Indonesian Speech Content De-Identification in Low Resource Transcripts

Rifqi Naufal Abdjul[1], Dessi Puji Lestari[1], Ayu Purwarianti[1], Candy Olivia Mawalim[2], Sakriani Sakti[3], Masashi Unoki[2],

[1]Bandung Institute of Technology, [2]Japan Advanced Insitute of Science and Technology, [3]Nara Institute of Science and Technology,

Correspondence: rifqi.abdjul23@gmail.com

Abstract

Advancements in technology and the increased use of digital data threaten individual privacy, especially in speech containing Personally Identifiable Information (PII). Therefore, systems that can remove or process privacy-sensitive data in speech are needed, particularly for low-resource transcripts. These transcripts are minimally annotated or labeled automatically, which is less precise than human annotation. However, using them can simplify the development of de-identification systems in any language. In this study, we develop and evaluate an efficient speech de-identification system. We create an Indonesian speech dataset containing sensitive private information and design a system with three main components: speech recognition, information extraction, and masking. To enhance performance in low-resource settings, we incorporate transcription data in training, use data augmentation, and apply weakly supervised learning. Our results show that our techniques significantly improve privacy detection performance, with approximately 29% increase in F1 score, 20% in precision, and 30% in recall with minimally labeled data.

1 Introduction

A considerable amount of private data is readily accessible online (Liu et al., 2021), often utilized for machine learning research leveraging publicly available information. While privacy concerns for text data have received attention (NAYAK et al., 2011), strategies to protect speech data remain underdeveloped. This imbalance highlights the critical need to implement robust privacy safeguards for all modality.

Speech privacy comprises two main categories: speaker identity and content privacy, with the latter, including sensitive spoken utterances like Personally Identifiable Information (PII), being relatively underexplored (Williams et al., 2021). This content may include spoken utterances that contain sensitive information, such as Personally Identifiable Information (PII). Exposure to PII risks severe consequences, such as losing control over personal information (Wright and Raab, 2014).

To protect the privacy of speech content, a method called speech content de-identification can be employed. This technique focuses on identifying private information and either removing it or substituting it with uniform noise. On the surface, de-identified data might seem unusable, but Flechl et al. (2022) have demonstrated that such data can still be useful for training a privacy-preserving speech recognition models without a significant drop in performance.

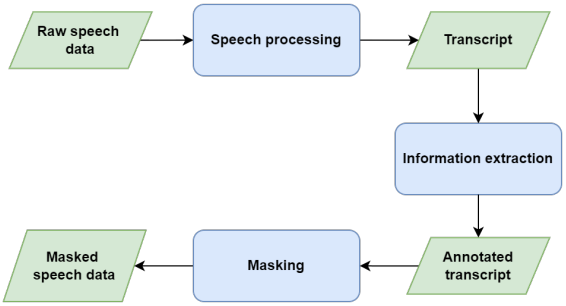

Figure 1: Overview of the speech content de-identification

Multiple studies on speech content de-identification have used text transcripts as intermediaries (Baril et al., 2022; Cohn et al., 2019; Kaplan, 2020). These systems typically consist of a speech recognition module, an information extraction module, and a masking module, as shown in Figure 1. Although prior research achieved positive results, their resource-intensive implementations are challenging to apply to low-resource languages, which often lack advanced privacy protection tools. Consequently, sensitive data in these languages are at greater risk of privacy breaches. To ensure privacy, de-identification

Proceedings of the Second Workshop in South East Asian Language Processing, pages 63–71
January 20, 2025.

systems must operate effectively in the target language despite limited resources.

The objective of this research is to develop a speech de-identification system that overcomes the challenges related to transcription and limited resources. To this end, we incorporate transcription data in the training, utilize data augmentation techniques, and apply weakly supervised learning. Our work contributes to improving the system efficiency, particularly when working with low-resource transcripts.

2 Related Work

Speech content de-identification involves the systematic removal of any PII from recorded speech, positioning it as a new entity recognition task (Cohn et al., 2019). For example, in a recorded speech that reads, "John came from Indonesia," a speech content de-identification system would process the data to redact any private information like "John" and "Indonesia." This ensures the anonymization of sensitive information within the speech data.

