OPEN ACCESS
International Journal of Asian Language Processing
Vol. 35, No. 1 (2025) 2450014 (18 pages)
© The Author(s)
DOI: 10.1142/S2717554524500140

World Scientific
www.worldscientific.com

Leveraging Large Language Models for Speech De-Identification

Priyanshu Dhingra
Rajiv Gandhi Institute of Petroleum Technology
Mubarakpur Mukhatiya, Uttar Pradesh 229305, India
priyanshudhingra1@gmail.com

Satyam Agrawal
National Institute of Technology Karnataka
Mangaluru, Karnataka 575025, India
lmcsatyam@gmail.com

Chandra Sekar Veerappan
Singapore Institute of Technology
Singapore 828608, Singapore
chandra.veerappan@singaporetech.edu.sg

Eng Siong Chng
Nanyang Technological University
Singapore 639798, Singapore
aseschng@ntu.edu.sg

Rong Tong*
Singapore Institute of Technology
Singapore 828608, Singapore
tong.rong@singaporetech.edu.sg

Received 9 November 2024
Accepted 17 December 2024
Published 27 January 2025

This paper presents a novel approach to address the scarcity of labeled data in speech de-identification, a critical task for protecting personal privacy. By leveraging a large language model, we propose a fully automated data augmentation strategy that generates synthetic speech text data enriched with diverse personally identifiable information (PII) entities. This augmented dataset is then used to train the speech-de-identification models, significantly

*Corresponding author.

improving its performance on spoken language. To further enhance de-identification accuracy, we explore both 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. Our 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.

Keywords: Speech De-identification; Data Augmentation; Named Entity Recognition; Large Language Model; End-to-End Model.

1. Introduction

The rapid digitization of our world has led to an unprecedented increase in the collection and dissemination of personal data. While this has enabled numerous technological advancements, it has also raised significant privacy concerns. De-identification, the process of removing personally identifiable information (PII) from data, is crucial for safeguarding sensitive information.

Speech data, particularly in the form of audio recordings, can contain a wealth of PII, including names, addresses, phone numbers, and social security numbers. While techniques exist to de-identify speaker-specific biometrics, such as voice prints, less attention has been given to the de-identification of nonbiometric information within speech content. This is particularly relevant in domains like customer service, healthcare, and law enforcement, where sensitive information is often captured in audio recordings.

This paper addresses the challenge of automatically de-identifying PII from speech data. By developing robust techniques to remove or mask sensitive information, we aim to enhance the privacy of individuals while enabling the ethical use of speech data for research and development. Our work builds upon the significant efforts made in the fields of speech processing and privacy, such as the ASVSpoof Challenge[1] and the Voice Privacy Challenge,[2] which have primarily focused on speaker biometrics. By extending these efforts to nonbiometric information, we aim to contribute to the advancement of privacy-preserving speech technologies.

Figure 1 illustrates a typical speech de-identification system. This system aims to protect sensitive information by identifying PII entities, locating their positions within the speech, and subsequently removing or redacting them.

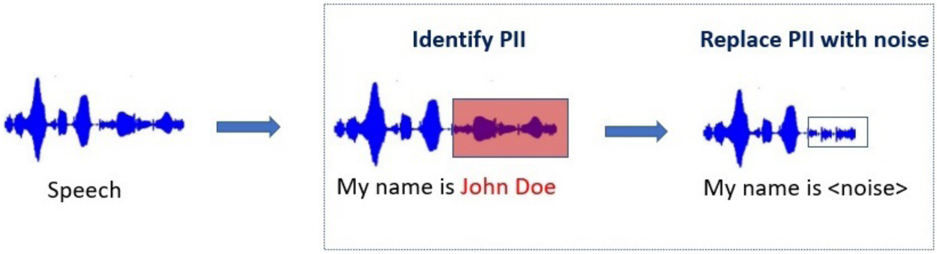

Fig. 1. Example of a speech de-identification system.

Automatic speech de-identification typically involves processing the speech signal, detecting PII entities, and finally obfuscating or removing them from the original audio data. However, constructing such a system necessitates speech data with extensive PII annotations. Publicly available datasets often prioritize common identifiers like names and public information, neglecting less frequent PIIs such as passport numbers or IDs. Moreover, generating large speech datasets with labeled PII is hindered by privacy concerns, as individuals are understandably reluctant to share recordings containing their personal information.[3]

A significant challenge in developing effective de-identification systems is the scarcity of publicly available speech datasets annotated with PII. This lack of labeled data limits the ability to train robust models. To address this issue, we propose leveraging the power of large language models (LLMs). Trained on massive amounts of text data, LLMs have demonstrated remarkable capabilities in various natural language processing tasks, including text generation, summarization, and translation. Their ability to understand and generate human-like text makes them promising candidates for assisting in the development of speech de-identification systems.

