Enhancing Speech De-identification with LLM-Based Data Augmentation

Priyanshu Dhingra*, Satyam Agrawal†, Chandra Sekar Veerappan‡, Eng Siong Chng§, Rong Tong‡
* Rajiv Gandhi Institute of Petroleum Technology, priyanshudhingra1@gmail.com
† National Institute of Technology Karnataka, lmcsatyam@gmail.com
‡ Singapore Institute of Technology, {chandra.veerappan, tong.rong}@singaporetech.edu.sg
§ Nanyang Technological University, aseschng@ntu.edu.sg

***Abstract*—This paper addresses the challenge of data scarcity in speech de-identification by introducing a novel, fully automated data augmentation method leveraging large language models. Our approach overcomes the limitations of human annotation, enabling the creation of extensive training datasets. To enhance de-identification performance, we compare pipeline and end-to-end models. While the pipeline approach sequentially applies speech recognition and named entity recognition, the end-to-end model jointly learns these tasks. Experimental results demonstrate the effectiveness of our data augmentation strategy and the superiority of the end-to-end model in improving PII detection accuracy and robustness.**

***Index Terms*—Data augmentation, speech recognition, named entity recognition, de-identification**

I. Introduction

The increasing use of voice-operated systems in healthcare, finance, and other sensitive sectors has amplified privacy worries, as these systems often handle highly confidential information. Preserving the confidentiality of personally identifiable details embedded in speech data is crucial. Speech de-identification, which involves identifying and masking personally identifiable elements from speech recordings, is a critical countermeasure. However, creating systems to remove private details from speech is difficult because there aren't enough publicly available speech recordings with marked personal information. This lack of data makes it a big problem and requires new methods to solve it while still protecting people's privacy.

This paper investigates speech de-identification specifically tailored for Singapore English. We propose a dual-pronged approach encompassing data augmentation and end-to-end modeling to address the challenges posed by limited annotated data.

To overcome the scarcity of training data, we employ data augmentation techniques. Starting with a relatively small corpus of Singapore English text data rich in personally identifiable information (PII), we generate additional training examples through various augmentation methods. These augmented text sequences are subsequently transformed into speech using voice conversion techniques, effectively expanding the dataset.

In our prior work [1], we investigated data augmentation using large language models to generate semantically and structurally similar text from seed data. However, this approach requires post PII annotation, limiting its scalability for large data augmentation. To eliminate the need for data annotation, we introduce two new fully automated data augmentation strategies: similar and diverse augmentation.

Similar augmentation involves replacing named entities within sentences with alternative terms belonging to the same entity category. For instance, substituting a specific person's name with another name while preserving the person entity type.

Diverse augmentation leverages a large language model to generate synthetic sentences containing a predefined set of named entity tags. This approach introduces a wider range of linguistic variations and contextual diversity into the augmented dataset.

We subsequently compare the performance of pipeline and end-to-end approaches for speech de-identification. Traditional approaches to speech de-identification typically involve a pipeline architecture, consisting of separate automatic speech recognition (ASR) and named entity recognition (NER) modules. This sequential process can amplify errors from preceding stages, impacting overall performance. The end-to-end model is trained concurrently on speech recognition and PII tagging, aiming to capture the dependencies between speech features and PII for enhanced accuracy and robustness.

II. Related Work

De-identification is the process of removing or masking personally identifiable information from a dataset. De-identification is crucial for safeguarding individual privacy while maximizing the utility of data. By removing or masking personally identifiable information from data, individuals are protected from identity theft and privacy breaches, the de-identified data can be safely shared for research, public health initiatives, and other valuable purposes without compromising privacy.

Text de-identification is widely used in the medical domain. A de-identification system has been developed to protect Personally Identifiable Information within free-text medical records such as nurse's notes and hospital discharge reports [2]. De-identification imposes some limitations on NLP models, but the overall impact on performance is modest [3].

Although text de-identification methods are well-developed, extending these approaches to video and audio data presents substantial challenges. De-identification of multimodal information presents a complex challenge. Visual data, for instance,

may contain visual cues that reveal sensitive information, such as facial recognition [4] [5] or identifiable locations Similarly, Speech data can contain explicit personally identifiable information or unique vocal characteristics that make de-identification challenging. In speech processing, many efforts are working on the de-identification of speaker's biometric information [6] [7], there are less works on de-identification of non-biometric information in speech content. The automatic de-identification on the speech content is under explored. This paper addresses the critical task of protecting sensitive, non-biometric information embedded in speech.