The main challenges in developing a speech de-identification system for low-resource languages like Indonesian include, but are not limited to, transcribing speech in these languages, processing the transcribed text, and effectively handling the unique characteristics of text in low-resource languages. Cohn et al. (2019) explains that the system performance is mostly dependent on the transcription result from the speech processing component. This is inline with Kumar et al. (2021); Hervé et al. (2022) as it states that the transcription text is a different domain than a normal text, which needs a transfer learning to improve the performance.

Numerous applications (Amazon; Microsoft) and research efforts have focused on speech de-identification systems in English (Kaplan, 2020; Cohn et al., 2019; Gouvêa et al., 2023) and other languages such as French (Baril et al., 2022). However, none of these studies address low-resource languages like Indonesian language, which suffer from a lack of annotated datasets and pre-trained models. This presents a significant problem, as such systems are highly language-dependent and may perform poorly when applied to languages that are either underrepresented in training data or fall outside the system's distribution.

There are ways to improve the system performance with multiple low-resource handling methods. Dai and Adel (2020) proves that using simple augmentation on a low-resource data could improve the performance on Named Entity Recognition (NER). This is because the variation sentence that it creates from augmentation can be learned as a new sentence by the language model. Other than that, methods like weakly-supervision training can improve the robustness with low quality data that usually can be seen on low-resource language data. A method like (Xu et al., 2023) to make model learn on so-called "predicted" false-negative data can be useful to increase the performance.

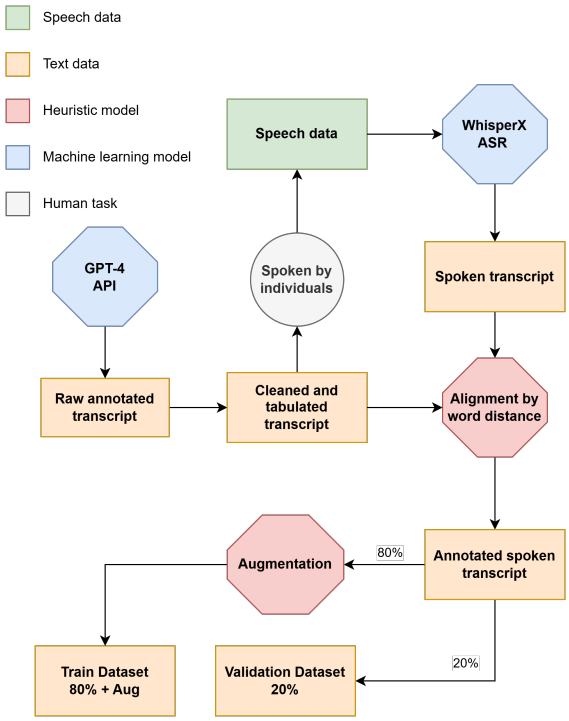

Figure 2: Speech dataset creation flowchart

Our research aims to combine, adapt, and improve multiple method to develop a speech content de-identification system and data creation pipeline tailored for low-resource languages, with the goal of enabling similar systems for languages with resource levels similar to Indonesian, ensuring privacy in speech-related contexts.

3 Proposed Method

To develop and improve the speech content de-identification system and methods, we need to create a dataset for the training and validation and a data processing pipeline that includes 3 main component, speech processing component, information extraction component, and masking component. After establishing the system, we imple-

mented three optimization methods tailored to the data domain: training on audio transcription text, dataset augmentation, and weakly supervised learning.

3.1 Dataset Creation

For model training and validation, we created both a transcribed speech dataset and a text-written dataset. The text dataset was generated using GPT-4, and individuals were recruited to speak the text, creating the speech dataset. The speakers included 10 personnel, consisting of 5 males and 5 females. Details of the dataset creation process are in Appendix A. Whisper Automatic Speech Recognition (ASR) (Radford et al., 2022) was used to transcribe the speech, and a modified WER algorithm aligned the generated and transcribed text, transferring labels while handling insertions, deletions, and substitutions. Figure 2 illustrates the process.

We created 86 dialogues across four topics: Job Interview, Medical Analysis, Bank Call Center, and Casual Conversation, totaling nearly two hours of speech data. The dialogues were carefully selected and relabeled to ensure the labels are of high quality and considered golden labels. Table 4 provides further details. While this may not be the most sophisticated approach, it is practical given the minimal effort required.

