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　 ５ 月 ２１ 日現在

機関番号：１３３０２
研究種目：研究活動スタート支援
研究期間：2022～2023
課題番号：２２Ｋ２１３０４

研究課題名（和文）Speech security on human-computer interaction

研究課題名（英文）Speech security on human-computer interaction

研究代表者
ＭＡＷＡＬＩＭ　ＣａｎｄｙＯｌｉｖｉａ（MAWALIM, CandyOlivia）

北陸先端科学技術大学院大学・先端科学技術研究科・助教

研究者番号：１０９６３７２０

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）：FY2022年度は、タイムスケール変換アルゴリズムを用いた話者匿名化手法を開発した。フェーズボーコーダ方式が音声特徴保持に有効で、プライバシーと明瞭度のバランス向上を実現した。また、機械学習モデルで匿名化処理と性別認識の影響を解析した。成果は3件の国際学会で発表した。
FY2023年度は、(1) 話者匿名化の目標設定と音声攻撃への対応、(2) なりすまし攻撃検出の新手法開発、(3) 音声知覚メカニズム解明による音声認識理解の研究を推進した。(3) の成果は「Journal of Applied Acoustics」掲載した。タイ語話者向けスプーフィングデータベース開発も計画している。

研究成果の学術的意義や社会的意義
Innovative techniques for speaker anonymization and spoofing detection open up new possibilities for voice privacy and security research. This research will greatly contributes to securing voice communication, strengthening authentication systems, and improving human-computer interaction.

研究成果の概要（英文）：In FY2022, we developed speaker anonymization methods using time-scale modification. The phase vocoder method is most effective for preserving voice characteristics. This method offered a better balance between privacy and speech intelligibility. Additionally, we analyzed the impact of anonymization on gender perception using a machine learning model. These findings were presented at three international conferences.
In FY2023, research focused on two areas: (1) addressing unclear goals in speaker anonymization and the variety of speech attacks. New methods for tackling spoofing in speaker verification systems were developed. These findings were presented at two conferences. (2) investigating human speech perception to understand how we perceive intelligibility. This research, published in the Journal of Applied Acoustics, lays the groundwork for detecting changes caused by speech synthesis. Finally, the project is expanding its scope to include developing a Thai language spoof database.

研究分野：speech security, voice privacy

キーワード：voice privacy　phase vocoder　speaker anonymization　speaker verification　spoof attacks　speech intelligibility　auditory model

１．研究開始当初の背景

Within human expressions, speech is considered one of the most important and effective ways of communication because people naturally communicate via spoken language. Consequently, speech technology is preferable to assist work efficiency and standard of living. For instance, an automatic speech recognition system could produce well-written transcription, which is beneficial in improving work productivity and assisting people with speech or visual impairment. Besides, a voice assistant is no longer uncommon for helping with human tasks, such as road navigation, shopping, and playing music.

As the English phrase "convenience comes with a price", the convenience of speech technology causes a critical issue related to security and privacy [1]. Speech data contains abundant information [2], including linguistic, paralinguistic, and non-linguistic information. Linguistic information is the information that can be explicitly represented as written text or implicitly inferred from the context. Paralinguistic information is the additional information apart from the linguistic information that the speaker intentionally adds. Non-linguistic information is uncontrollable information that is not directly related to both linguistic and paralinguistic information, such as age, gender, and physical states. Once the speech is exposed to the public, private information encapsulated in the speech will also be exposed, causing privacy violations. Hence, research in securing speech communication is essential in human-computer interaction (HCI).

２．研究の目的

The ultimate goal of this study is to provide a solution to the privacy violation problems in HCI, especially in dealing with speech data. Speech is usually transmitted via a public switched telephone network (PSTN) or voice over Internet protocol (VoIP). This transmission channel is vulnerable to various interceptions or attacks. Consequently, speech security countermeasures are indispensable, which are addressed as the subject of research in this study.

Speech security countermeasures involve the protection of voice privacy by speaker anonymization without degrading speech intelligibility and quality. The personal data encapsulated in voice data includes personality traits, emotions, skills, etc. This information is also part of social signals (informative or communicative signals provided through social interactions, emotions, behavior, and relationships [3]). Understanding social signals is essential for improving HCI. It helps the computer not only to accurately do its tasks (as most existing speech technology does) but also to better sense, understand, and respond intelligently and naturally to human emotional feedback. Accordingly, the social signals that are beneficial to HCI are maintained while anonymizing the speaker identity to protect voice privacy.