Traditional approaches to speech de-identification often involve a two-step process: speech-to-text transcription followed by named entity recognition (NER). This sequential approach can amplify errors from both systems. To address this limitation, we propose an end-to-end (E2E) approach that jointly models speech and entity categories, eliminating the need for intermediate transcriptions.

This paper presents a novel approach to speech de-identification that leverages the power of LLMs. We propose a data-driven framework that combines LLM-based data augmentation with advanced de-identification models. Our contributions include the following:

- A novel data augmentation strategy: We introduce a method to generate synthetic speech data enriched with PII using LLMs.
- Development of E2E de-identification models: We explore the use of E2E models that jointly learn speech recognition, language understanding, and de-identification.

2. Related Work

2.1. *Text-based system*

Researchers have made significant strides in text de-identification over the years. An early breakthrough came when[4] introduced artificial neural networks (ANN) to tackle this challenge — a novel approach at the time. They focused specifically on de-identification of personal details from medical records, and their neural network proved more effective than the CRF-based systems that dominated the field previously. The results were particularly impressive across key performance metrics including $F1$ scores, precision, and recall.

Building on this foundation, Liu *et al.*[5] later developed a more sophisticated approach. Rather than relying on a single method, they created a ensemble system comprised several individual subsystems which combined a bi-directional LSTM neural network with both CRF modeling and rule-based components. By integrating these different approaches, their system could more effectively identify and remove sensitive information from clinical documents.

Text-based de-identification often relies on NER to detect sensitive information. Deep learning has become a cornerstone in text-based NER tasks. Li *et al.*[6] provided a comprehensive taxonomy of existing approaches, categorizing them based on input representations, context encoders, and tag decoders.

LLMs have also been explored for NER. Wang *et al.*[7] introduced GPT-NER, a method that leverages the capabilities of LLMs by transforming NER into a text generation task. This approach allows LLMs to directly generate text sequences with embedded entity labels, bridging the gap between traditional sequence labeling and the strengths of LLMs.

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.[8,9] Similarly, speech data can contain explicit PII 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,[10,11] there are less works on de-identification of nonbiometric information in speech content. The automatic de-identification on the speech content is under explored. This paper addresses the critical task of protecting sensitive, nonbiometric information embedded in speech.

2.2. *Data augmentation*

Real-world language applications often encounter challenges such as typos, informal language, slang, and diverse sentence structures. These complexities can hinder the performance of models trained on clean, annotated data. Data augmentation addresses this issue by introducing these variations into the training set, enhancing the model's robustness and adaptability to different writing styles and potential errors.

Ding *et al.*[12] proposed a novel augmentation method to generate high-quality synthetic data for low-resource tagging tasks using language models trained on linearized labeled sentences. Pellicer *et al.*[13] conducted a comprehensive study of NLP data augmentation techniques, comparing their relative performance under various conditions. Their analysis revealed that data augmentation is particularly beneficial in data-scarce scenarios, significantly improving model performance. Building on previous work,[14] the authors adapted existing data augmentation techniques for sentence-level and sentence-pair NLP tasks to the specific needs of NER.

Recent advancements in data augmentation have also been explored for document-level relation extraction and NER. Sun *et al.*[15] employed a chain-of-prompt method with LLMs, while Hemateja *et al.*[16] proposed a data augmentation method with a sanity-checker to improve transformer-based NER models. Cai et *al.*[17] introduced the graph propagated data augmentation (GPDA) framework, leveraging graph propagation to build relationships between labeled and unlabeled data for NER.

Data augmentation has also been explored in speech processing. Ko *et al.*[18] evaluated audio augmentation techniques with low implementation costs. Their experiments demonstrated that speed perturbation, which emulates both vocal tract length perturbation (VTLP) and tempo perturbation, significantly improves automatic speech recognition (ASR) performance. Meng *et al.*[19] proposed a simple yet effective data augmentation method based on mixup for ASR. To address the scarcity of patient data, Vachhani *et al.*[20] explored data augmentation using temporal and speed modifications to healthy speech to simulate dysarthric speech. Geng *et al.*[21] investigated the impact of VTLP, tempo perturbation, and speed perturbation on speech data augmentation, demonstrating improvements in disordered speech recognition. Liang *et al.*[22] introduced a novel data augmentation technique by employing text-based speech editing models. This approach generates augmented speech that is more coherent, diverse, and closely resembles natural speech.

