様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科研費 KAKENHI

科学研究費助成事業　　研究成果報告書

令和　５ 年　６ 月　６ 日現在

機関番号：８２６３６

研究種目：若手研究

研究期間：2021～2022

課題番号：２１Ｋ１７８３７

研究課題名（和文）Phantom in the Opera: the Vulnerabilities of Speech Interface for Robotic Dialogue System

研究課題名（英文）Phantom in the Opera: the Vulnerabilities of Speech Interface for Robotic Dialogue System

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・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・研究員

研究者番号：７０８４０９４０

交付決定額（研究期間全体）：（直接経費）　1,600,000円

研究成果の概要（和文）：このプロジェクトでは、音声認識システムの原理を研究し、攻撃の詳細を調査。調査結果をまとめ、音声認識システムの改善方法を提案しました。研究範囲を普遍的視点で拡大し、音声関連システムにも同様の攻撃が共存できることを示しました。敵対的攻撃をノイズとみなし、音声強化・モデリング・ポストプロセッシング法を組み合わせて対処。InterspeechやICASSPなどが成果を認め、2冊の本に紹介され、関西国立図書館に収められました。これはAIシステムの安全性と信頼性の確保に貢献しています。

研究成果の学術的意義や社会的意義
The development of deep neural networks has been progressing rapidly and the evolution of speech recognition systems has been incredibly fast. The study aims to provide researchers with ideas on improving system security in light of the increasingly severe security issues.

研究成果の概要（英文）：In this project, we carefully studied the principles of speech recognition systems and researched all possible attack details. We summarized our findings in a review and proposed methods for improving the front-end and back-end of speech recognition systems.
We expanded our research scope with a universal point of view. Similar attacks can co-exist in speech-related systems, not just speech recognition systems. We also consider adversarial attacks as particular noise, then combining traditional speech enhancement, modeling, and post-processing methods in system development can sufficiently deal with this attack.
Top journals and conferences in the speech field accepted our achievements, such as Interspeech and ICASSP. Above two years of research achievement have been introduced into two books (ISBN: 978-4-904020-26-5, ISBN: 978-4-904020-28-9) by NICT and stored in the national library Kansai. These efforts are our contribution to ensuring the security and reliability of AI systems.

研究分野：知覚情報処理

キーワード：speech recognition　adversarial attack　privacy perserving　deepfake detection　spoken dialogue　federated learning　security

※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。

１．研究開始当初の背景

Nowadays, spoken dialogue systems have been widely used in our society. The automatic speech recognition (ASR) module, as these dialogue systems' most natural human-machine interface, is based on deep neural networks (DNNs). The development of deep neural networks has been progressing rapidly, from basic DNNs to end-to-end models, to the popular self-supervised learning models in recent years, and now to the revolutionary large language models (LLMs). However, DNN is known to be vulnerable to adversarial examples (or attacks). This is a severe problem.

２．研究の目的

The study of adversarial attacks on DNN-based speech recognition systems serves as a starting point for our research, aiming to provide speech development researchers with ideas on improving system security in light of the increasingly severe security issues related to speech-based systems.

３．研究の方法

Current academia is focusing on designing individual purposed perfect adversarial attacks on Automatic Speech Recognition (ASR) systems while still sounding normal to human listeners with the following focuses:

1. Black-box attacks: In black-box attacks, the attacker has limited or no access to the internals of the target model, such as its architecture or weights. They craft adversarial examples by exploiting the transferability of adversarial perturbations across models or by querying the target model and using its output to estimate its behavior.

2. Untargeted attacks: Adversarial attacks on speech systems can be categorized into targeted and untargeted attacks. Targeted attacks aim to make the system produce a specific, desired output, while untargeted attacks cause incorrect output without a specific target.

3. Real-world applicability: Researchers are increasingly focusing on adversarial attacks that can be effective in real-world settings, such as over-the-air attacks or attacks robust to various environmental conditions like background noise.

