Data-Centric Optimization of Enrollment Selection in Speaker Identification

Long-Quoc Le[1,2][a] and Minh-Nhut Ngo[1,2*][b]
[1]*Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam*
[2]*Vietnam National University, Ho Chi Minh City, Vietnam*

Keywords: Data-Centric, Speaker Identification, Enrollment Selection.

Abstract: In this paper, we introduce a novel method for optimizing enrollment selection in speaker identification systems, with a particular focus on low-resource languages. Unlike traditional approaches that rely on random enrollment samples, our method systematically analyzes pair-wise similarities between enrollment utterances to eliminate poor-quality samples often impacted by noise or adverse environments. By retaining only high-quality and representative utterances, we ensure a more robust speaker profile. This innovative approach, applied to the Vietnam-Celeb dataset using the state-of-the-art ECAPA-TDNN model, delivers substantial performance improvements. Our method boosts accuracy from 73.38% in bad scenarios to 93.62% and increases the F1-score from 72.91% to 95.48%, demonstrating the effectiveness of focusing on quality-driven enrollment selection even in low-resource contexts.

1 INTRODUCTION

Speech communication has become an increasingly popular interface in virtual assistants, especially with advancements in large language models that enable understanding of high-level knowledge. Speaker recognition has garnered significant attention as it enhances speech communication by providing added functionality and security. By enabling speaker recognition, virtual assistants and smart interaction systems can respond more naturally and customize interactions for specific users, improving the overall user experience (Mohd Hanifa et al., 2021). Additionally, speaker recognition strengthens security by preventing unauthorized users from executing critical commands.

Prior to the deep learning era, most speaker recognition systems relied on i-vector based models (Dehak et al., 2010), which utilized Mel-Frequency Cepstral Coefficients (MFCC) and universal background models (UBM) built with Gaussian Mixture Models (GMM). These i-vector approaches projected speaker information from a high-dimensional UBM space into a lower-dimensional speaker space. However, i-vector models suffered from limited performance due to their reliance on handcrafted features, which struggled to capture the complex variations in human voice characteristics, especially under challenging conditions.

However, the advent of deep learning has driven remarkable advancements in speaker recognition performance, with early deep neural embedding-based models such as x-vectors (Snyder et al., 2018) marking a notable leap from traditional i-vector approaches. The x-vector model pioneered the use of deep neural networks for generating speaker embeddings, laying the groundwork for later innovations. Building on the Time Delay Neural Network (TDNN) (Peddinti et al., 2015) which is a frame-level feature extractor, the ECAPA-TDNN architecture (Dawalatabad et al., 2021) introduced refined feature extraction layers and when combined with the ArcFace loss function (Deng et al., 2019), achieved an Equal Error Rate (EER) of 0.87% on the VoxCeleb1 test set (Zeinali et al., 2019), representing a considerable enhancement in accuracy. In addition, ResNet architectures (He et al., 2016), adapted specifically for speaker recognition, have demonstrated remarkable performance by utilizing their powerful feature extraction capabilities. Various ResNet configurations have yielded impressive outcomes. For instance, the Thin ResNet-34 model (Chung et al., 2019), paired

[a] https://orcid.org/0009-0007-9838-2260
[b] https://orcid.org/0009-0001-1185-2394
* Corresponding author.

Le, L.-Q. and Ngo, M.-N.
Data-Centric Optimization of Enrollment Selection in Speaker Identification.
DOI: 10.5220/0013256800003905
In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 344-351
ISBN: 978-989-758-730-6; ISSN: 2184-4313

with the Angular Prototypical loss function (Chung et al., 2020), achieved an EER of 2.21% on the VoxCeleb1 test set. Further pushing the boundaries, RawNet3 (Jung et al., 2022) utilized raw audio signals directly through 1D convolutional layers, attaining an EER of 0.89% on VoxCeleb1. These models underscore the remarkable evolution and effectiveness of end-to-end deep learning architectures in speaker recognition.

Although deep learning models have advanced speaker recognition, significant challenges persist. While speaker recognition systems face the common challenge of capturing the natural diversity in human vocal characteristics—including variations in tone, pitch, and speaking style within the same individual—these difficulties. This challenge becomes especially pronounced in low-resource languages, where data scarcity limits the ability to create comprehensive speaker representations.

To address this issue, we propose a data-centric approach that emphasizes selecting high-quality and representative enrollment samples, especially tailored for low-resource languages. Our method prioritizes samples that accurately capture the speaker's dominant vocal characteristics and eliminates those affected by noise or distortions. This strategy is particularly crucial in low-resource settings, where limited data prevents comprehensive coverage of all vocal traits. Instead, our approach ensures a well-defined speaker profile by selecting only the most consistent and reliable vocal traits, which in turn, enhances the system's robustness and reliability for speaker identification tasks.