2.3. *Speech de-identification system*

Early speech de-identification methods typically employed a pipeline architecture: ASR to convert audio to text, followed by NER to identify and remove sensitive information. Cohn *et al.*[23] made significant contributions to this field, proposing a novel audio de-identification framework and an innovative evaluation metric.

A major challenge in these pipeline approaches was handling out-of-vocabulary (OOV) words. Parada *et al.*[24] addressed this by augmenting NER systems within speech recognition, treating OOV words as distinct regions and utilizing contextual information for entity detection.

Szymański *et al.*[25] provided a comprehensive analysis of ASR-NER systems, highlighting several critical limitations. They identified shortcomings in traditional evaluation metrics and emphasized the lack of representative annotated datasets that capture the nuances of spontaneous speech. Moreover, their analysis revealed a fundamental challenge: errors in the ASR component often cascade through the pipeline, adversely affecting NER performance.

These limitations in pipeline approach led researchers to explore E2E approaches, where ASR and NER components could be trained jointly.[26,27] In contrast to the traditional pipeline approach, these methods aim to learn both speech recognition and NER tasks simultaneously within a single neural network architecture. In Ref. 27, a deep neural network is first trained for the ASR task. Subsequently, the final layer of the network is reinitialized and fine-tuned for the NER task. This allows the model to leverage the learned features from speech recognition during the NER stage. Chen *et al.*[28] proposed to incorporate 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.

The development of publicly available annotated speech datasets has been crucial for advancing speech de-identification. Yadav *et al.*[26] introduced the first publicly available English speech dataset with NER annotations, enabling the joint training of ASR and NER models. Chen *et al.*[28] presented a similar dataset for Chinese speech.

3. Proposed Solution

3.1. *Challenges*

Speech processing is particularly difficult in relation to text processing due to its informal nature and nonstandard structures. Spoken language often contains slang, contractions, and incomplete phrases. The use of disfluencies like *um* and *uh* can be a hindrance to the fluent speech which makes it difficult to accurately extract and identify PII. Consider these two spoken sentences as examples:

(1) *The day before yesterday, Ram received another email from r e m y at outlook dot sg.*
(2) *My phone number is (uh) eight five eight two nine three one one.*

In the first sentence, the email address: ***r e m y at outlook dot sg*** is spelled out phonetically, which is a format rarely encountered in written text. Similarly, the second sentence presents challenges for phone number recognition due to the inclusion of the filler word (***uh***) and the phone number's representation.

Singapore English, or Singlish, is a unique language variety that blends standard English with elements of Malay, Mandarin, Tamil, and other local dialects. While it shares similarities with standard English, Singlish has its own distinct grammar and vocabulary. For instance, it often omits subject pronouns and uses "can" to signal possibility. To effectively train speech processing models for Singlish, it is crucial to generate synthetic data that accurately reflects its linguistic characteristics.

Traditional speech recognition and NER datasets often focus on common PII types, such as names, addresses, and phone numbers. This can lead to under representation of less common but important entities like passport numbers. Additionally, privacy concerns limit the availability of large-scale, high-quality, labeled PII datasets, making it challenging to train robust speech de-identification models.

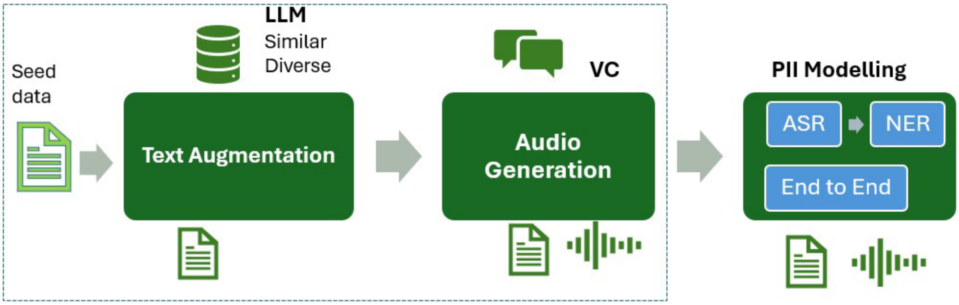

Fig. 2. Proposed speech de-identification system training framework.