３．研究の方法

The purposes of this study are: (O1) to improve speech security in HCI and (O2) to investigate the social signals encapsulated in voice data. Initially, we address the methods for dealing with privacy violation issues in public speech communication channels using speaker anonymization (O1). The speech anonymization suppresses the speaker identity while maintaining other factors, such as quality and intelligibility, of the given input [1]. Generally, speech anonymization is comprised of speech analysis, feature modification, and synthesis processes. Input speech is analyzed prior to extraction of the acoustic features that perceptually related to personal data [2,4]. Subsequently, these acoustic features are modified for anonymization. Lastly, the speech synthesis by a generative model neural vocoder is performed to obtain the anonymized speech.

Understanding the relationship between speech and social signals is crucial [5], especially for speech technology that assists human activities. Yet, the definition and annotation of these social signals were often obscure [3]. Hence, we investigate the social signals in voice data that are beneficial in enhancing HCI (O2). In contrast to other existing studies, the phenomena and characteristics of the human perceptual system [4] will be considered in this study.

４．研究成果

Our first year focused on improving speaker anonymization using time-scale modification (TSM) algorithms. We explored different TSM methods and found that the phase vocoder-based approach was most effective due to its ability to preserve the harmonic structure of human voices, crucial for maintaining naturalness. To evaluate our proposed methods, we employed standard objective metrics from the VoicePrivacy Challenge. The results demonstrated that our phase vocoder-based TSM (PV-TSM) method offered a superior balance between privacy and speech utility compared to baseline systems. This was

particularly evident in anonymized enrollment and anonymized test scenarios (a-a) evaluated using an automatic speaker verification (ASV) system. Notably, our approach outperformed the x-vector-based speaker anonymization method, which suffers from limitations such as a complex training process, compromised privacy in a-a scenarios, and reduced speaker distinctiveness. The research findings were presented at the Voice Privacy Challenge 2022, held jointly with Interspeech 2022 [6].

We further investigated the impact of anonymization on gender perception using a gender classifier model trained on x-vector speaker embeddings. Objective evaluation confirmed that our proposed method effectively anonymized gender information. Additionally, compared to signal processing baselines, our methods achieved superior speaker anonymization (as measured by x-vectors in ASV evaluations) while maintaining good speech intelligibility. These results were presented at APSIPA 2022 conference [7].

Looking ahead, several key areas require further exploration. First, clearly defining the target level of speaker anonymization is crucial. Additionally, addressing the limitations of current anonymization methods against various attack models is critical for real-world applications. We aim to develop more robust and secure speaker anonymization techniques that can withstand potential attacks, thereby enabling broader and more secure deployments. Finally, ethical considerations and user privacy will remain paramount throughout the development process to ensure responsible anonymization practices.

Our second year addressed two key challenges in spoof detection: defining the target level of anonymization and mitigating diverse speech attacks. While humans can often distinguish genuine from spoofed speech, machines struggle due to difficulties in separating speech content from vocal tract information.

To tackle spoofing in speaker verification systems, we proposed novel methods utilizing linear frequency cepstral coefficients (LFCCs) and spectro-temporal modulation representations (STM). Evaluated on benchmark datasets from ASVspoof 2019 and ADD2023, our methods achieved impressive results with equal error rates (EER) of 8.33% and 42.10%, respectively. Notably, STM demonstrated strong performance in differentiating real and deepfake speech across both datasets. These findings were presented at APSIPA 2023 [8] and iSAI-NLP 2023 conferences [9].

In parallel, we explored human speech perception through an auditory periphery model. This research, published in the Journal of Applied Acoustics [10], lays the groundwork for future studies on detecting speech and speaker identity changes caused by speech synthesis. Recognizing the critical need for multilingual resources, we initiated the development of a Thai language spoof database (ThaiSpoof) [11]. This database encompasses genuine and various types of spoofed speech signals generated using text-to-speech tools, synthesizers, and modification tools. To showcase the potential of ThaiSpoof, we implemented a simple convolutional neural network (CNN) model taking LFCCs as input. Trained and evaluated on ThaiSpoof, the model achieved an accuracy of 95% and an EER of 4.67%. These results demonstrate the value of ThaiSpoof as a resource for advancing spoof detection research.