These works studying individual threats are very important contributions to the community. However, **other substantial threats co-exist in real-world scenarios and adversarial attacks**. These attacks can compromise the integrity of machine learning models and seriously affect critical applications, such as autonomous vehicles, healthcare, and security systems.

In our proposed method, we propose **a universal point of view on dealing with these threats**.

1. **Defense of the whole system pipeline and taking adversarial audio as noise**: The speech recognition module is not isolated in the industry system pipeline. Taking adversarial audio as a type of noise, we confirm that three traditional methods can effectively deal with adversarial attacks: 1. Changing/detecting front-end input signal (denoising, adding noise, or detect-then-reject), 2. Changing models (tuning structure or retraining), and 3. Correct by back-end language models. All these methods exist in current speech systems.

2. **Extend to concerning the robustness of other speech-related models**: Similar to the adversarial attack, other substantial threats exist. **(1) Deepfakes** are manipulated media, particularly images and videos, created using advanced

machine-learning techniques. Deepfakes can spread misinformation, manipulate public opinion, and harm reputations, posing a severe threat to individuals, organizations, and society. **(2) Privacy leakage** refers to unauthorized access or disclosure of sensitive personal information, often resulting from inadequate data protection measures or malicious activities. Privacy leakage can lead to identity theft, financial loss, and damage to an individual's reputation. **(3) Adversarial attacks to other speech models:** the same immediate impact as deepfakes or privacy leakage, adversarial attacks on speaker recognition systems can still be a significant threat.

４．研究成果

In the first year of this project, we carefully studied the principles of speech recognition systems and researched all possible attack details. We summarized our findings in a review and proposed methods for improving the front-end and back-end of speech recognition systems. Our achievements were accepted by top conferences in the speech field, such as Interspeech and ICASSP, as followings:

- We construct speech recognition systems with recent popular training toolkits and neural network types (accepted in Journals and conferences, e.g., ICASSP2022).
- We did surveys for the current attack methods (accepted in Journal).
- We implement robust adversarial attacks using the Kaldi-based ASR systems (accepted in SLT2020).
- We are also happy to see that this framework can be used to protect sensitive speech content (accepted in LREC 2022).
- To defend against attacks, we find that adversarial audios are very sensitive. Moreover, the feature of its spectrogram is very different from the human voice, and it can be treated as a special kind of noise. We constructed speech enhancement systems and studied their mechanism this year (accepted in Journals and conferences, e.g., ICASSP2022).

In the project's second year, we discovered that attacks on speech recognition systems are not only limited to the content of speech recognition but also involve more attacks related to speaker attributes. Therefore, we expanded our research scope with a universal point of view.

- We continued research on the front-end method about synthesized speech quality estimation, signal-to-noise ratio estimation framework (accepted in conferences, such as APSIPA2022, EUSIPCO2022, and Interspeech2022), and back-end of the language model, speech recognition/translation systems and other down-streaming systems (accepted in Journals and conferences, e.g., JNLP, International Journal of Asian Language Processing, Interspeech2022, ICASSP2023).
- We participated in several privacy attack competitions, proposing our deepfake detection (accepted by Interspeech2022). Moreover, we published our first database following security protection law (O-COCOSDA2022).
- At the same time, we expanded our research from speech recognition systems to spoken dialogue systems (accepted in Sigdial2022).
- We also studied the state-of-the-art modeling training method, self-supervised training (accepted in Interspeech2022), and federated learning (accepted in ICASSP2023).

Above two years of research achievement have been introduced into two books (ISBN: 978-4-904020-26-5, ISBN: 978-4-904020-28-9) by NICT and stored in the national library Kansai.