3.2 Baseline Implementation

As shown in the data processing flow diagram in Fig. 3, we developed three main components for speech recognition, information extraction, and masking to obfuscate speech data containing privacy-sensitive information.

The speech processing component was implemented using the WhisperX library (Bain et al., 2023). We chose the Whisper model due to its superior performance and its capability to predict punctuation, thereby enhancing the subsequent text processing stages. WhisperX also offers flexibility in the selection of forced alignment models, allowing the use of models specifically trained for the Indonesian language to ensure accurate forced alignment[1].

The information extraction component employs the mLUKE (Ri et al., 2022) language model for Named Entity Recognition (NER), which leverages the entity attention mechanism and entity embedding capabilities to process text. This approach enhances the performance and can be applied in weakly supervised methods to improve the model training efficiency on the dataset later.

We utilized a heuristic to transform speech segments into pink noise with intensity matched to the original speech, ensuring minimal disturbance to the listeners or users of the speech data (Cooper et al., 1985; Saeki et al., 2004). The procedure can be adjusted as needed, such as cutting out private information if noise replacement is unnecessary.

3.3 Training on Audio Transcription Text

Hervé et al. (2022) experimented on the use of transcript text, written text, and the combination of both for training a language model. Their evaluation showed that the most significant performance increase occurred when using a mix of transcript text and written text. We also found that this is true for the current environment. We therefore mixed the dataset using 50% of each dataset to make sure the NER model had a clear understanding of the grammar structure and able to consider the vocabulary of the spoken transcript.

Example 1
Original Sentence: Selamat pagi, saya Dokter **Surya** (B-PER). Anda datang untuk pemeriksaan rutin hari ini?
Translation: Good morning, I am Doctor **Surya** (B-PER). Are you here for a routine check-up today?
Mention Replacement: Selamat pagi, saya Dokter **Lisa Pratama** (B-PER, I-PER). Anda datang untuk pemeriksaan rutin hari ini?
Example 2
Original Sentence: Nomor telepon saya **081234567890** (B-TEL).
Translation: My phone number is **081234567890**.
Mention Replacement: Nomor telepon saya **082198765432** (B-TEL).

Table 1: Example of Mention Replacement Augmentation

3.4 Dataset Augmentation

For a simple augmentation on the dataset, we used the neraug library (Dai and Adel, 2020) to perform a mention replacement augmentation, with example on Table 1. This method was chosen over the more powerful augmentation ones because they re-

[1] https://huggingface.co/indonesian-nlp/wav2vec2-large-xlsr-indonesian

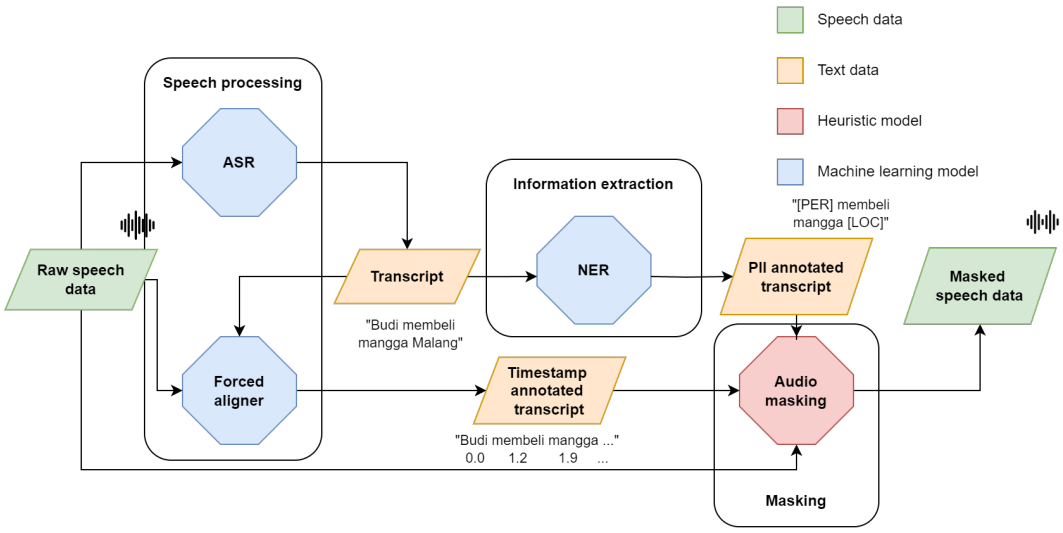

Figure 3: Detailed data flow on speech content de-identification system

move the context of the transcript and affect the performance (Giridhara et al., 2019). Also, privacy-labeled data usually lacks significant correlation between the labels themselves and with the context.

