improving End-to-End speech recognition systems.

３．学会等名
IEEE-ICASSP, 2021（国際学会）

４．発表年
2021年

１．発表者名
H. Huang, K. Wang, Y. Hu, S. Li,

２．発表標題
Encoder-Decoder based pitch tracking and joint model training for Mandarin tone classification.

３．学会等名
IEEE-ICASSP, 2021（国際学会）

４．発表年
2021年

１．発表者名
H. Zhang, S. Li, X. Ma, Y. Zhao, Y. Cao, T. Kawahara,

２．発表標題
Phantom in the Opera: Effective Adversarial Music Attack on Keyword Spotting Systems.

３．学会等名
IEEE-SLT, 2021（国際学会）

４．発表年
2021年

１．発表者名
K. Soky, S. Li, M. Mimura, C. Chu, T. Kawahara,

２．発表標題
Comparison of End-to-End Models for Joint Speaker and Speech Recognition

３．学会等名
IEICE-SP, 2021.

４．発表年
2021年

１．発表者名
K. Soky, S. Li, T. Kawahara, S. Seng,

２．発表標題
Multilingual transformer training for Khmer automatic speech recognition

３．学会等名
Interspeech 2020 Satellite Workshop (SLIMTS2020)（招待講演）（国際学会）

４．発表年
2020年

１．発表者名
S. Shimizu, C. Chu, S. Li, S. Kurohashi,

２．発表標題
End-to-End Speech Translation with Cross-lingual Transfer Learning

３．学会等名
NLP, 2021.

４．発表年
2020年

１．発表者名
S. Guo, L. Wang, S. Li, J. Zhang, C. Gong, Y. Wang, J. Dang, K. Honda

２．発表標題
Effectively Synthesizing Code-switched Speech Using Highly Imbalanced Mix-lingual Data and mask embedding

３．学会等名
Interspeech 2020 Satellite Workshop (SLIMTS2020)（招待講演）（国際学会）

４．発表年
2020年

１．発表者名
H. Zhang, S. Ueno, M. Mimura, S. Li, W. Zhang, T. Kawahara,

２．発表標題
A Mixture of Character and Word End-to-End System for Keyword Spotting

３．学会等名
Interspeech 2020 Satellite Workshop (SLIMTS2020)(full paper).（招待講演）（国際学会）

４．発表年
2020年

１．発表者名
S. Guo, L. Wang, S. Li, J. Zhang, C. Gong, Y. Wang, J. Dang, K. Honda.

２．発表標題
Effectively Synthesizing Code-switched Speech Using Highly Imbalanced Mix-lingual Data

３．学会等名
In Proc. ICONIP, 2020（国際学会）

４．発表年
2020年

１．発表者名
Y. Lin, L. Wang, S. Li, J. Dang, and C. Ding.

２．発表標題
Staged Knowledge Distillation for End-to-End Dysarthric Speech Recognition and Speech Attribute Transcription

３．学会等名
In Proc. INTERSPEECH, 2020 (Travel Granted by ISCA).（国際学会）

４．発表年
2020年

１．発表者名
A. Thida, N. Han, S. Oo, S. Li and C. Ding.

２．発表標題
VOIS: The First Speech Therapy App in the World for Myanmar Hearing-Impaired Children.

３．学会等名
In Proc. O-COCOSDA, 2020.（国際学会）

４．発表年
2020年

１．発表者名
Y. Han, Y. Cao, S. Li, Q. Ma, M. Yoshikawa.

２．発表標題
Voice-Indistinguishability: Protecting Voiceprint in Privacy-Preserving Speech Data Release,

３．学会等名
Interspeech 2020 Satellite Workshop (SLIMTS2020) (invited report).（招待講演）（国際学会）

４．発表年
2020年

１．発表者名
Y. Han, Y. Cao, S. Li, Q. Ma, M. Yoshikawa.

２．発表標題
Voice-Indistinguishability: Protecting Voiceprint with Differential Privacy under an Untrusted Server.

３．学会等名
ACM conference on Computer and Communications Security (CCS), demo, 2020.（国際学会）

４．発表年
2020年

１．発表者名
Y. Han, S. Li, Y. Cao, M. Yoshikawa,

２．発表標題
System Description for Voice Privacy Challenge (Kyoto Team).

３．学会等名
In special session of INTERSPEECH 2020 (VoicePrivacy challenge 2020).（国際学会）

４．発表年
2020年

１．発表者名 H. Shi, L. Wang, S. Li, C. Ding, M. Ge, N. Li, J. Dang, and H. Seki.
２．発表標題 Singing Voice Extraction with Attention based Spectrograms Fusion.
３．学会等名 In Proc. INTERSPEECH, 2020 (Travel Granted by ISCA).（国際学会）
４．発表年 2020年

〔図書〕　計1件

１．著者名 X. Lu, S. Li, M. Fujimoto	４．発行年 2020年
２．出版社 Springer Singapore	５．総ページ数 18
３．書名 Automatic speech recognition	

〔出願〕　計4件

産業財産権の名称 推論器、推論プログラムおよび学習方法	発明者 李勝、 ルーシュガン、 丁塵辰、 河原達也、 河井恒	権利者 国立研究開発法人情報通信研究機構
産業財産権の種類、番号 特許、特願2019-163555	出願年 2019年	国内・外国の別 国内

産業財産権の名称 推論器および推論器の学習方法	発明者 李勝、ルーシュガン、河井恒	権利者 国立研究開発法人情報通信研究機構
産業財産権の種類、番号 特許、特願2020-059962	出願年 2020年	国内・外国の別 国内

産業財産権の名称 言語識別モデルの訓練方法及び装置、並びにそのためのコンピュータプログラム	発明者 沈 鵬, ルー シュガン , 李 勝 , 河井恒	権利者 国立研究開発法人情報通信研究機構
産業財産権の種類、番号 特許、特願2019-086005	出願年 2019年	国内・外国の別 国内

産業財産権の名称 推論器、学習方法および学習プログラム	発明者 李勝、 ルーシュガン、 ダブレラジ、 河井恒	権利者 国立研究開発法人情報通信研究機構
産業財産権の種類、番号 特許、特願2019-051008	出願年 2019年	国内・外国の別 国内

〔取得〕　計0件

〔その他〕

publication information on DBLP
https://dblp.dagstuhl.de/pid/23/3439-10.html
Google scholar homepage
https://scholar.google.com/citations?hl=en&user=zHAhs0IAAAAJ
researchmap homepage
https://researchmap.jp/listen
NICT researcher's homepage
https://ast-astrec.nict.go.jp/aboutus/member/sheng-li/index.html
researchgage researcher's homepage
https://www.researchgate.net/profile/Sheng-Li-60
The paper urls can be found in these pages.

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			
中国	Tianjin University	Xinjiang University	Hithink RoyalFlush AI	