SpeeDF - A Speech De-identification Framework

Chandra Sekar Veerappan
Singapore Institute of Technology
chandra.veerappan@singaporetech.edu.sg

Priyanshu Dhingra
Rajiv Gandhi Institute of Petroleum Technology
priyanshudhingra1@gmail.com

Daniel Zhengkui Wang
Singapore Institute of Technology
zhengkui.wang@singaporetech.edu.sg

Rong Tong
Singapore Institute of Technology
tong.rong@singaporetech.edu.sg

***Abstract*—This paper proposes SpeeDF, a novel three-step framework for anonymizing speech data, particularly focusing on Singaporean English (Singlish). SpeeDF tackles the challenge of protecting less-studied Personally Identifiable Information (PII) like NRIC and passport numbers, which often go overlooked by traditional de-identification methods. Unlike approaches focused solely on entity extraction, SpeeDF leverages a combination of automatic speech recognition (ASR), named entity recognition (NER), and information anonymization. This comprehensive approach ensures thorough PII redaction while preserving the naturalness and usability of the anonymized speech data for research and various downstream applications.**

***Index Terms*—automatic speech recognition, speech de-identification, named entity recognition, deep learning, framework**

I. Introduction

The explosion of digital applications has transformed how we live. These applications, from chatbots to voice assistants, collect and share a vast amount of personal data (PII) across digital networks. This raises concerns about potential misuse, leading to privacy breaches, identity theft, and other harm. The data can even be used to train machine learning models, further amplifying privacy risks [3].

Speech de-identification is the process of anonymizing spoken language data by removing or modifying personally identifiable information while preserving the overall content and usability of the speech. It has become a hot topic in spoken language understanding research [1]. It allows us to protect sensitive information, like IDs, from spoken conversations. This technology has valuable applications, such as anonymizing sensitive information in medical recordings [4].

In this paper, we propose a three-step speech de-identification framework for anonymizing Singapore English speech. Speech de-identification faces the following challenges:

Variations in Natural Speech: Speech data, unlike formal text, is inherently informal and can include slang, contractions, and incomplete sentences. Additionally, spoken language often carries accents, especially in multicultural societies like Singapore. These natural variations in speech can pose challenges for speech recognition and entity recognition systems.

Data Scarcity and privacy concerns: Training effective de-identification models requires a large amount of high-quality speech data annotated with PII. However, the very nature of PII necessitates careful handling of speech data. Obtaining large, diverse datasets can be difficult due to ethical considerations and privacy regulations.

Privacy-Utility Tradeoff: While anonymization protects individuals' privacy, it can also limit the usefulness of the data for research and development purposes. Finding the right balance is crucial to ensure both privacy protection and the advancement of speech-related technologies. This paper proposes a novel speech de-identification framework, SpeeDF, that tackles the challenges of anonymizing spoken data in Singaporean English (Singlish).

SpeeDF addresses these challenges through a three-pronged approach:

1) Customized Automatic Speech Recognition: SpeeDF tailors ASR to recognize local accents and Singlish language patterns for improved accuracy.
2) Enhanced PII Detection with Data Augmentation: The framework utilizes data augmentation techniques to create synthetic speech samples rich in diverse PII examples, enhancing its ability to detect sensitive information.
3) Multi-Level PII Anonymization: SpeeDF employs different anonymization techniques based on the sensitivity level of the detected PII, ensuring a balance between privacy protection and data usability [2].

II. Related works

An early speech de-identification system is explored in [14], this paper investigated the task of removing sensitive information from audio data in various domains, including medical settings. The system is a pipeline system: named entity recognition is performed after automatic speech recognition to convert the speech into text. End to end system is also explored, in this type of systems, the speech recognition and named entity recognition are joint trained. For example, Chen et al. [15] presented thier work on combining entity aware ASR with NER taggers for Chinese dataset.

Accented speech recognition has been a well-documented challenge due to factors like unique vocabulary, pronunciation variations, and code-switching. Leveraging text-to-speech (TTS) for generating rare words from text data, as proposed by [6], holds promise for improving ASR performance in

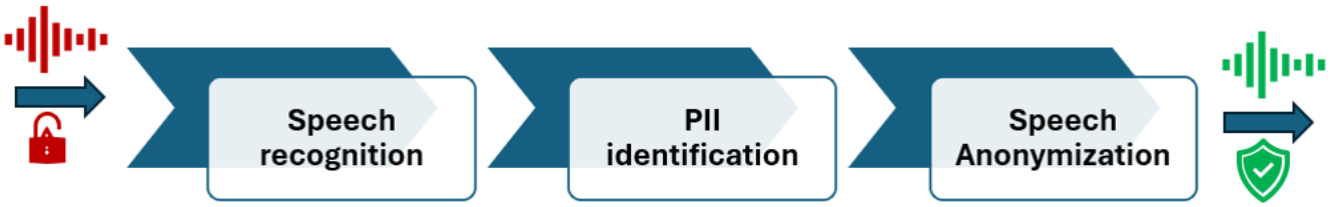

Fig. 1. Three step framework for speech de-identification

Singlish. Additionally, [5] worked on accent-robust ASR for code-switching Mandarin-English speech recognition.