3.5 Weakly Supervised Learning

We adapted the Xu et al. (2023) method for weakly supervised learning based on weakly annotated data. The main idea of the process is creating an assumption that the false negative data have a high similarity with the true positive data. On the process of training loss calculation, we assign 10% of the most similar negative data to the positive batch and calculate them as positive data. To improve the method, we take advantage of the built-in mLUKE (Ri et al., 2022) entity embedding mechanism rather than using a separate model like Xu et al. (2023) did. Utilizing the entity embedding model from the mLUKE (Ri et al., 2022) language model offers several advantages:

- The process of training the model is more simplified where we can accomodate everything in a single loop rather than training the entity embedding model and the language model separately.
- The entity embedding model is typically more mature and more in line with the main language model that is being trained. The entity embedding itself can learn alongside the model giving a dynamic improvement rather than a static one.
- The entity embedding model fine-tuned on the specific dataset that is used can learn the specific domain (e.g., privacy data).

4 Experimental Setup

For the experiment, we split the current dataset into 80:20 for training and testing. To create variations of the systems, we created tags as follows.

- **A** means the training dataset is augmented using the augmentation process.
- **T** means the training dataset is mixed using the spoken transcript dataset. If this variation is mixed with the augmentation, this will be done first.
- **W** means the NER model is trained using the weakly supervised method.

These variants can be combined and used interchangeably: for example, the WAT variant means that the training dataset is mixed with the spoken transcript dataset and then augmented, and the model is trained using the weakly supervised method.

To simulate a low-resource environment (distantly annotated data), we removed percentages of labels on the dataset based on the "missing label" variable. The variable varies from 0 to 0.8 with 0.2 steps, where 0 means the label is complete and 0.8 means 80% of the label is randomly missing. To make sure there were no random variables in the experiment, we performed the evaluation five times and averaged the results.

To evaluate the speech content de-identification system, we utilized multiple evaluation metrics as follows:

- **WER and CER**, to evaluate the error from the speech recognition component. Word Error Rate (WER) measures the rate of errors in transcribed words, while Character Error Rate (CER) quantifies errors at the character level. We evaluated the speech dataset based on the written text and the spoken transcript dataset created by the Whisper ASR.
- **Seqeval (Nakayama, 2018)**, to evaluate the information extraction component based on precision, recall, and F1. We evaluated the component with the spoken transcript dataset as input.
- **Nerval (Blanche and Kermorvant, 2021)**, to evaluate the overall system based on precision, recall, and F1. We used 30% as the threshold for the CER in the library. We evaluated the overall system using the audio dataset as input and the spoken transcript dataset as reference.

5 Results

This section summarizes the experimental results, including the ASR evaluation, text component evaluation, reliability evaluation, error analysis, variant performance analysis, and overall system performance.

5.1 ASR Evaluation

Evaluation results of the ASR demonstrate a high WER with relatively low CER, as shown on Table 3. This occured because of the standardization of the spoken language: for example, the word 'nggak' was transcribed as 'ga'. Although this can lead to a higher WER, it should not affect the information extraction too much.

5.2 Text Component Evaluation

The experimental results depicted in Fig. 4 reveal that the system maintains a relatively high performance in both F1 and recall metrics, even with 40% to 60% missing labels. The WAT variant consistently exhibits a higher recall compared to other variants, indicating that combining various methods enhance the overall performance. The augmentation method shows the most significant performance improvement, especially when the missing label rate decreases, making the dataset more complete. The utilization of domain transcription data increases the performance only with perfect data or when combined with other methods. Weakly supervised learning notably enhances recall but reduces precision with perfect data. This method enables