3.2. *Speech de-identification framework*

Figure 2 presents a flowchart of our speech de-identification system training framework, highlighting the proposed spoken data augmentation (depicted in the dotted box).

We initiate the process with a small, carefully curated seed text dataset that accurately represents Singaporean English, incorporating a diverse range of PII entities.[29] Leveraging these seed data, we introduce fully automated data augmentation methods targeting both similar and diverse variations.[30]

Subsequently, we generate speech data from the augmented text data using voice conversion (VC) techniques. The resulting speech and text pairs serve as input for PII modeling.

We explore two approaches for PII modeling. The pipeline method involves a two-step process: first, an ASR system converts speech into text transcriptions, and then an NER system identifies PII entities within the text. In contrast, the E2E method directly predicts PII entities from the speech input, eliminating the need for intermediate text transcriptions.

3.3. *Spoken data augmentation*

3.3.1. *Seed data*

To create a seed dataset that reflects the unique linguistic features of Singaporean English, we utilize the NSC corpus.[31] This open-source speech dataset, specifically developed for Singaporean English, provides human-labeled text transcriptions. We leverage these transcriptions to identify and incorporate a diverse range of PII entities, including less common types like passport numbers and IDs, into our seed dataset. This ensures that our model is trained on a dataset that is both representative of Singaporean English and rich in PII information.

Table 1 lists the customized entity categories representing PII information specific to the Singaporean context.

Table 1. Customized PII categories.

PII entity	Description
NRIC	Start and end with alphabet, 7 digits
PASSPORT_NUM	Start with K, 7 digits, 1 alphabet
PHONE	Telephone number, 8 digits, starts with 6, 8, or 9
EMAIL	Email address
CREDIT_CARD	Credit number, 16 digits, first digits is 3, 4, 5, or 6
CAR_PLATE	Car plate number, S with 2–3 alphabets 4 digits
BANK_ACCOUNT	Singapore bank account number

3.3.2. *Text augmentation*

To address the scarcity of real-world annotated speech data, we propose a data augmentation method involving the following steps:

(1) **Example Data Extraction**: Sentences containing target NER tags are extracted from the seed dataset to serve as templates for generating new sentences with similar PII types.
(2) **LLM Prompting**: PII tags and optional structural information are provided as prompts to the LLM to guide the generation of coherent sentences with accurate entity alignment.
(3) **Sentence Generation**: The LLM generates new sentences based on the prompts, mimicking natural conversational patterns in Singaporean English.
(4) **Tag Replacement**: A tag replacement process is applied to the synthetic sentences to diversify the dataset and introduce variations in entity presentation, while maintaining consistent labeling.

To ensure comprehensive coverage of both typical and edge cases, our data augmentation approach incorporates two strategies:

- **Similar-based Augmentation (Similar)**: This approach provides the LLM with examples containing specific entity tags, instructing it to generate new sentences that adhere closely to the original structure. This ensures that the generated text aligns with the linguistic patterns of the seed dataset, making it suitable for scenarios requiring precise entity placement. Figure 3 shows the prompt used for similar data generation.
- **Diverse-based Augmentation (Diverse)**: This strategy provides the LLM with PII tags but allows greater freedom in sentence structure and word choice. By generating more diverse examples, this approach enriches the dataset with a broader range of linguistic patterns and contexts, improving the model's ability to handle diverse input variations. This is particularly valuable for capturing the inherent variability of natural speech. Figure 4 shows the prompt used for diverse data generation.

```
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. 3. Data augmentation prompt with Similar criteria.

```
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. 4. Data augmentation prompt with Diverse criteria.

Integrating both augmentation methods with the LLM's automatic annotation feature offers significant advantages. By incorporating PII tags into generated sentences, the process becomes highly efficient, reducing the need for manual annotation and accelerating dataset creation. This automated approach surpasses traditional manual methods, which are often time-consuming and limited in scale.

The proposed data augmentation method combining Similar and Diverse augmentation techniques, this enables the production of a large, high-quality, annotated dataset. This enhanced dataset empowers the speech de-identification model to accurately identify PII in spontaneous speech, improving its performance and adaptability to diverse linguistic contexts like Singaporean English. This automated approach effectively addresses the challenge of data scarcity and ensures the model's suitability for real-world applications.