While our research has shown promising results with a strong focus on spoof detection methods, there are two key areas requiring further exploration for real-world application. First, integrating these methods with human-computer interaction (HCI) systems is crucial. This will involve investigating how to seamlessly incorporate spoof detection into user interfaces and ensure a smooth user experience while maintaining security. Second, we need to broaden the scope of voice privacy protection beyond gender anonymization. Future research should explore methods to protect other speaker-related features, such as age, ethnicity, or emotional state. This will require a deeper understanding of how these features are encoded in speech signals and how to anonymize them effectively without compromising intelligibility. By addressing these limitations, we can pave the way for robust and user-centric spoof detection solutions that safeguard user privacy in various HCI applications.

＜引用文献＞

[1] Tomashenko, et al. (2020). Introducing the VoicePrivacy Initiative. *Interspeech 2020*, pp.1693–1697.

[2] Fujisaki (2004). Information, prosody, and modeling - with emphasis on tonal features of speech -. *Speech Prosody 2004*, pp. 1–10.

[3] Poggi and D'Errico. (2012). Social signals: a framework in terms of goals and beliefs. *Cognitive processing, 13 Suppl 2*, pp. 427–445.

[4] Plack, C.J. (2018). The Sense of Hearing (3rd ed.). *Routledge*.

[5] Schuller, et al. (2013). Paralinguistic in speech and language – State-of-the-art and the challenge. *Computer Speech & Language, 27*, pp. 4–39.

[6] Mawalim, C.O., et al. (2022). Speaker Anonymization by Pitch Shifting Based on Time-Scale Modification. *ISCA SPSC 2022*, pp. 35–42.

[7] Mawalim, C.O., et al. (2022). F0 Modification via PV-TSM Algorithm for Speaker Anonymization Across Gender. *APSIPA 2022*, pp. 196–203.

[8] Cheng, H, et al. (2023). Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection. *APSIPA 2023*, pp. 1822–1829.

[9] Mon, K.Z., et al. (2023). Voice Contribution on LFCC feature and ResNet-34 for Spoof Detection. *iSAI-NLP 2023*.

[10] Mawalim, C. O., et al. (2023). Non-intrusive speech intelligibility prediction using an auditory periphery model with hearing loss. *Applied Acoustics, 214*, pp. 1–11.

[11] Galajit, K., et al. (2023). ThaiSpoof: A Database for Spoof Detection in Thai Language. *iSAI-NLP 2023*.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　3件／うちオープンアクセス　3件）

１．著者名 Mawalim Candy Olivia、Okada Shogo、Nakano Yukiko I.、Unoki Masashi	４．巻 -
２．論文標題 Personality trait estimation in group discussions using multimodal analysis and speaker embedding	５．発行年 2023年
３．雑誌名 Journal on Multimodal User Interfaces	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） 10.1007/s12193-023-00401-0	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 Hayashi Takato、Mawalim Candy Olivia、Ishii Ryo、Morikawa Akira、Fukayama Atsushi、Nakamura Takao、Okada Shogo	４．巻 11
２．論文標題 A Ranking Model for Evaluation of Conversation Partners Based on Rapport Levels	５．発行年 2023年
３．雑誌名 IEEE Access	６．最初と最後の頁 73024～73035
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2023.3287984	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 Mawalim Candy Olivia、Titalim Benita Angela、Okada Shogo、Unoki Masashi	４．巻 214
２．論文標題 Non-intrusive speech intelligibility prediction using an auditory periphery model with hearing loss	５．発行年 2023年
３．雑誌名 Applied Acoustics	６．最初と最後の頁 109663～109663
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.apacoust.2023.109663	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13件（うち招待講演　0件／うち国際学会　13件）

１．発表者名 Ohba Tomoya、Mawalim Candy Olivia、Katada Shun、Kuroki Haruki、Okada Shogo
２．発表標題 Multimodal Analysis for Communication Skill and Self-Efficacy Level Estimation in Job Interview Scenario
３．学会等名 The 21st International Conference on Mobile and Ubiquitous Multimedia (MUM 2022)（国際学会）
４．発表年 2022年

１．発表者名
Mawalim Candy Olivia、Okada Shogo、Unoki Masashi

２．発表標題
F0 Modification via PV-TSM Algorithm for Speaker Anonymization Across Gender