To tackle the data sparsity problem, a data augmentation approach is proposed by Hemateja et al. [8] for named entity recognition, to improve the accuracy of named entity recognition. This approach can act as a sanity check on the correctness of synthesized speech samples used for training. Additionally, incorporating external data, as explored by [9], holds promise for further enhancing speech NER performance. Their findings suggest that external data can be particularly beneficial for end-to-end models.

To protect PII from clinical notes [10] utilized dictionaries to automatically annotate a training dataset for a named entity recognition model to de-identify narrative clinical text. Textwash [11] is an open source tool for text anonymisation, it replace the detected sensitive information with category-specific and meaning-preserving tokens. For example, it may use 'firstname1' to replace a person's first name Joe . Various data anonymization methods are compared in [12] to evaluate the effectiveness of data anonymization. Those methods are classified into two groups: randomization based techniques, such as noise addition, permutation and differential privacy; generalization based techniques, include aggregation, k-anonymity, l-diversity and t-proximity.

III. Proposed Method

Figure 1 illustrates the proposed SpeeDF framework for speech de-identification. The framework operates in three steps:

- Speech Recognition: The input speech signal is processed by an automatic speech recognition model, converting it into text.
- PII Identification: The text transcription is then analyzed by a PII identification module to detect and locate any sensitive personally identifiable information within the text.
- Speech Anonymization: The identified PII undergoes an anonymization process within a dedicated module. This anonymization protects the sensitive information while preserving the overall content and usability of the speech data. Finally, the anonymized speech is produced as the output.

A. Automatic speech recognition

The first step in speech de-identification involves extracting the speech content. We achieve this by leveraging an automatic speech recognition (ASR) system to convert the spoken audio into text. However, accurately transcribing accented speech can be challenging.

This work leverages Whisper, an open-source speech recognition model[1], to convert spoken language into text. To better capture the unique characteristics of Singaporean English (Singlish), we fine-tuned the Whisper model on a dataset of conversational samples from the Singapore English National Speech Corpus (NSC) [16].

B. PII identification

To addresses the challenge of limited real-world speech data for de-identification tasks, specifically for Singaporean English. We propose a data augmentation approach that leverages large language models (LLMs) as described in our recent work [7]. Our approach begins with a handcrafted seed dataset containing sample PII information. We then utilize the LLM to learn the stylistic patterns of the seed data. This allows the LLM to generate new, realistic speech data rich in diverse PII examples. We have released our seed and augmented datasets to the research community[2].

While F1 score is commonly used for entity recognition tasks, it can be insufficient for analyzing specific errors encountered during entity detection in spontaneous speech data. To provide a more granular performance analysis, we propose alternative metrics consist of: the percentage of perfect entity match (Perfect), the percentage of partial match (Partial), the percentage of substitution (Substitution), the percentage of insertion (Hallucination) and the percentage of deletion (Omission). These new metrics offer a more detailed understanding of the error types encountered compared to the traditional F1 score approach.

C. Speech Anonymization

In Speech anonymization step, given the identified PIIs within the speech data, along with their location, we propose two methods to redact the PII:

- Simple Replacement: This method directly replaces PII with white noise or a beep sound. It's easy to implement

[1]https://github.com/openai/whisper/blob/main/model-card.md
[2]https://github.com/e102sg/speedf

and works for all PII types. However, with frequent PII occurrences, the audio can become unusable and unpleasant.

- Category-Preserving Replacement: This approach prioritizes preserving the original audio quality and speaker characteristics. PII is first replaced with fictional data of the same category (e.g., a fake phone number for a real one).

Following PII anonymization in the text domain, the anonymized text can optionally be converted back into synthetic speech using a text-to-speech or voice conversion system. This approach offers a two-layer privacy safeguard: protecting the content PII and the speaker's acoustic characteristics. This ensures a natural listening experience while maintaining privacy.

IV. EXPERIMENTS AND ANALYSIS

A. Automatic speech recognition

We extracted speech samples from Part 6 of the National Corpus of Singapore (NSC) corpus. This part consists of conversational sessions across six common scenarios: hotel booking, travel itinerary planning, restaurant reservation, banking services, insurance services, and telecommunication services. We randomly selected 500 conversations for fine tuning Whisper medium model. The adaptation is performed using a low-rank adaptation (LoRA) approach [17].