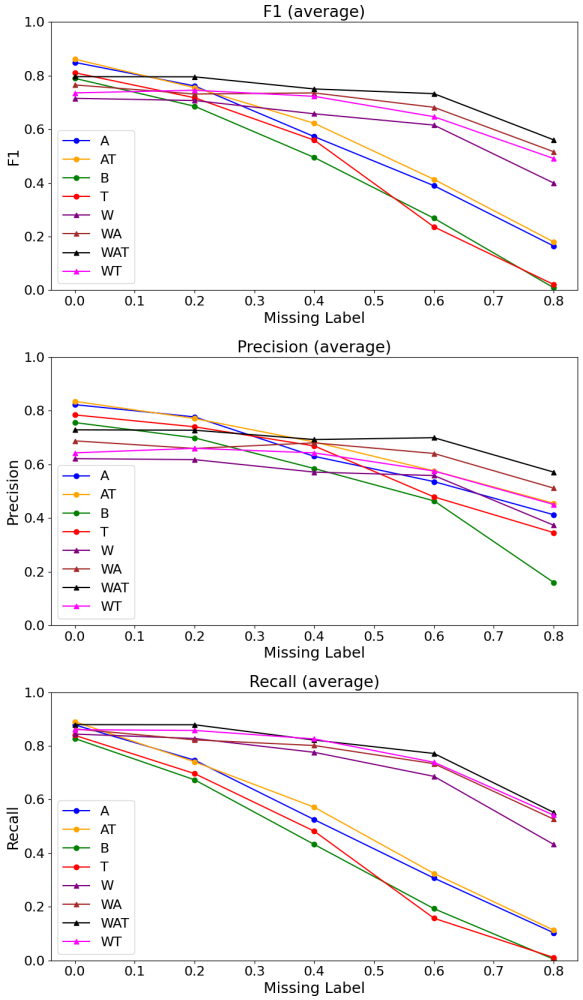

Figure 4: Evaluation results for information extraction component

the model to learn from only 20% of the total data annotations.

5.3 Reliability Evaluation

A standard deviation analysis was conducted to assess system reliability, categorizing deviations as low (<5%), moderate (5–10%), and high (>10%). These thresholds align with widely accepted standards, where deviations below 5% are negligible and those exceeding 10% are significant. The variation in the system is directly related to the model's stability and robustness with respect to changing data. The standard deviation values for each metric are provided in Fig. 5, showing low variation for recall and F1 metrics except in the baseline variant.

5.4 Error Analysis

The WAT variant with 0% missing labels exhibited several types of errors. A major problem was the omission of common nouns when label-

Type	Metric (Avg)	Portion of Missing Label									
		Text Component					Overall System				
		0	**0.2**	**0.4**	**0.6**	**0.8**	**0**	**0.2**	**0.4**	**0.6**	**0.8**
B	F1	0.790	0.686	0.495	0.268	0.009	0.739	0.640	0.456	0.246	0.004
	Precision	0.756	0.699	0.585	0.464	0.160	0.794	0.736	0.630	0.535	0.200
	Recall	0.828	0.674	0.432	0.193	0.005	0.692	0.566	0.362	0.165	0.002
A	F1	0.849	0.762	0.573	0.390	0.165	0.778	0.717	0.549	0.371	0.139
	Precision	0.823	**0.777**	0.630	0.536	0.412	0.824	0.826	0.715	0.647	0.499
	Recall	0.878	0.747	0.525	0.307	0.103	0.738	0.634	0.447	0.261	0.082
T	F1	0.811	0.717	0.559	0.236	0.021	0.756	0.687	0.530	0.228	0.018
	Precision	0.785	0.740	0.669	0.479	0.346	0.821	0.819	0.767	0.705	0.593
	Recall	0.839	0.697	0.482	0.157	0.011	0.701	0.593	0.405	0.140	0.009
AT	F1	**0.861**	0.755	0.622	0.413	0.180	**0.794**	0.723	0.605	0.400	0.167
	Precision	**0.835**	0.771	0.684	0.576	0.455	**0.847**	**0.846**	**0.799**	0.735	0.602
	Recall	**0.890**	0.741	0.571	0.324	0.113	0.748	0.631	0.487	0.274	0.096
W	F1	0.715	0.707	0.658	0.616	0.400	0.691	0.689	0.646	0.607	0.408
	Precision	0.621	0.618	0.571	0.558	0.373	0.671	0.685	0.583	0.530	0.448
	Recall	0.844	0.828	0.776	0.686	0.433	0.716	0.696	0.659	0.586	0.368
WA	F1	0.765	0.731	0.736	0.682	0.516	0.720	0.691	0.711	0.662	0.522
	Precision	0.688	0.660	0.680	0.641	0.512	0.706	0.674	0.744	0.682	0.531
	Recall	0.863	0.823	0.801	0.734	0.527	0.736	0.710	0.682	0.627	0.413
WT	F1	0.736	0.746	0.723	0.647	0.491	0.700	0.728	0.712	0.631	0.493
	Precision	0.643	0.660	0.643	0.575	0.451	0.677	0.732	0.736	0.675	0.549
	Recall	0.860	0.858	**0.826**	0.739	0.542	0.726	0.726	**0.696**	0.633	0.435
WAT	F1	0.796	**0.796**	**0.750**	**0.733**	**0.561**	0.753	**0.757**	**0.729**	**0.731**	**0.567**
	Precision	0.729	0.728	**0.693**	**0.700**	**0.571**	0.756	0.767	0.770	**0.817**	**0.700**
	Recall	0.879	**0.879**	0.822	**0.772**	**0.552**	**0.752**	**0.747**	0.694	**0.662**	**0.478**