In this work, we introduce an empirical method to optimize enrollment sample selection for speaker identification, aimed at maximizing the effectiveness of limited data in low-resource settings. Our experiments demonstrate that selecting high-quality enrollment samples leads to significant performance improvements in speaker identification. The proposed iterative selection method identifies and removes poor-quality samples, resulting in a high-confidence enrollment set. By configuring the total number of enrollment samples, our approach also allows for customization based on application needs.

The rest of this paper is organized as follows. Section 2 discusses the existing literature and related work in models of speaker recognition. Section 3 describes the proposed method of optimizing selection of enrolling utterances in speaker identification, illustrating our idea, explaining the underlying principles, and detailing the key steps. Section 4 presents the experimental setup and configurations, the results of evaluations, and analyzes the results. Section 5 gives some concluding remarks and suggests some directions for future work.

2 RELATED WORK

Before the era of deep learning, speaker recognition models predominantly used the i-vector method (Dehak et al., 2010), which extracted speaker features based on Mel-Frequency Cepstral Coefficients (MFCC) and universal background models (UBM). In this approach, a Gaussian Mixture Model (GMM) was employed to map the high-dimensional UBM space to a lower-dimensional i-vector space, providing a compact representation of each speaker. Despite its utility, the i-vector approach was limited by its vulnerability to variations in speaking style, background noise, and other environmental conditions. These limitations reduced its robustness in real-world applications, as it struggled to consistently differentiate speakers across diverse and unpredictable acoustic environments.

The shift to deep learning introduced a new era of speaker recognition architectures, beginning with x-vectors (Snyder et al., 2018), a deep neural network (DNN) embedding model designed for text-independent speaker recognition. The x-vector model consists of three main components: a Time Delay Neural Network (TDNN) (Peddinti et al., 2015) for frame-level feature extraction from MFCC inputs, a statistics pooling layer that aggregates segment-level statistics, and a soft-max output layer trained with cross-entropy loss to classify speakers. This architecture laid the foundation for further advancements in speaker embedding. Building on this structure, the ECAPA-TDNN architecture (Dawalatabad et al., 2021) introduced refined feature extraction layers, further enhancing accuracy, particularly when paired with the ArcFace loss function (Deng et al., 2019), achieving an EER of 0.87% on the VoxCeleb1 test set (Zeinali et al., 2019).

ResNet (He et al., 2016), initially popularized in computer vision, has also become a prominent architecture in speaker recognition. Unlike its use in image tasks, ResNet in audio processing is customized to work with speech spectrograms, capturing speaker-specific patterns effectively. Several ResNet configurations, such as Thin ResNet-34 (Chung et al., 2019) combined with the Angular Prototypical loss function (Chung et al., 2020), have demonstrated impressive performance, with Thin ResNet-34 achieving an EER of 2.21% on the VoxCeleb1 test set.

Further advancements in deep learning for speaker recognition include the use of raw audio data, exem-

plified by RawNet3 (Jung et al., 2022). This model employs 1D convolutional layers to directly process raw audio signals, eliminating the need for spectrogram conversion and enabling the capture of more granular acoustic features. Combined with the ArcFace loss, RawNet3 achieved an EER of 0.89% on the VoxCeleb1 test set, underscoring the potential of end-to-end deep learning models in speaker recognition.

Several techniques of adaptation and normalization were proposed to deal with limited enrollment data (Kimball et al., 1997) and training, enrollment and test mismatching (Mak et al., 2006), (Glembek et al., 2014), (Wang et al., 2018), (Aronowitz, 2014), (Li et al., 2022). Some approaches have been proposed to deal with enrollment of utterances for later uses in verification or identification processes. (Li et al., 2024) proposed an augmentation technique that applies to enrolling utterances which results in consistent performance improvement. (Mingote et al., 2020) directly trained speaker enrollment models for each speaker by leveraging an embedding dictionary stored during the training phase in the last layer of a deep neural network. The verification scores are obtained directly from the speaker enrollment models without using another comparison metric.

Our method prioritizes the selection of high-quality samples that accurately reflect the speaker's dominant vocal traits, while filtering out those affected by noise or inconsistencies. By capturing the most consistent and reliable vocal characteristics, we create an effective and representative speaker profile for recognition tasks.