These efforts have laid the foundation for further research to ensure the security and reliability of AI systems.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　2件／うちオープンアクセス　4件）

１．著者名 Shimizu Shuichiro、Chu Chenhui、Li Sheng、Kurohashi Sadao	４．巻 29
２．論文標題 Cross-Lingual Transfer Learning for End-to-End Speech Translation	５．発行年 2022年
３．雑誌名 Journal of Natural Language Processing	６．最初と最後の頁 611～637
掲載論文のDOI（デジタルオブジェクト識別子） 10.5715/jnlp.29.611	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Soky Kak、Mimura Masato、Kawahara Tatsuya、Chu Chenhui、Li Sheng、Ding Chenchen、Sam Sethserey	４．巻 31
２．論文標題 TriECCC: Trilingual Corpus of the Extraordinary Chambers in the Courts of Cambodia for Speech Recognition and Translation Studies	５．発行年 2022年
３．雑誌名 International Journal of Asian Language Processing	６．最初と最後の頁 1～21
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S2717554522500072	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Chen Xiaojiao、Li Sheng、Huang Hao	４．巻 11
２．論文標題 Adversarial Attack and Defense on Deep Neural Network-Based Voice Processing Systems: An Overview	５．発行年 2021年
３．雑誌名 Applied Sciences	６．最初と最後の頁 8450-8450
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/app11188450	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Qin Siqing、Wang Longbiao、Li Sheng、Dang Jianwu、Pan Lixin	４．巻 -
２．論文標題 Improving low-resource Tibetan end-to-end ASR by multilingual and multilevel unit modeling	５．発行年 2022年
３．雑誌名 EURASIP Journal on Audio, Speech, and Music Processing	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s13636-021-00233-4	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計28件（うち招待講演　0件／うち国際学会　28件）

１．発表者名
Kai Li, Xugang Lu, Masato Akagi, Jianwu Dang, Sheng Li, Masashi Unoki

２．発表標題
Relationship Between Speakers' Physiological Structure and Acoustic Speech Signals: Data-Driven Study Based on Frequency-Wise Attentional Neural Network

３．学会等名
30th European Signal Processing Conference (EUSIPCO)（国際学会）

４．発表年
2022年

１．発表者名
Kak Soky, Sheng Li, Masato Mimura, Chenhui Chu, Tatsuya Kawahara

２．発表標題
Leveraging Simultaneous Translation for Enhancing Transcription of Low-resource Language via Cross Attention Mechanism

３．学会等名
INTERSPEECH 2022（国際学会）

４．発表年
2022年

１．発表者名
Longfei Yang, Wenqing Wei, Sheng Li, Jiyi Li, Takahiro Shinozaki

２．発表標題
Augmented Adversarial Self-Supervised Learning for Early-Stage Alzheimer's Speech Detection

３．学会等名
INTERSPEECH 2022（国際学会）

４．発表年
2022年

１．発表者名
Kai Li, Sheng Li, Xugang Lu, Masato Akagi, Meng Liu, Lin Zhang, Chang Zeng, Longbiao Wang, Jianwu Dang, Masashi Unoki

２．発表標題
Data Augmentation Using McAdams-Coefficient-Based Speaker Anonymization for Fake Audio Detection

３．学会等名
INTERSPEECH 2022（国際学会）

４．発表年
2022年

１．発表者名
Zhengdong Yang, Wangjin Zhou, Chenhui Chu, Sheng Li, Raj Dabre, Raphael Rubino, Yi Zhao

２．発表標題
Fusion of Self-supervised Learned Models for MOS Prediction

３．学会等名
INTERSPEECH 2022（国際学会）

４．発表年
2022年

１．発表者名
Hao Shi, Longbiao Wang, Sheng Li, Jianwu Dang, Tatsuya Kawahara

２．発表標題
Monaural Speech Enhancement Based on Spectrogram Decomposition for Convolutional Neural Network-sensitive Feature Extraction

３．学会等名
INTERSPEECH 2022（国際学会）

４．発表年
2022年

１．発表者名
Longfei Yang, Jiyi Li, Sheng Li, Takahiro Shinozaki

２．発表標題
Multi-Domain Dialogue State Tracking with Top-k Slot Self Attention

３．学会等名
SIGdial Meeting Discourse and Dialogue 2022（国際学会）

４．発表年
2022年

１．発表者名
Kak Soky, Zhuo Gong, Sheng Li

２．発表標題
Nict-Tib1: A Public Speech Corpus Of Lhasa Dialect For Benchmarking Tibetan Language Speech Recognition Systems