Table 2: Evaluation Results for only text component (left) and the overall system (right). The highest value per metric and per missing label value are in bold.

Evaluation metric	Value
Word error rate (WER)	5.16%
Character error rate (CER)	2.22%

Table 3: Transcription evaluation results

ing locations and professions. Terms like 'hotel' in 'hotel harris' and 'cafe' in 'cafe kenangan' were frequently excluded from location labels. Similarly, professional terms like 'designer' in 'freelance graphic designer' and 'software' in 'software engineer' were often overlooked. These errors stem from the weakly supervised learning method, which can lead the model to misinterpret these terms as false positives because of their resemblance to non-private terms.

Another significant error was the misclassification of educational data as professional data. In the test data for the WAT variant, 142 out of 500 educational labels were incorrectly identified as professional data. This issue likely arises from the similarity between educational and professional terms, which can be difficult to distinguish without additional context.

Identification numbers were also frequently misclassified as phone numbers. In the WAT variant test data, 50 out of 136 identification number labels were incorrectly identified as phone numbers. This error stems from the model's inability to correctly interpret the context of these numbers, indicating a need for models with more parameters to enhance contextual understanding.

Furthermore, informal date or time expressions were often not detected by the model. Of 385 date labels, many informal expressions like "nanti tanggal 15 ya" (translated as "Later at the 15th") and "hari senin minggu depan" (translated as "Monday, next week") were missed. This shortfall highlights the model's limited proficiency in understanding informal Indonesian language, suggesting that training with more varied Indonesian text data could improve detection.

The error analysis for different variants based on Table 2 provided detailed insights into their performance and limitations. For the baseline variant, the most frequent errors were false positive detections of privacy data. This issue is likely due to the model overfitting to clean text domains, causing it

to misclassify transcription errors as entities. The higher performance of the B variant rather than the T variant with missing labels further illustrates this tendency.

5.5 Variant Performance Analysis

In the augmentation variant (variant A), the system performance is improved in general, but augmentation sometimes disrupted the context of the data, resulting in errors not present in the normal variant. Despite this, the overall performance of the augmentation variant was consistently higher than that of the normal variant, indicating the benefit of this method.

The transcription variant (variant T) was trained using a combined dataset of normal and transcribed data. This training allowed the system to recognize and account for transcription errors, thus improving the performance when utilized with other methods or perfect datasets. However, this variant's performance declined with datasets having minimal labels, highlighting its dependency on comprehensive data for optimal functionality.

The weakly supervised variant (variant W) aimed to improve the model's efficiency with incomplete datasets by learning from false negatives and unlabeled data. This method significantly boosted the performance with imperfect datasets, as the model could still extract valuable information despite missing labels. However, the variant became overly sensitive to data similar to true positives, leading to an increase in false positives with datasets containing minimum missing labels. This sensitivity suggests that while weakly supervised learning is advantageous for incomplete data, it requires careful calibration to prevent over-sensitivity to similar but incorrect data points.