Traditional speech de-identification systems often relied on a pipeline architecture. An automatic speech recognition model was initially used to transcribe speech into text, followed by named entity recognition to identify and potentially replace personally identifiable information.

Cohn et al. [8] introduced the audio de-identification task, emphasizing the importance of removing sensitive information from audio datasets across various domains. Their approach involved a two-step pipeline: first, ASR was employed to convert speech to text, followed by NER to identify entities. Finally, an alignment process was applied. To overcome the negative effects of out of vocabulary (OOV) problem, [9] proposed to augment NER by including OOV indictive terms in speech recognition system. In this way, any OOV words are treated as in an OOV region, the NER detection is relied on the context of the OOV region. This approach shows some improvements in the NER detection. Szymanski et al. [10] conducted a critical analysis of ASR-NER systems, acknowledging the limitations of both ASR and NER components.

In pipeline approaches, errors introduced by the ASR component can propagate to the NER model, leading to decreased performance. To address the limitation of the pipeline de-identification approach, end to end method is proposed to train the ASR and NER together. In early end to end approach, the model training process is still in a sequential fashion: a deep neural network is trained on the ASR task first, the last layer of the neural network is then reinitialized and trained on NER task [11] [12].

Chen et al. [13] introduced a Chinese speech NER dataset, serving as a critical resource for the field. By incorporating NER tags as special tokens within ASR transcriptions, they demonstrated a significant improvement in speech NER performance through the integration of entity-aware ASR and pre-trained NER taggers. This work underscores the potential of combining these components for enhanced entity extraction from speech.

A de-identification systems often face the challenge of data scarcity, particularly regarding speech data containing personally identifiable information. To address this, data augmentation has emerged as a common strategy. Recent studies have explored various data augmentation techniques. For instance, a data augmentation approach with a sanity-checker was proposed to improve transformer-based NER models [14]. Furthermore, the Graph Propagated Data Augmentation (GPDA) framework [15] leverages graph propagation to establish relationships between labeled and unlabeled text for named entity recognition.

III. Proposed Solution

Figure 1 presents a flowchart illustrating the proposed system. We initiate the process with a small, carefully curated seed text dataset that accurately represents Singaporean oral English, incorporating a diverse range of PII entities. The seed data is expanded using the Gretel tool [16] to generate synthetic data based on semantic and structural similarities. Given the requirement for post PII annotations in this text expansion process, we introduce two fully automated data augmentation methods targeting distinct aspects: similar and diverse. Subsequently, speech data is derived from the augmented text data via voice conversion (VC). The resulting speech and text pairs serve as input for PII modeling.

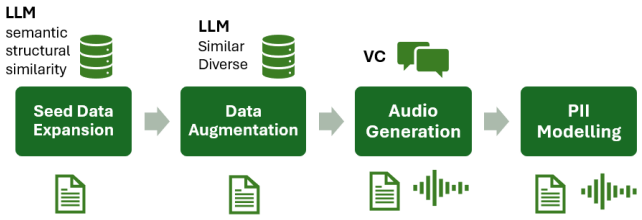

Fig. 1. System flowchart

A. Text Data Augmentation

In our previous work [1], we described the creation of a seed dataset by revising sentences randomly selected from the NSC corpus [17]. The seed data is expanded by using Gretel tool. However the generated text lacked annotations. To address this, we adopted a semi-automatic approach. This involved training a custom NER model on a combination of a base spaCy model[1] and our seed data. While this approach achieved some success, the need for human verification and correction of annotations created a bottleneck, limiting the volume of data that could be produced.

This paper introduces a data augmentation method that bypasses the need for human expert review, thereby facilitating the generation of a substantially larger augmented dataset. The proposed method comprises the following steps:

1) **Example Data Extraction**: Sentences containing target NER annotations are selected from the seed dataset.
2) **Language Model Prompting**: For each chosen sentence, NER tags, and optional prompts or structural information are provided to the Large Language Model.
3) **Sentence Generation**: The LLM generates new sentences annotated with the specified NER tags.
4) **Tag Replacement**: Generated sentences undergo a tag replacement process, substituting original NER tags with alternative ones while preserving label consistency.

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

During the large language model prompting, we propose two approaches that focus on distinct augmentation criteria for generating augmented data.