３．学会等名
25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech
Databases and Assessment Techniques (O-COCOSDA)（国際学会）

４．発表年
2022年

１．発表者名
Hao Shi, Longbiao Wang, Sheng Li, Jianwu Dang, Tatsuya Kawahara

２．発表標題
Subband-based Spectrogram Fusion for Speech Enhancement by Combining Mapping and Masking Approaches

３．学会等名
Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)（国際学会）

４．発表年
2022年

１．発表者名
Zhuo Gong, Saito Daisuke, Sheng Li, Hisashi Kawai, Minematsu Nobuaki

２．発表標題
Can We Train a Language Model Inside an End-to-End ASR Model? - Investigating Effective Implicit Language Modeling

３．学会等名
Proceedings of the Second Workshop on When Creative AI Meets Conversational AI（国際学会）

４．発表年
2022年

１．発表者名
Chao Tan, Yang Cao, Sheng Li and Masatoshi Yoshikawa

２．発表標題
GENERAL OR SPECIFIC? INVESTIGATING EFFECTIVE PRIVACY PROTECTION IN FEDERATED LEARNING FOR SPEECH EMOTION RECOGNITION

３．学会等名
ICASSP（国際学会）

４．発表年
2023年

１．発表者名
Kak Soky, Sheng Li, Chenhui Chu, Tatsuya Kawahara

２．発表標題
DOMAIN AND LANGUAGE ADAPTATION USING HETEROGENEOUS DATASETS FOR WAV2VEC2.0-BASED SPEECH RECOGNITION OF LOW-RESOURCE LANGUAGE

３．学会等名
ICASSP（国際学会）

４．発表年
2023年

１．発表者名
Z. Gong, D. Saito, L. Yang, T. Shinozaki, S. Li, H. Kawai and N. Minematsu

２．発表標題
Self-Adaptive Multilingual ASR Rescoring with Language Identification and Unified Language Model

３．学会等名
ISCA-Odyssey (The Speaker and Language Recognition Workshop)（国際学会）

４．発表年
2022年

１．発表者名
S. Li, J. Li, Q. Liu and Z. Gong

２．発表標題
Adversarial Speech Generation and Natural Speech Recovery for Speech Content Protection

３．学会等名
LREC (Language Resources and Evaluation Conference)（国際学会）

４．発表年
2022年

１．発表者名
Y. Lv, L. Wang, M. Ge, S. Li, C. Ding, L. Pan, Y. Wang, J. Dang, K. Honda

２．発表標題
Compressing Transformer-based ASR Model by Task-driven Loss and Attention-based Multi-level Feature Distillation

３．学会等名
in Proc. IEEE-ICASSP, pp. 7992--7996, 2022.（国際学会）

４．発表年
2022年

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

２．発表標題
Mining Hard Samples Locally and Globally for Improved Speech Separation

３．学会等名
in Proc. IEEE-ICASSP, pp. 6037--6041, 2022.（国際学会）

４．発表年
2022年

１．発表者名
H. Shi, L. Wang, S. Li, C. Fan, J. Dang, and T. Kawahara

２．発表標題
Spectrograms Fusion-based End-to-End Robust Automatic Speech Recognition

３．学会等名
In Proc. APSIPA ASC, pp. 438--442, 2021.（国際学会）

４．発表年
2021年

１．発表者名
Y. Peng, J. Zhang, H. Zhang, H. Xu, H. Huang, S. Li, and E.S. Chng

２．発表標題
Multilingual Approach to Joint Speech and Accent Recognition with DNN-HMM Framework

３．学会等名
In Proc. APSIPA ASC, pp. 1043--1048, 2021.（国際学会）

４．発表年
2021年

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

２．発表標題
On the Use of Speaker Information for Automatic Speech Recognition in Speaker-imbalanced Corpora

３．学会等名
In Proc. APSIPA ASC, pp. 433--437, 2021.（国際学会）

４．発表年
2021年

１．発表者名
D. Wang, S. Ye, X. Hu, S. Li, and X. Xu

２．発表標題
An End-to-End Dialect Identification System with Transfer Learning from a Multilingual Automatic Speech Recognition Model