3 PROPOSED METHOD

Figure 1 illustrates the 3D representation of embeddings for seven randomly selected speakers from the Vietnam-Celeb (Thanh et al., 2023) dataset, after dimensionality reduction through the Principal component analysis (PCA) (Kurita, 2019) algorithm. These embeddings were generated using the ECAPA-TDNN model (Dawalatabad et al., 2021), fine-tuned on the Vietnam-Celeb (Thanh et al., 2023) data. While there is clear separation among most speakers, some points remain indistinct, reflecting cases where embeddings overlap. Further auditory analysis reveals that these ambiguous samples are often of lower quality, likely due to noise or variations in voice tone and pronunciation that deviate from the speaker's usual patterns. These represent outlier embeddings that could benefit from filtering in the data preprocessing phase.

Our proposed approach, which emphasizes the careful selection of high-quality samples, proves particularly advantageous in low-resource languages, where it enables performance levels comparable to systems using extensive datasets. This efficiency makes the method highly suited to low-resource contexts, where obtaining large datasets can be challenging. While this method requires tonal consistency during enrollment, potentially leading to authentication rejections when a user's voice tone varies significantly, this trade-off enhances system robustness—a compromise acceptable in low-resource language contexts, where consistent enrollment data significantly improves performance without requiring large datasets.

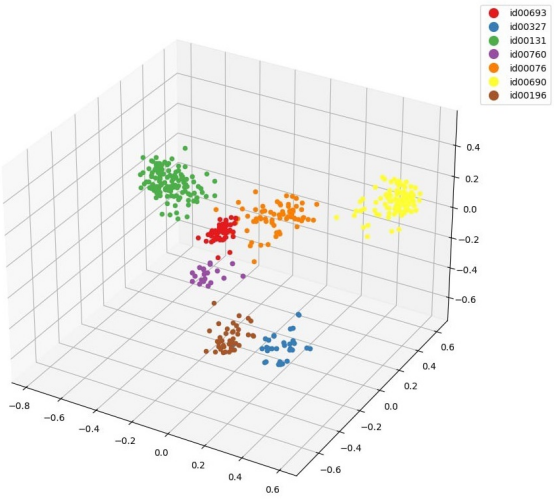

Figure 1: Representation of Speaker Utterances.

Based on these observations, our method relies on two key assumptions:

- Higher similarity among samples within the enrollment set likely indicates convergence toward the speaker's standard voice under optimal conditions, minimizing variability in vocal traits.
- Increasing high-quality samples in the enrollment set enhances system robustness, allowing a smaller set of samples in low-resource languages to achieve comparable robustness to larger datasets in high-resource languages.

Initially, we hypothesized that high-quality samples alone would be sufficient. However, our findings showed that a more nuanced approach is necessary. Thus, we propose a data-centric solution that involves optimizing sample selection based on a fine-tuned positive threshold.

The positive threshold optimizes Equal Error Rate (EER) during training, ensuring that selected enrollment samples have a pairwise similarity meeting or exceeding this threshold. This approach creates a

consistent and representative enrollment set, reducing noise and environmental variability.

In our experiments, we also observed diminishing returns with increasing sample quantity, revealing that more data does not necessarily improve performance. We therefore propose selecting an optimal sample quantity at which performance stabilizes, achieving a balance between complexity and accuracy. In summary, our solution comprises three key steps:

- Define a positive threshold during training to ensure high-quality, consistent samples.
- Ensure each sample pair in the enrollment set meets or exceeds the positive threshold, forming a homogeneous and representative set.
- Determine the optimal sample quantity to prevent redundancy or noise, balancing performance and complexity.

Figure 2 presents an example enrollment set containing four utterances. When adding a new sample (5), we calculate its similarity to each existing sample. Sample (5) is closely aligned with samples (1) and (2), indicated by solid connections, while sample (2) is also close to (1). These samples form a high-similarity cluster, from which we can select (1), (2), and (5) as candidates. However, sample (5) lacks similarity with samples (3) and (4), represented by dashed lines, and thus does not meet the positive threshold for a cohesive enrollment set.

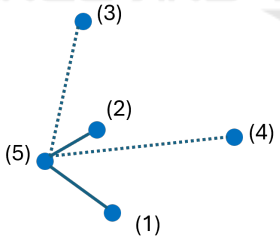

Figure 2: Example of Enrollment Selection Based on Pairwise Similarity.

These steps aim to create a robust enrollment process that represents dominant vocal traits, enhancing system performance while resisting variations in speech style and environmental factors.