3.3.3. *Audio generation*

To generate the required speech data for our training process, we leverage WhisperSpeech,[a] a state-of-the-art open-source text-to-speech (TTS) system. Beyond its core TTS capabilities, WhisperSpeech offers valuable voice cloning abilities. By providing an audio sample of a specific speaker, the system can accurately mimic their voice characteristics. While this opens doors for personalized audio content creation, our primary focus here is on its ability to generate speech that reflects the nuances of Singaporean English.

We utilize WhisperSpeech to convert our augmented text data, rich in diverse PII entities and Singaporean English characteristics, into high-quality speech. This process creates a text-and-audio paired dataset directly suited for training our speech de-identification model. Furthermore, the voice cloning allows us to synthesize speech samples from real-world Singaporean English recordings. This enriches the dataset with natural variations and authentic speech patterns, enhancing the model's ability to generalize to real-world scenarios.

3.4. ***PII modeling***

This work explores two distinct approaches for speech de-identification: pipeline and E2E models. Each method offers advantages and challenges, particularly when dealing with the complexities of Singaporean English and the accurate identification of PII.

3.4.1. *Pipeline approach*

This method tackles de-identification in two sequential steps: ASR followed by NER.

For the ASR component, we utilize the Whisper model,[b] fine-tuned on conversational datasets from the Singaporean English National Speech Corpus (NSC). This adaptation enables the model to better handle the unique linguistic features of Singapore English, resulting in more accurate transcriptions.

For the NER component, we employ a custom spaCy model[c] trained on an augmented dataset, encompassing a wide range of PII entities. This model is designed to effectively identify PII within the transcriptions generated by the ASR system.

The modular design of the pipeline approach allows for independent fine-tuning of each component. However, it is susceptible to error propagation. Inaccuracies in the ASR transcriptions can negatively impact subsequent NER performance.

[a] https://github.com/collabora/whisperspeech
[b] https://github.com/openai/whisper
[c] https://spacy.io/api/entityrecognizer/

3.4.2. *End-to-end approach*

The E2E methodology aims to integrate speech recognition and entity recognition into a single model. This unified approach allows the model to directly process audio files, producing transcriptions with embedded entity tags. By leveraging acoustic information, the model can more accurately identify and locate PII within the speech input. The E2E model follows these steps:

(1) **Data Preparation**:
The training data for the E2E model consist of audio files paired with corresponding transcriptions containing annotated entity tags. For instance, a transcription might be labeled as follows:

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

These annotations guide the model to recognize and tag entities directly from the speech input during training.

(2) **Tokenization and Special Tokens**:
The text is broken into smaller units called tokens, with each word and entity tag treated as distinct tokens. Since entity tags like `[PERSON]` or `[CAR_PLATE]` are not part of the standard English vocabulary, they are added to the ASR model's tokenizer as special tokens. This allows the model to identify these tags as unique entities during training, improving its ability to accurately detect boundaries for PII entities within the transcribed text.

(3) **Model Training**:
During training, each audio file is paired with its annotated transcription. The model learns to associate audio features with the sequence of tokens, including entity tags. The loss function penalizes errors in both transcription and entity tagging, ensuring the model learns to minimize errors in both tasks simultaneously. This unified training approach enables the E2E model to directly generate transcriptions with embedded entity labels from speech input.

4. Experiments and Analysis

4.1. *Data*

- Seed Data
An initial dataset was extracted from the NSC and enriched with diverse PII entities, including names, email addresses, phone numbers, and other specific categories listed in Table 1. This dataset, comprising 992 sentences, was split into training (80%) and testing (20%) sets.

The training portion was further expanded using the Gretel tool.[32] A model was trained on the seed data to learn its linguistic patterns, enabling it to generate synthetic sentences that closely resemble the original style and structure. These synthetic sentences were then subjected to a two-step annotation process: automatic annotation followed by human expert review to ensure accurate PII labeling.[29] For simplicity, we refer to the combined set of original and synthetic sentences as the seed dataset.

The seed dataset comprises a total of 3592 sentences, combining the original 992 sentences with 2600 expanded sentences generated using the Gretel tool.