1) **Similar-based Augmentation (Similar):** This approach involves providing the language model with an NER tag, along with example sentences where the tag appears. The LLM is instructed to replicate the sentence structure of the provided examples. This method aims to produce sentences that closely resemble the original examples, ensuring consistency in style and structure. Figure 2 shows the prompt used for similar data generation. The prompts were crafted to guide the model in generating new sentences that closely adhere to the structure and style of these examples, ensuring that the augmented sentences retain a high degree of similarity to the original data.

```
You are given a set of named entity recognition (NER)
tags. Your task is to generate a set of 20 new
sentences that contain exactly the same NER tags as
given to you. Each new sentence should:
 1.Use exactly the same NER tags in appropriate
   contexts without adding any new NER tags.
 2.Reflect spoken language, as if someone is speaking
   to another person.
 3.Follow the structure as given by
   {sentence_examples}.
The format for the input will be such that you will
get the NER tags in close brackets. Given this kind of
input, you need to now generate 20 more sentence that
contain exactly the same NER tags as given in the
prompt. DO NOT ADD BULLET NUMBERS IN THE OUTPUT.
```

Fig. 2. Data augmentation prompt for Similar

2) **Diverse-based Augmentation (Diverse):** In this method, the LLM is given the NER tag and encouraged to generate sentences with greater creative freedom. Figure 3 shows the prompt used for diverse data generation. By prompting the model to create diverse examples, we aimed to enrich our dataset with a broad range of variations, which helps in capturing a wider array of linguistic patterns and contexts.

```
You are given a set of named entity recognition (NER)
tags. Your task is to generate a set of 20 new
sentences that contain exactly the same NER tags as
given to you. Each new sentence should:
 1.Use exactly the same NER tags in appropriate
   contexts without adding any new NER tags.
 2.Reflect spoken language, as if someone is speaking
   to another person.
 3.Be as creative and diverse as you can with the
   structure of sentences you come up with. Talk about
   the NER tag in different contexts and themes.
The format for the input will be such that you will
get the NER tags in close brackets. Given this kind of
input, you need to now generate 20 more sentence that
contain exactly the same NER tags as given in the
prompt. DO NOT ADD BULLET NUMBERS IN THE OUTPUT.
```

Fig. 3. Data augmentation prompt for Diverse

The proposed Diverse and Similar augmentation methods benefit from automatic annotation. Since the language model incorporates pre-existing NER tags and labels from the seed sentence into its response, tags in the generated sentences can be easily replaced with semantically equivalent tags. In contrast, the Gretel-generated data initially lacked annotations, requiring a labor-intensive semi-automatic refinement process involving human experts. This manual intervention created a bottleneck, limiting data volume. Our LLM-based approach eliminates this bottleneck, enabling efficient and scalable generation of annotated data.

B. Audio Generation

To generate the required speech data for our training process, we leverage WhisperSpeech[2], a state-of-the-art open-source Text-to-Speech (TTS) system. By inverting the Whisper speech recognition model, WhisperSpeech produces high-quality speech output from textual input. Beyond standard text-to-speech capabilities, WhisperSpeech excels at voice cloning, accurately mimicking a specific speaker's voice based on a provided audio sample. By leveraging this tool, we can generate high-quality speech from augmented text data and sample Singapore English audios, creating text and audio paired dataset for speech de-identification modelling.

C. PII Modelling

In this work, we explore two speech de-identification modelling methods: pipeline approach and end to end approach.

1) Pipeline approach: Our pipeline approach consists of two main components: Automatic Speech Recognition and Named Entity Recognition. This method processes speech data sequentially, first converting speech to text and then identifying PII within the transcribed text using named entity recognition model.

- Automatic Speech Recognition
  We utilize the Whisper model[3], an open-source speech recognition system, as our base automatic speech recognition component. To better capture the unique characteristics of Singaporean English (Singlish), we fine-tune the Whisper model on a dataset of conversational samples from the Singapore English National Speech Corpus (NSC) [17]. The adaptation is performed using a low-rank adaptation (LoRA) approach, which allows for efficient fine-tuning while preserving the model's general knowledge. This fine-tuning process improves the model's ability to accurately transcribe Singlish speech patterns and vocabulary.
- Named Entity Recognition
  For the NER component, we employ a custom spaCy model specifically tailored to identify Personally Identifiable Information (PII) in speech transcripts. This augmented dataset is used to train the NER model, improving its ability to recognize a wide range of PII entities in transcribed speech.
- Pipeline Integration

[2]https://github.com/collabora/whisperspeech
[3]https://github.com/openai/whisper

In the operational pipeline, the ASR module first processes the input speech, generating a text transcription. This transcription is then passed to the NER module, which identifies and tags the PII entities within the text. The pipeline approach allows for modular development and optimization of each component separately. However, it's important to note that this approach may be subject to error propagation between modules and may not fully leverage the acoustic features present in the original speech signal for entity recognition.