4 EXPERIMENTS

4.1 Dataset and Experimental Setup

In this study, we used the **Vietnam-Celeb** dataset (Thanh et al., 2023), which consists of voice samples collected from 1,000 distinct speakers. This dataset was split into three subsets: 900 speakers for training, 50 speakers for validation, and 50 speakers for testing. The primary goal of our experiments was to evaluate the effectiveness of our proposed method of enrollment selection with speaker recognition model **ECAPA-TDNN** (Dawalatabad et al., 2021) which was found efficient for Vietnamese (Ngo and Le, 2024).

We chose this dataset because Vietnamese is a low-resource language, making it an ideal choice to rigorously evaluate our algorithm's effectiveness under challenging conditions. By using Vietnamese, we can accurately assess the algorithm's performance in optimizing enrollment selection and speaker recognition accuracy when data is inherently limited.

4.1.1 Data Quality Control and Mislabeled Samples

During the initial data inspection, we discovered that a significant number of samples in the test set were mislabeled. This issue posed a potential threat to the reliability of the experimental results, as inaccurate labeling could lead to biased evaluation metrics. Upon further investigation, we found that out of 7,351 utterances in the test set, 182 were mislabeled, which accounts for **2.48%** of the total test data.

To ensure the validity of the evaluation, we manually re-labeled the entire test set, correcting the mislabeling errors. The re-labeling process was critical for the integrity of the results, as it directly impacted the accuracy of performance metrics like Equal Error Rate (EER). However, due to resource constraints, we did not re-label the training and validation sets. We believe that the impact of mislabeled data in the training and validation sets is minimal, given that only 2.48% of the test set samples were mislabeled. This relatively small percentage of mislabels is unlikely to significantly influence the model's ability to learn. The ECAPA-TDNN model (Dawalatabad et al., 2021) is robust and can effectively generalize, allowing it to focus on key speaker characteristics and ignore occasional mislabeled samples. Furthermore, the model's resilience to mislabeled data is enhanced when working with larger datasets, as the model can still capture meaningful patterns from the majority of correctly labeled data.

4.1.2 Embedding Model and Similarity Measurement

For embedding extraction, we used the ECAPA-TDNN model (Dawalatabad et al., 2021), a state-of-the-art architecture widely recognized for its supe-

rior performance in speaker recognition tasks. This model was specifically designed to effectively capture speaker-specific features and handle variations in speech signals, making it ideal for this task. The ECAPA-TDNN model (Dawalatabad et al., 2021) was chosen not only for its strong performance but also because it was used in the original Vietnam-Celeb paper (Thanh et al., 2023), ensuring consistency in comparison and enabling a fair evaluation of results across different studies.

The ECAPA-TDNN model (Dawalatabad et al., 2021) was trained using the 79,789 training samples, which consist of voice data from 900 distinct speakers. The model's architecture was fine-tuned during the training process to optimize its ability to differentiate between speakers based on their unique vocal characteristics.

To measure similarity between speaker embeddings, we employed *cosine similarity*. Cosine similarity calculates the cosine of the angle between two embedding vectors, producing a score in the range $[-1, 1]$, where values closer to one indicate higher similarity. This metric allows for a precise comparison of embeddings, as it quantifies the alignment of speaker-specific features in the enrollment and test samples, contributing to accurate speaker identification.

4.1.3 Threshold Optimization and Evaluation

After calculating the similarity between two embeddings, a positive threshold is applied to determine whether they belong to the same person. If the similarity score between two embeddings meets or exceeds this threshold, they are classified as belonging to the same individual; otherwise, they are considered as coming from different individuals. Setting an appropriate positive threshold is essential for balancing the *False Acceptance Rate (FAR)*, which measures the rate of mistakenly accepting embeddings from different individuals and the *False Rejection Rate (FRR)*, which represents the rate of incorrectly rejecting embeddings from the same individual.

To optimize this threshold, we fine-tuned it using a validation set of 50 speakers, aiming to minimize the *Equal Error Rate (EER)* — the point where *FAR* and *FRR* are equal. By iteratively adjusting the threshold, we identified the value that balances these error rates, thus reducing both types of errors and improving the system's overall accuracy. This process ensures that the model is neither overly lenient (which would increase *FAR*) nor too strict (which would increase *FRR*), resulting in a robust and reliable classification.

Once the optimal positive threshold was established, we evaluated the model's performance on a carefully re-labeled test set to ensure data accuracy. This final evaluation provided a comprehensive assessment of the model's capability to correctly classify identities under realistic conditions. The model's performance was then compared with other state-of-the-art methods, demonstrating the effectiveness of our approach in terms of both accuracy and error rates.