- Synthetic Data

  The synthetic data are generated from the seed data following the text augmentation process described in Sec. 3.3.2. To generate synthetic data, we employed the Meta-Llama-3.1 model[d] with temperature settings of 0.3 and 0.6. Two distinct augmentation strategies were utilized:

  Similar Augmentation: The model was prompted to generate sentences that closely adhered to the structure and style of the seed data. By using a lower temperature setting (0.3), the model produced more focused and coherent sentences, ensuring that the generated data remained aligned with the linguistic characteristics of the original seed data.

  Diverse Augmentation: In contrast, the model was prompted with different prompts to generate sentences with a broader range of linguistic patterns and contexts. A higher temperature setting (0.6) was employed to encourage the model to explore more creative and diverse sentence structures, thereby enriching the dataset with a wider variety of language styles.

  By combining these two augmentation techniques, we were able to create a synthetic dataset that effectively captured the nuances of Singaporean English while also introducing sufficient diversity to enhance the model's generalization capabilities.
- Test Data

  A separate hold-out test set of 150 sentences, enriched with PII entities, was created. These sentences were distinct from those used for augmentation. Audio recordings of these sentences were obtained by having speakers read them aloud, simulating real-world conditions. This test set was used to evaluate the performance of the final de-identification model on unseen data.

4.2. *Evaluation metrics*

The $F1$ measure was used to assess the overall accuracy of PII detection, combining precision and recall. This metric provides a balanced evaluation of both the model's ability to correctly identify PII entities (precision) and its ability to identify all relevant entities (recall).

To evaluate the performance of detecting PIIs from speech, we calculate $F1$ scores using the following procedure:

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

- Timestamp-based Alignment: We utilized the Vosk generic English model[e] 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 nonmatches were labeled as 'O' (no entity).
- Scoring: Predicted entities overlapping with true entities were counted as correct predictions, while nonoverlapping predictions were considered false positives. These counts were used to compute precision, recall, and $F1$ scores for both overall performance and individual entity types.

4.3. *Experimental results*

For our experiments, we utilized the small variation of Whisper ASR model in the pipeline method, fine-tuning the ASR model on the Singaporean English NSC to better capture the nuances of local speech patterns. This adaptation resulted in a word error rate of approximately 17% on the test set.

In the E2E approach, we further fine-tuned this model for our specific speech de-identification task.

A baseline model was constructed using the original seed data only. We compared the performance of this baseline model with models trained on data augmented using the Similar and Diverse augmentation techniques.

Tables 2 and 3 present PII detection results for pipeline and E2E models, respectively. The PII detection performance are reported for various training data

Table 2. Pipeline model performance.

Model	Precision %	Recall %	$F1$ %
Baseline	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 3. End-to-end model performance.

Model	Precision %	Recall %	$F1$ %
Baseline	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

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

combinations. 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 LLMs, 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.

Comparing the pipeline and E2E results, it is clear that the E2E model substantially outperforms the pipeline model, even when trained on the same dataset. This highlights the benefits of jointly optimizing speech recognition and entity tagging within a single model. The E2E model's ability to directly learn the mapping between audio input and entity labels leads to improved accuracy and robustness.

5. Conclusion and Future Works

5.1. *Conclusion*

This paper presents a novel approach to automatic speech de-identification that leverages the power of LLMs. To address the scarcity of labeled speech data, we introduce a data augmentation technique that generates high-quality synthetic data, enriched with diverse PII entities. We then explore two modeling approaches: the traditional pipeline approach and an E2E approach.

Our experimental results demonstrate the superiority of the E2E approach over the pipeline approach. The E2E model, by jointly optimizing speech recognition and PII tagging, achieves significantly better performance in identifying and removing PII from speech data. This improvement is particularly evident in the reduced error rates, especially in terms of false positives and false negatives.

This work contributes to the advancement of speech de-identification techniques, particularly in scenarios with limited annotated data. By leveraging the capabilities of LLMs and exploring innovative modeling approaches, we aim to enhance the privacy and security of speech data.

5.2. *Future works*

Future research directions include the following.

(1) Cross-lingual and Code-switching Scenarios: Extending our approach to handle code-switching scenarios, where multiple languages are mixed within a single

utterance, would be a valuable contribution. This would require incorporating language identification and translation techniques into the de-identification pipeline.

(2) Multimodal Data: Exploring the integration of multimodal data, such as visual cues from videos or images, can provide additional context for more accurate PII detection. This could involve leveraging techniques from computer vision to extract relevant visual information.

(3) Overfitting and Generalization: Addressing potential overfitting issues associated with using similar synthesized data is crucial. Techniques like data augmentation and regularization can be explored to mitigate this risk. Additionally, evaluating the impact of random variations introduced by our augmentation methods on model performance is essential.