5.6 Overall System Performance

The overall system performance, as shown in Table 2, follows a similar pattern to the performance of the information extraction components. Generally, the overall system performance is lower than the performance of individual components. This discrepancy is due to the accumulation of errors at each component stage, which aggregate throughout the data processing pipeline.

The system evaluation results point to a higher performance in the precision metric for variation W compared to the evaluation of the information extraction components. This improvement is due to the evaluation method accommodating the Character Error Rate (CER), which omits predictions labeled O, thus reducing the number of false positives and enhancing the precision.

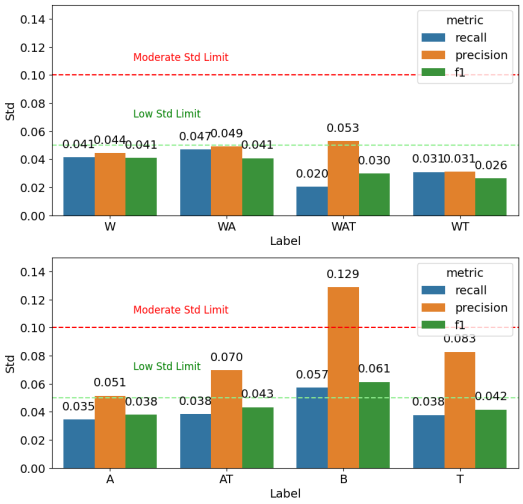

Figure 5: Standard deviation of the result per label

6 Conclusion

We successfully developed a de-identification system for Indonesian speech comprised of speech recognition, information extraction, and masking components. Using a dataset without missing labels, the system achieved a recall of 69.2%, precision of 79.4%, and an F1 score of 73.9%. When tested on a dataset with 60% labeled data, the performance showed a recall of 36.2%, precision of 63.0%, and an F1 score of 45.6%. However, on a dataset with only 20% labeled data, the system's performance dropped significantly, achieving a recall of 0.00%, precision of 0.20%, and an F1 score of 0.00%. The system's performance decreased with the percentage of labeled data, showing that the system gained its knowledge from the given data.

The addition of various techniques into the baseline model resulted in improved performance. Specifically, the combination of domain-specific transcription data, dataset augmentation, and weakly supervised learning methods yielded a significant performance boost. The de-identification system incorporating all techniques achieved a recall of 75.2%, precision of 75.6%, and an F1 score of 75.3% on perfect data; a recall of 69.4%, precision of 77.0%, and an F1 score of 72.9% on 60% labeled data; and a recall of 47.8%, precision of 70.0%, and an F1 score of 56.7% on 20% labeled data. These results indicate a significant improve-

ment over the baseline system.

7 Future Works

Future research directions for enhancing the de-identification system include exploring its scalability for larger datasets and complex scenarios, such as integration with tools like Hadoop or Spark. Additionally, adding diarization support is advised due to the common occurrence of speaker overlap in conversational speech data.

References

Amazon. Redacting or identifying personally identifiable information - amazon transcribe.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio.

Guillaume Baril, Patrick Cardinal, and Alessandro Lameiras Koerich. 2022. Named entity recognition for audio de-identification.

Blanche and Christopher Miret Kermorvant. 2021. Nerval: a python library for named-entity recognition evaluation on noisy texts.

Ido Cohn, Itay Laish, Genady Beryozkin, Gang Li, Izhak Shafran, Idan Szpektor, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2019. Audio de-identification - a new entity recognition task. pages 197–204. Association for Computational Linguistics.

William E. Cooper, Nancy Tye-Murray, and Stephen J. Eady. 1985. Acoustical cues to the reconstruction of missing words in speech perception. *Perception Psychophysics*, 38:30–40.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. pages 3861–3867. International Committee on Computational Linguistics.

Martin Flechl, Shou-Chun Yin, Junho Park, and Peter Skala. 2022. End-to-end speech recognition modeling from de-identified data.

Praveen Giridhara, Chinmaya Mishra, Reddy Venkataramana, Syed Bukhari, and Andreas Dengel. 2019. A study of various text augmentation techniques for relation classification in free text. pages 360–367. SCITEPRESS - Science and Technology Publications.