３．学会等名
in Proc. INTERSPEECH, pp. 3266--3270, 2021.（国際学会）

４．発表年
2021年

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

２．発表標題
End-to-End Speech Separation Using Orthogonal Representation in Complex and Real Time- Frequency Domain

３．学会等名
in Proc. INTERSPEECH, pp. 3046--3050, 2021.（国際学会）

４．発表年
2021年

１．発表者名
S. Li, R. Dabre, R. Raphael, W. Zhou, Z. Yang, C. Chu, Y. Zhao

２．発表標題
The System Description for VoiceMOS Challenge 2022 (KK team, main/ood tasks)

３．学会等名
VoiceMOS Challenge 2022（国際学会）

４．発表年
2022年

１．発表者名
D. Wang, S. Ye, X. Hu, S. Li

２．発表標題
The RoyalFlush-NICT System Description for AP21-OLR Challenge (Silk-road team, full tasks)

３．学会等名
OLR2021 (oriental language recognition challenge)（国際学会）

４．発表年
2021年

１．発表者名
W. Wei, R. Wong, S. Li, Y. Guo and H. Huang

２．発表標題
System description of Alzheimer's disease early detection (Silk-road team, short speech track)

３．学会等名
In special session of NCMMSC2021 (Alzheimer's disease detection challenge), 2021（国際学会）

４．発表年
2021年

１．発表者名
X. Chen, H. Huang, and S. Li

２．発表標題
Adversarial Attack and Defense on Deep Neural Network-based Voice Processing Systems: An Overview

３．学会等名
National Conference on Man-Machine Speech Communication (NCMMSC), 2021. (report is selected to publish in Applied Sciences, Special Issues of Machine Speech Communication)（国際学会）

４．発表年
2021年

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

２．発表標題
Speech Dereverberation Based on Scale-aware Mean Square Error Loss

３．学会等名
International Conference on Neural Information Processing (ICONIP2021), pp 55-63, Springer, 2021.（国際学会）

４．発表年
2021年

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

２．発表標題
Simultaneous Progressive Filtering-based Monaural Speech Enhancement

３．学会等名
International Conference on Neural Information Processing (ICONIP2021), pp 213-221, Springer, 2021.（国際学会）

４．発表年
2021年

１．発表者名
D. Liu, L. Wang, S. Li, H. Li, C. Ding, J. Zhang and J. Dang

２．発表標題
Exploring Effective Speech Representation via ASR for High-Quality End-to-End Multispeaker TTS

３．学会等名
International Conference on Neural Information Processing (ICONIP2021), pp 110-118, Springer, 2021.（国際学会）

４．発表年
2021年

〔図書〕　計2件

1．著者名 Sheng Li	4．発行年 2022年
2．出版社 NICT	5．総ページ数 112
3．書名 Voices of the Himalayas: Investigation of Speech Recognition Technology for the Tibetan Language	

1．著者名 Sheng Li	4．発行年 2022年
2．出版社 NICT	5．総ページ数 110
3．書名 Phantom in the Opera: The Vulnerabilities of Speech-based Artificial Intelligence Systems	

〔産業財産権〕

〔その他〕

The international collaboration with China is based on NICT's international collaboration activity in 2021. In 2022, we turn to Singapore according to NICT's rule.

情報通信研究機構の研究成果として、各年ごとの発表論文を日付順で紹介します。
https://www.nict.go.jp/outcome/journals/journals_2021_j.html
情報通信研究機構の研究成果として、各年ごとの発表論文を日付順で紹介します。
https://www.nict.go.jp/outcome/proceedings/proceedings_2021_j.html
google scholar of Sheng Li
https://scholar.google.com/citations?user=zHAhs0IAAAAJ&hl=en
Lab homepage of Sheng Li
https://ast-astrec.nict.go.jp/member/sheng-li/index.html

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	Royal Flush AI Research Inc.	
シンガポール	Nanyang Technological University			