３．学会等名
2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference（国際学会）

４．発表年
2022年

１．発表者名
Mawalim Candy Olivia、Okada Shogo、Unoki Masashi

２．発表標題
Speaker Anonymization by Pitch Shifting Based on Time-Scale Modification

３．学会等名
The 2nd SPSC joined with 2nd VoicePrivacy Challenge Workshop, as a satellite to Interspeech 2022（国際学会）

４．発表年
2022年

１．発表者名
Titalim Benita Angela、Mawalim Candy Olivia、Okada Shogo、Unoki Masashi

２．発表標題
Speech Intelligibility Prediction for Hearing Aids Using an Auditory Model and Acoustic Parameters

３．学会等名
2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference（国際学会）

４．発表年
2022年

１．発表者名
Mawalim Candy Olivia、Titalim Benita Angela、Okada Shogo、Unoki Masashi

２．発表標題
OBISHI: Objective Binaural Intelligibility Score for the Hearing Impaired

３．学会等名
The 18th Australasian International Conference on Speech Science and Technology（国際学会）

４．発表年
2022年

１．発表者名
Candy Olivia Mawalim, Benita Angela Titalim, Shogo Okada, and Masashi Unoki

２．発表標題
Auditory Model Optimization with Wavegram-CNN and Acoustic Parameter Models for Nonintrusive Speech Intelligibility Prediction in Hearing Aids

３．学会等名
The 31st European Signal Processing Conference (EUSIPCO 2023), Helsinki, Finland（国際学会）

４．発表年
2023年

１．発表者名
Hung Le, Sixia Li, Candy Olivia Mawalim, Hung-Hsuan Huang, Chee Wee Leong, and Shogo Okada

２．発表標題
Investigating the Effect of Linguistic Features on Personality and Job Performance Predictions

３．学会等名
The 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, July 23-28, 2023（国際学会）

４．発表年
2023年

１．発表者名
Xiguang Li, Shogo Okada, and Candy Olivia Mawalim

２．発表標題
Inter-person Intra-modality Attention Based Model for Dyadic Interaction Engagement Prediction

３．学会等名
The 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, July 23-28, 2023（国際学会）

４．発表年
2023年

１．発表者名
Xiajie Zhou, Candy Olivia Mawalim, and Masashi Unoki

２．発表標題
Incorporating the Digit Triplet Test in A Lightweight Speech Intelligibility Prediction for Hearing Aids

３．学会等名
The 15th Asia-Pasific Signal and Information Processing Association (APSIPA ASC 2023), Taipei, Taiwan, 31 October - 3 November 2023（国際学会）

４．発表年
2023年

１．発表者名
Haowei Cheng, Candy Olivia Mawalim, Kai Li, Lijun Wang, and Masashi Unoki

２．発表標題
Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection

３．学会等名
The 15th Asia-Pasific Signal and Information Processing Association (APSIPA ASC 2023), Taipei, Taiwan, 31 October - 3 November 2023 (国際学会)

４．発表年
2023年

１．発表者名
Khaing Zar Mon, Kasorn Galajit, Candy Olivia Mawalim, Jessada Karnjana, Tsuyoshi Isshiki, and Pakinee Aimmanee

２．発表標題
Voice Contribution on LFCC feature and ResNet-34 for Spoof Detection

３．学会等名
The 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing and The International Conference on Artificial Intelligence and Internet of Things (iSAI-NLP 2023) (国際学会)

４．発表年
2023年

１．発表者名
Kasorn Galajit, Thunpisit Kosolsriwiwat, Candy Olivia Mawalim, Pakinee Aimmanee, Waree Kongprawechnon, Win Pa Pa, Anuwat Chaiwongyen, Teeradaj Racharak, Hayati Yassin, Jessada Karnjana, Surasak Boonkla, and Masashi Unoki

２．発表標題
ThaiSpoof: A Database for Spoof Detection in Thai Language

３．学会等名
The 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing and The International Conference on Artificial Intelligence and Internet of Things (iSAI-NLP 2023) (国際学会)

４．発表年
2023年

１．発表者名
Aulia Adila, Candy Olivia Mawalim, Isoyama Takuto, and Masashi Unoki

２．発表標題
Study on Inaudible Speech Watermarking Method Based on Spread-Spectrum Using Linear Prediction Residue

３．学会等名
The 2024 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, Hawaii, 27 February - 1 March 2024 (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