Evandro Gouvêa, Ali Dadgar, Shahab Jalalvand, Rathi Chengalvarayan, Badrinath Jayakumar, Ryan Price, Nicholas Ruiz, Jennifer McGovern, Srinivas Bangalore, and Ben Stern. 2023. Trustera: A live conversation redaction system. *Preprint*, arXiv:2303.09438.

Nicolas Hervé, Valentin Pelloin, Benoit Favre, Franck Dary, Antoine Laurent, Sylvain Meignier, and Laurent Besacier. 2022. Using asr-generated text for spoken language modeling. pages 17–25. Association for Computational Linguistics.

Micaela Kaplan. 2020. May i ask who's calling? named entity recognition on call center transcripts for privacy law compliance. pages 1–6. Association for Computational Linguistics.

Ayush Kumar, Mukuntha Narayanan Sundararaman, and Jithendra Vepa. 2021. What bert based language models learn in spoken transcripts: An empirical study.

Yizhi Liu, Fang Yu Lin, Mohammadreza Ebrahimi, Weifeng Li, and Hsinchun Chen. 2021. Automated pii extraction from social media for raising privacy awareness: A deep transfer learning approach. pages 1–6. IEEE.

Microsoft. What is the personally identifying information (pii) detection feature in azure ai language? - azure ai services.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

NAYAK, GAYATRI, and Swagatika. 2011. A survey on privacy preserving data mining: approaches and techniques. *International Journal of Engineering Science and Technology*, 3.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mluke: The power of entity representations in multilingual pretrained language models. pages 7316–7330. Association for Computational Linguistics.

Tetsuro Saeki, Takahiro Tamesue, Shizuma Yamaguchi, and Kazuya Sunada. 2004. Selection of meaningless steady noise for masking of speech. *Applied Acoustics*, 65:203–210.

Jennifer Williams, Junichi Yamagishi, Paul-Gauthier, Cassia Valentini-Botinhao, and Jean-François Bonastre. 2021. Revisiting speech content privacy. pages 42–46. ISCA.

David Wright and Charles Raab. 2014. Privacy principles, risks and harms. *International Review of Law, Computers Technology*, 28:277–298.

Lu Xu, Lidong Bing, and Wei Lu. 2023. Better sampling of negatives for distantly supervised named entity recognition.

A Dataset Creation

A.1 Generation Parameters

To generate the initial dataset, we used the API version of GPT-4 with these settings based on the results of our manual testing.

- Model name: gpt-4-1106-preview
- Temperature: 0.7
- Top P: 0.8

A.2 Prompt

System:

You are a system that creates natural and detailed speech transcripts in Indonesian tailored to specific contexts. Follow these rules:

1. Maintain a natural flow and adhere to a 400-word limit for each transcript.

2. Separate transcripts with triple newlines.

3. Clearly annotate all personal information within the transcripts in this format:

[Personal Information: {information}, Relation: {class}, Entity: {entity}]

- Relation classes include: name, address, date, datetime, location, birthplace, birthdate, phone number, email, professiontitle, professioncompany, educationlevel, educationplace, educationyear, banknumber, bankcvv, bankexpiry, and identification number.

- Entity refers to the person the information pertains to.

- Link even simple nicknames but avoid annotating the aspect itself (e.g., do not annotate "nickname").

- Reuse existing annotations for duplicate personal information.

User:

Create three distinct speech transcripts in Indonesian, each tailored to a specific context:

1. Job interview

2. Medical anamnesis

3. Bank call center

Incorporate fictional personal information naturally, such as names, addresses, dates, phone numbers, emails, professions, education details, locations, and financial or identification details.

A.3 Dataset Statistics

Parameter	Value
Dialogues	86
Utterances	912
PII count	Person's Name: 508 Location: 162 Date: 59 Email: 39 Profession: 106 Telephone number: 59 Bank Number (Number, CVV, Exp Date): 20 Identification Number (SSN, Healthcare, etc.): 13 Education Information: 16
Speaker	10 (5 male, 5 female)
Duration	6617 seconds
Sampling rate	16000 Hz
Dialogue topics	Casual Conversation: 30 Medical Anamnesis: 19 Job interviews: 19 Bank Call Center: 18

Table 4: Generated Data Statistics