2) End-to-End approach: Our end-to-end model training approach involves fine-tuning the Whisper speech recognition model to perform both speech recognition and named entity recognition simultaneously. The process consists of the following key steps:

- Data Preparation:
  The training data consists of audio files and their corresponding transcriptions, where the transcriptions are annotated with entity tags. For example: A transcription might be annotated like this:

  [PERSON] John Doe [PERSON] went to [GPE] New York [GPE] on [DATE] January 1st [DATE]

- Tokenization and Special Tokens:
  The text is broken down into smaller units (tokens). Each word and entity tag (like [PERSON] and [GPE]) is treated as a separate token. As the entity tags are not standard English vocabulary, they are newly introduced for entity recognition. The entity tags are added to the ASR model's vocabulary as special tokens, allowing the model to recognize these tags as distinct entities during training and prediction.
- Model Training:
  Each audio file is paired with its annotated transcription. As the entity tagging information is presented as special tags in the transcription, a standard ASR model training process is performed on those data. The model learns the associations between sound patterns in the audio and the sequence of tokens in the text, including recognizing entity boundaries and context. The loss function penalizes incorrect transcriptions and incorrect entity tag placements.

This end-to-end approach allows the model to learn the complex relationships between speech input and named entity recognition simultaneously, potentially leveraging acoustic cues that might be lost in a pipeline approach.

IV. Experiments and Analysis

A. Data

1) Seed Data: The initial dataset was extracted from National Speech Corpus(NSC) samples and subsequently enriched with diverse PII entities based on predefined patterns. This seed dataset comprises 992 sentences, serving as the foundation for our data augmentation process.

2) Augmented Data: The seed dataset is initially expanded using the Gretel tool [16] to generate synthetic data that preserves the original dataset's structural integrity. We refer to the combined seed and Gretel data as the baseline dataset.

Building upon this baseline, two new data augmentation techniques are explored. For creating similar augmented data, we employed the Meta-Llama-3.1 model[4] with two temperature settings: 0.3 and 0.6. In contrast, for generating diverse augmented data, we utilized the Meta-Llama-3.1 language model with the same two temperature settings: 0.3 and 0.6.

3) Test Data: To evaluate speech de-identification effectiveness, a hold-out set of 150 entity-rich sentences was extracted from the seed data, ensuring no overlap with the data used for augmentation. Each sentence was read by one of four speakers to obtain corresponding audio data.

B. Evaluation metrics

In building a speech de-identification system, the following evaluation metrics are utilized:

1) Word Error Rate (WER): For ASR performance evaluation, we calculated the word error rate. The WER was computed by comparing the original transcription files (true) with the output transcriptions from the Whisper small model (hypothesis). Both sets of files underwent identical preprocessing steps before WER calculation. The resulting WER on the test data was approximately 17%.

2) F1 Score Calculation: To evaluate the performance of detecting PIIs from speech, we calculate F1 scores using the following procedure:

- Timestamp-based Alignment: We utilized the Vosk generic US English model[5] to align spoken words in the audio files with their corresponding timestamps, producing aligned transcriptions.
- Entity Matching: The alignment process matched predicted entities with true entities based on timestamp overlaps. For each true entity, we searched for a predicted entity with overlapping timestamps. Matching entities were added to the aligned entities list, while non-matches were labeled as 'O' (no entity).
- Scoring: Predicted entities overlapping with true entities were counted as correct predictions, while non-overlapping predictions were considered false positives. These counts were used to compute precision, recall, and F1 scores for both overall performance and individual entity types.

C. Experiments

For our experiments, we utilized the small model of Whisper in pipeline method. In end-to-end approach, we fine-tuned this model for our specific speech de-identification task.

A baseline model is constructed using a combined dataset of seed and Gretel-generated data. We compare the performance of the proposed Similar and Diverse augmentation approaches by adding augmented data to this baseline.

[4]https://llama.meta.com/

[5]https://alphacephei.com/vosk/

Tables I and II present PII detection results for pipeline and end-to-end models, respectively. The PII detection performance are reported for various training data combinations.

A comparative analysis of Tables I and II clearly demonstrates the superior performance of the end-to-end model across all training data configurations. The results consistently show that incorporating augmented data significantly improves PII detection compared to the baseline, validating the effectiveness of our proposed augmentation methods.