To evaluate the performance of fine-tuned ASR, we have invited four speakers (3 male 1 female) to record the 150 sentences from our NER test set. All the speakers are not present in the NSC corpus.

Table I presents the performance of the fine-tuned model on Singlish speech recognition. The word error rate (WER) reduced from 17.7% to 13.56%.

TABLE I
SPEECH RECOGNITION PERFORMANCE ON TEST SET

ASR model	WER%
Whisper-medium.en	17.7
Fine-tuned on NSC data	13.56

Here's an example sentence before and after adaptation. The baseline ASR model incorrectly recognized the word "singel" as a company name, likely due to its unfamiliarity with Singaporean English. However, after fine-tuning the model on Singaporean English data, it accurately identified the word.

Reference can i *have* your email please sure customer dot service at *singtel dot* com

Hypothesis(Whisper.medium.en): can i *help* your email please sure customer service *centel* com

Hypothesis (fine-tuned Whisper.medium.en): can i *have* your email please sure customer dot service at *singtel dot* com

B. PII identification

Leveraging the text transcription from the ASR system, we transform the speech PII identification task into a text-based named entity recognition (NER) problem. Our NER model targets both common entities like names and Singapore-specific entities like NRIC, passport numbers, phone numbers, and various financial details.

Addressing the scarcity of highly sensitive entities in public datasets, we leverage a large language model to generate synthetic data tailored to training our named entity recognition model. For both NER model training and identification of Personally Identifiable Information entities, we utilize spaCy[3]. The PII detection model is trained on both the initial seed data and the LLM-generated data.

Table II compares the overall PII detection performance between a baseline model and our data-augmented model. We present the conventional F1 score alongside five additional metrics we propose. The results demonstrate that the augmented model surpasses the baseline model in achieving a higher overall perfect match rate, a lower hallucination rate, and a lower omission rate. While the baseline model exhibits higher partial match and lower substitution rates, these scores might be inflated by false positives stemming from hallucinations.

TABLE II
PII DETECTION RESULTS

Overall metrics	Base model	Data augmented model
F1 ↑	0.67	**0.8**
Perfect(%) ↑	66.77	**74.34**
Partial(%) ↑	9.59	0.95
Substitution(%) ↓	8.07	14.23
Hallucination(%) ↓	73.19	**26.77**
Omission(%) ↓	15.56	**10.48**

C. Speech Anonymization

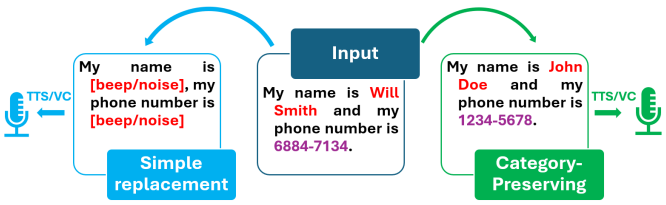

Fig. 2. Example of speech anonymization using proposed methods

Figure 2 illustrates two speech anonymization approaches within SpeeDF. The example demonstrates the identification of two PIIs: a person's name and a phone number.

Simple Replacement Method: This method directly modifies the original audio. It can either: Remove the identified PIIs entirely, or replace them with noise (e.g., static or beeps) to mask the information.

Category-Preserving Anonymization: This approach replaces the PIIs with anonymized information that maintains the

[3]https://spacy.io/api/entityrecognizer/

same category: The person's name is replaced with a generic name like "John Doe." The phone number is replaced with a different phone number sequence (e.g., "1234-5678").

The anonymized text can then be fed into a text-to-speech or voice cloning system to generate new audio with a different voice, offering an additional layer of privacy protection.

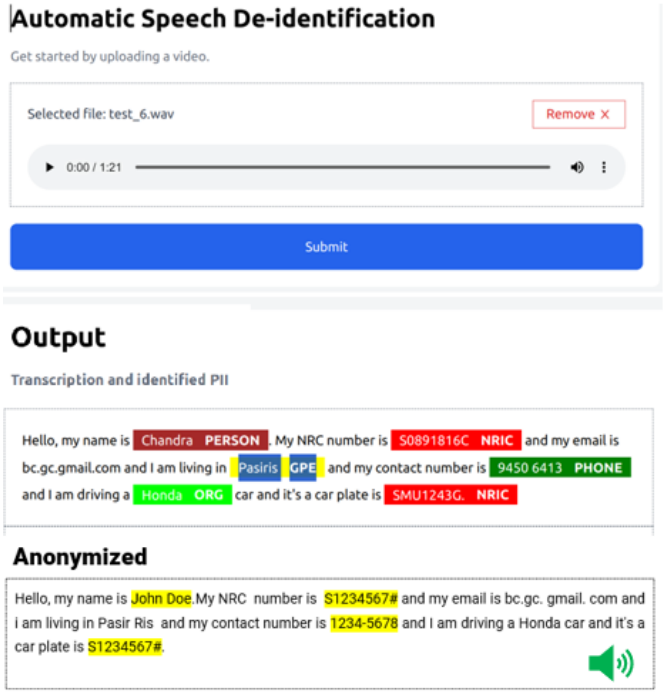

Fig. 3. SpeeDF prototype system

D. system prototype

Figure 3 showcases a screenshot of the SpeeDF system's user interface. Users can upload an audio file for processing. SpeeDF then displays the following information:

- Transcription: The transcription of the spoken content derived from automatic speech recognition.
- Identified PII: Sensitive information (PII) highlighted with color coding for easy identification.
- Anonymized Speech content: The original audio with PII replaced by anonymized information that maintains the same category

Additionally, users can download the processed audio file.

V. Conclusion and Future Works

This paper introduces SpeeDF, a novel three-step framework for safeguarding privacy in speech data, with a specific focus on Singaporean English (Singlish). By integrating automatic speech recognition, named entity recognition, and information anonymization, SpeeDF effectively protects sensitive Personally Identifiable Information while preserving speech naturalness. SpeeDF's unique strength lies in its dual-layer privacy protection, safeguarding both speech content and speaker identity.

Future work will explore end-to-end speech PII identification, aiming to optimize speech recognition and entity extraction simultaneously. Such an approach requires a larger dataset of speech data rich in PII information. To realize this, we will address the challenge of data scarcity through innovative data augmentation techniques and speech synthesis methods.

Acknowledgment

This work is supported by Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant No. R-R12A405-0009 and Ignition grant No. R-IE2-A405-00006.

References

[1] Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin, "Where are we in named entity recognition from speech?" in Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 4514–4520.

[2] Ghosheh, Ghadeer O., Jin Li, and Tingting Zhu. "A survey of generative adversarial networks for synthesizing structured electronic health records." ACM Computing Surveys 56.6 (2024): 1-34

[3] Lauren Leffer, "Your Personal Information Is Probably Being Used to Train Generative AI Models", October 19, 2023 (online article) https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/

[4] Zengjian Liu et al. "De-identification of clinical notes via recurrent neural network and conditional random field". In: Jour nal of biomedical informatics 75 (2017),S34–S42

[5] Yang, Yuhang, et al. "Adapting OpenAI's Whisper for Speech Recognition on Code-Switch Mandarin-English SEAME and ASRU2019 Datasets." arXiv preprint arXiv:2311.17382 (2023).

[6] Yuen, Kwok Chin, Li Haoyang, and Chng Eng Siong. "ASR Model Adaptation for Rare Words Using Synthetic Data Generated by Multiple Text-To-Speech Systems." APSIPA ASC. IEEE, 2023.

[7] Priyanshu Dhingra, Satyam Agrawal, Chandra Sekar Veerappan, Ho Thi Nga, Eng Siong Chng, Rong Tong, "Speech de-identification data augmentation leveraging large language model", International Conference on Asian Language Processing (IALP) 2024

[8] Hemateja, Aluru VNM, et al. "Novel data augmentation for named entity recognition." International Journal of Speech Technology 26.4 (2023): 869-878.

[9] Pasad, Ankita, et al. "On the Use of External Data for Spoken Named Entity Recognition." Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022.

[10] Laursen, Martin Sundahl, et al. "Automatic Annotation of Training Data for Deep Learning Based De-identification of Narrative Clinical Text." The First Workshop on Context-aware NLP in eHealth:(WNLPe-Health 2022). CEUR Workshop Proceedings, 2023.

[11] Kleinberg, Bennett, Toby Davies, and Maximilian Mozes. "Textwash–automated open-source text anonymisation." arXiv preprint arXiv:2208.13081 (2022).

[12] Ni, Chunchun, et al. "Data anonymization evaluation for big data and IoT environment." Information Sciences 605 (2022): 381-392.

[13] Flechl, Martin, et al. "End-to-end speech recognition modeling from de-identified data." INTERSPEECH 2022

[14] Ido Cohn et al. "Audio De-identification - a New Entity Recognition Task". In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). June 2019, pp. 197–204. DOI: 10.18653/v1/N19-2025.

[15] Boli Chen et al. "Aishell-ner: Named entity recognition from chinese speech". In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2022, pp. 8352–8356.

[16] Jia Xin Koh et al. "Building the Singapore English national speech corpus". In: Malay 20.25.0 (2019), pp. 19–3.

[17] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W., "Lora: Low-rank adaptation of large language models," in arXiv preprint arXiv:2106.09685, 2021.