(4) Privacy-preserving Techniques: Investigating privacy-preserving techniques to protect sensitive information during the training and inference phases of the de-identification system is important. This could involve differential privacy or federated learning approaches.

(5) Real-world Deployment: Developing robust and efficient deployment strategies for our de-identification system in real-world applications, such as call centers or healthcare settings, is a practical challenge. This would involve considering factors like computational resources, latency, and the need for continuous adaptation to evolving language patterns and PII types.

Acknowledgment

This work is supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 Grant R-R12-A405-0009 and Ignition Grant R-IE2-A405-00006.

ORCID

Priyanshu Dhingra https://orcid.org/0009-0007-7650-0756
Satyam Agrawal https://orcid.org/0009-0002-9032-0739
Chandra Sekar Veerappan https://orcid.org/0000-0003-0948-0568
Eng Siong Chng https://orcid.org/0000-0001-6257-7399
Rong Tong https://orcid.org/0000-0003-3410-8354

References

1. H. Delgado, N. Evans, J. Jung, T. Kinnunen, I. Kukanov, K. A. Lee, X. Liu, H.-J. Shim, M. Sahidullah, H. Tak, M. Todisco, X. Wang, J. Yamagishi and ASVspoof Consortium, ASVspoof 5 Evaluation Plan 2024.
2. N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi and M. Todisco, The VoicePrivacy 2024 Challenge Evaluation Plan, arXiv:2404.02677 (2024).

3. M. Flechl, S.-C. Yin, J. Park and P. Skala, End-to-end speech recognition modeling from de-identified data, in *Proc. Interspeech* (2022), pp. 1382–1386, doi: 10.21437/Interspeech.2022-10484.
4. F. Dernoncourt, J. Y. Lee, Ö. Uzuner and P. Szolovits, De-identification of patient notes with recurrent neural networks, *J. Am. Med. Inform. Assoc.* **24**(3) (2017) 596–606, doi:10.1093/jamia/ocw156.
5. Z. Liu, B. Tang, X. Wang and Q. Chen, De-identification of clinical notes via recurrent neural network and conditional random field, *J. Biomed. Inform.* **75**(2017) S34–S42.
6. J. Li, A. Sun, J. Han and C. Li, A survey on deep learning for named entity recognition, *IEEE Trans. Knowl. Data Eng.* **34**(1) (2020) 50–70.
7. S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li and G. Wang, Gptner: Named entity recognition via large language models, arXiv:2304.10428 (2023).
8. Y. Wen, B. Liu, J. Cao, R. Xie and L. Song, Divide and conquer: A two-step method for high quality face de-identification with model explainability, in *Proc. IEEE/CVF Int. Conf. Computer Vision* (IEEE, 2023), pp. 5148–5157.
9. B. Zhu, H. Fang, Y. Sui and L. Li, Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation, in *Proc. AAAI/ ACM Conf. AI, Ethics, and Society* (ACM, 2020), pp. 414–420.
10. N. Abdi, X. Zhan, K. M. Ramokapane and J. Such, Privacy norms for smart home personal assistants, in *Proc. 2021 CHI Conf. Human Factors in Computing Systems* (ACM, 2021), pp. 1–14.
11. F. M. Espinoza-Cuadros, J. M. Perero-Codosero, J. Antón-Martín and L. A. Hernández-Gómez, Speaker de-identification system using autoencoders and adversarial training, arXiv:2011.04696 (2020).
12. B. Ding, L. Liu, L. Bing, C. Kruengkrai, T. H. Nguyen, S. Joty, L. Si and C. Miao, DAGA: Data augmentation with a generation approach for low-resource tagging tasks, in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2020), pp. 6045–6057.
13. L. F. A. O. Pellicer, T. M. Ferreira and A. H. R. Costa, Data augmentation techniques in natural language processing, *Appl. Soft Comput.* **132**(2023) 109803.
14. X. Dai and H. Adel, An analysis of simple data augmentation for named entity recognition, arXiv:2010.11683 (2020).
15. Q. Sun, K. Huang, X. Yang, R. Tong, K. Zhang and S. Poria, Consistency guided knowledge retrieval and denoising in LLMs for zero-shot document-level relation triplet extraction, in *Proc. ACM on Web Conf. 2024* (ACM, 2024), pp. 4407–4416.
16. A. V. Hemateja, G. Kondakath, S. Das, M. Kothandaraman, S. Shoba, A. Pandey, R. Babu and A. Jain, Novel data augmentation for named entity recognition, *Int. J. Speech Technol.* **26**(4) (2023) 869–878.