We subsequently investigated the influence of the temperature parameter on data augmentation. In large language models, temperature controls the degree of randomness in generated text. Higher temperatures produce more diverse and unpredictable outputs.

Comparison of results revealed that lower temperature values yielded better performance for similar augmentation, while higher temperatures excelled in diverse augmentation. This aligns with our intuition: similar augmentation benefits from closely resembling the seed data, thus requiring lower randomness, whereas diverse augmentation thrives on introducing variability, hence the preference for higher temperatures.

TABLE I
PIPELINE MODEL PERFORMANCE

Model	Precision %	Recall %	F1 %
Baseline (Seed+Gretel)	79.64	55.70	63.21
Baseline+Diverse0.3	85.36	55.03	65.69
Baseline+Diverse0.6	78.43	57.05	65.32
Baseline+Similar0.3	81.66	57.04	65.99
Baseline+Similar0.6	86.01	59.73	68.67

TABLE II
END-TO-END MODEL PERFORMANCE

Model	Precision %	Recall %	F1 %
Baseline (Seed+Gretel)	86.05	56.38	67.72
Baseline+Diverse0.3	81.61	61.07	69.55
Baseline+Diverse0.6	87.70	65.77	74.68
Baseline+Similar0.3	93.71	69.13	78.80
Baseline+Similar0.6	83.51	61.07	69.96

V. CONCLUSION

This paper introduces a fully automated data augmentation method for speech de-identification that leverages large language models. We explore two augmentation strategies and investigate the impact of the temperature parameter on their effectiveness. Experimental results validate the efficacy of data augmentation in improving PII detection performance. Additionally, our findings demonstrate the superiority of end-to-end modeling over traditional pipeline approaches for speech de-identification by jointly optimizing speech recognition and PII tagging.

ACKNOWLEDGMENT

This work is supported by Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant R-R12-A405-0009.

REFERENCES

[1] 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

[2] Neamatullah, Ishna, Margaret M. Douglass, Li-Wei H. Lehman, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford. "Automated de-identification of free-text medical records." BMC medical informatics and decision making 8 (2008): 1-17.

[3] Lothritz, Cedric, Bertrand Lebichot, Kevin Allix, Saad Ezzini, Tegawendé Bissyandé, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. "Evaluating the Impact of Text De-Identification on Downstream NLP Tasks." In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pp. 10-16. 2023.

[4] Wen, Yunqian, Bo Liu, Jingyi Cao, Rong Xie, and Li Song. "Divide and conquer: a two-step method for high quality face de-identification with model explainability." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5148-5157. 2023.

[5] Zhu, Bingquan, Hao Fang, Yanan Sui, and Luming Li. "Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 414-420. 2020.

[6] Abdi, Noura, Xiao Zhan, Kopo M. Ramokapane, and Jose Such. "Privacy norms for smart home personal assistants." In Proceedings of the 2021 CHI conference on human factors in computing systems, pp. 1-14. 2021.

[7] Espinoza-Cuadros, Fernando M., Juan M. Perero-Codosero, Javier Antón-Martín, and Luis A. Hernández-Gómez. "Speaker de-identification system using autoencoders and adversarial training." arXiv preprint arXiv:2011.04696 (2020).

[8] Ido Cohn et al. "Audio De-identification - a New Entity Recognition Task". ACL 2019, pp. 197–204. DOI: 10.18653/v1/N19-2025.

[9] Parada, C., Dredze, M., Jelinek, F. (2011) OOV sensitive named-entity recognition in speech. Proc. Interspeech 2011, 2085-2088, doi: 10.21437/Interspeech.2011-547

[10] Piotr Szymanski et al. "Why Aren't We NER Yet? Artifacts of ASR Errors in Named Entity Recognition in Spontaneous Speech Transcripts". ACL 2023, pp. 1746–1761. DOI: 10.18653/v1/2023.acl-long.98.

[11] Yadav, Hemant, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. "End-to-end named entity recognition from english speech." arXiv preprint arXiv:2005.11184 (2020).

[12] Ghannay, S. C. et.al, "End-to-end named entity and semantic concept extraction from speech". SLT 2018, IEEE, pp. 692–699.

[13] 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

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

[15] Jiong Cai et al. "Improving Low-resource Named Entity Recognition with Graph Propagated Data Augmentation". In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2023, pp. 110–118

[16] Gretel Synthetics. https : / / github. com / gretelai / gretel synthetics. 2024 (accessed May 20, 2024).

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