17. J. Cai, S. Huang, Y. Jiang, Z. Tan, P. Xie and K. Tu, Improving low-resource named entity recognition with graph propagated data augmentation, in *Proc. 61st Annu. Meeting of the Association for Computational Linguistics* Short Papers, Vol. 2 (Association for Computational Linguistics, 2023), pp. 110–118.
18. T. Ko, V. Peddinti, D. Povey and S. Khudanpur, Audio augmentation for speech recognition, in *Proc. Interspeech* (2015), pp. 3586–3589, doi: 10.21437/Interspeech.2015-711.
19. L. Meng, J. Xu, X. Tan, J. Wang, T. Qin and B. Xu, Mixspeech: Data augmentation for low-resource automatic speech recognition, in *2021 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2021), pp. 7008–7012.
20. B. Vachhani, C. Bhat and S. K. Kopparapu, Data augmentation using healthy speech for dysarthric speech recognition, in *Proc. Interspeech* (2018), pp. 471–475, doi: 10.21437/Interspeech.2018-1751.
21. M. Geng, X. Xie, S. Liu, J. Yu, S. Hu, X. Liu and H. Meng, Investigation of data augmentation techniques for disordered speech recognition, arXiv:2201.05562 (2022).
22. Z. Liang, Z. Song, Z. Ma, C. Du, K. Yu and X. Chen, Improving code-switching and named entity recognition in ASR with speech editing based data augmentation, arXiv:2306.08588 (2023).
23. I. Cohn, I. Laish, G. Beryozkin, G. Li, I. Shafran, I. Szpektor, T. Hartman, A. Hassidim and Y. Matias, Audio de-identification — A new entity recognition task, in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Industry Papers, Vol. 2 (Association for Computational Linguistics, Minneapolis, MN, 2019), pp. 197–204.
24. C. Parada, M. Dredze and F. Jelinek, OOV sensitive named-entity recognition in speech, in *Proc. Interspeech* (2011), pp. 2085–2088, doi: 10.21437/Interspeech.2011-547.
25. P. Szymański, L. Augustyniak, M. Morzy, A. Szymczak, K. Surdyk and P. Żelasko, Why aren't we NER yet? Artifacts of ASR errors in named entity recognition in spontaneous speech transcripts, in *Proc. 61st Annu. Meeting of the Association for Computational Linguistics,* Long Papers, Vol. 1 (Association for Computational Linguistics, Toronto, Canada, 2023), pp. 1746–1761.
26. H. Yadav, S. Ghosh, Y. Yu and R. R. Shah, End-to-end named entity recognition from English speech, arXiv:2005.11184 (2020).
27. S. Ghannay, A. Caubrière, Y. Estève, N. Camelin, E. Simonnet, A. Laurent and E. Morin, End-to-end named entity and semantic concept extraction from speech, in *2018 IEEE Spoken Language Technology Workshop (SLT)* (IEEE, 2018), pp. 692–699.
28. B. Chen, G. Xu, X. Wang, P. Xie, M. Zhang and F. Huang, Aishell-ner: Named entity recognition from Chinese speech, in *2022 IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP*) (IEEE, 2022), pp. 8352–8356.

29. P. Dhingra, S. Agrawal, C. S. Veerappan, T. N. Ho, E. S. Chng and R. Tong, Speech de-identification data augmentation leveraging large language model, in *2024 Int. Conf. Asian Language Processing (IALP)* (IEEE, 2024), pp. 97–102.
30. P. Dhingra, S. Satyam, C. S. Veerappan, C. Eng Siong and R. Tong, Enhancing Speech De-identification with LLM-Based Data Augmentation, in *2024 11th Int. Conf. Advanced Informatics: Concept, Theory and Application (ICAICTA)* (IEEE, 2024), pp. 1–5.
31. J. X. Koh, A. Mislan, K. Khoo, B. Ang, W. Ang, C. Ng and Y. Tan, Building the Singapore English national speech corpus, in *Proc. Interspeech* (2019), pp. 321–325, doi: 10.21437/Interspeech.2019-1525.
32. Gretel Synthetics (2024), https://github.com/gretelai/gretel-synthetics.