4.2 Experimental Configurations

Our experiments were divided into three main configurations, each designed to assess the impact of enrollment data quality and consistency on speaker identification performance:

- **Bad Case:** In this configuration, enrollment samples have low pairwise similarity, with values falling below a predefined positive threshold. This threshold represents the minimum similarity score needed to consider samples as consistent representations of the speaker's primary vocal characteristics. The Bad Case simulates a worst-case scenario, where enrollment samples are likely to be impacted by noise, distortions, or inconsistent speaker tones. Such poor-quality data introduces variability and can compromise the reliability of the speaker profile, leading to decreased identification accuracy and increased error rates.
- **Random Case:** In the Random Case, enrollment samples are selected without any consideration for pairwise similarity, meaning samples are chosen randomly from the dataset. This setup represents a typical real-world scenario where no specific enrollment strategy is applied, serving as a baseline for comparison. Some samples may, by chance, meet the positive threshold, while others may fall below it, creating inconsistencies in data quality. The mixed-quality dataset in this configuration reflects common, unsupervised enrollment conditions and helps gauge how our method compares against a standard, uncontrolled selection process.
- **Optimal Case:** In the Optimal Case, only samples that meet or exceed the positive threshold are included in the enrollment set. This careful selection process ensures a high level of quality and consistency, producing a dataset that strongly represents the speaker's dominant vocal characteristics. By filtering out samples that do not meet the similarity criteria, this configuration aims to provide the most reliable speaker profile, with minimal noise or variability. As a result, the Opti-

mal Case is expected to deliver the highest performance in speaker identification, highlighting the benefits of a controlled, similarity-based enrollment strategy.

An illustration of the algorithm setup for all configurations can be found in Figure 3. For each configuration, we systematically varied the number of enrollment samples from one to 10 to observe its effect on performance metrics. This approach allowed us to analyze how the quality and quantity of enrollment data interact to influence accuracy and to determine the optimal number of samples required to achieve high identification accuracy with a minimal enrollment set size.

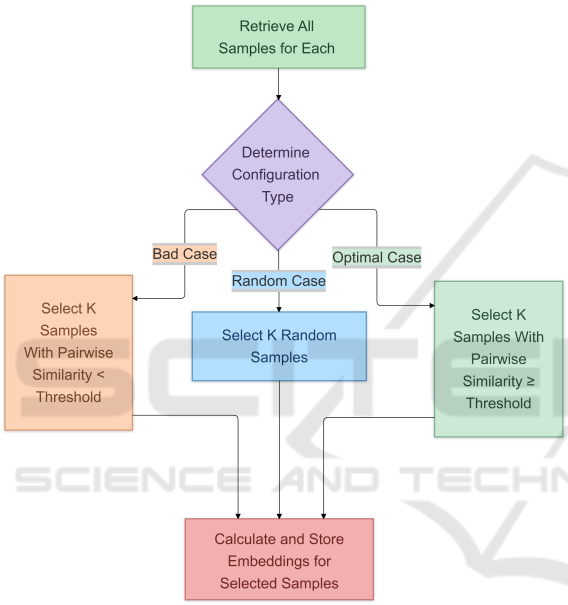

Figure 3: Illustration of the algorithm setup across configurations.

The experimental results, summarized in Table 1, Table 2, and Table 3, demonstrate the substantial impact of enrollment data quality on speaker identification performance across three configurations: Bad Case, Random Case, and Optimal Case. For each configuration, we evaluated the model's performance by analyzing Accuracy, Precision, Recall, and F1-score as the number of enrollment samples increased from one to 10. We randomly selected the samples for each sample size with 30 iterations and calculated mean and standard deviation (std) of the scores. A clear trend is observed: the **Optimal Case** yields the highest performance across all metrics, substantiating the effectiveness of our proposed solution.

4.3 Results and Analysis

In the Bad Case configuration, where enrollment samples fall below the positive threshold, model performance is notably reduced. Accuracy begins at 68.73% with one sample, increasing only marginally to 73.38% with 10 samples, while high standard deviations across metrics reveal unstable and inconsistent outcomes. These results underscore the adverse effects of low-quality and inconsistent samples, which introduce variability and noise, ultimately degrading identification accuracy.

Table 1: Performance Metrics for Different Enrollment Sample Sizes in Bad Case.

Size	Accuracy		Precision		Recall		F1-score	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	0.687	0.016	0.868	0.028	0.646	0.019	0.694	0.018
2	0.700	0.020	0.887	0.034	0.658	0.022	0.705	0.021
3	0.708	0.023	0.899	0.035	0.665	0.024	0.712	0.023
4	0.715	0.024	0.900	0.033	0.673	0.026	0.719	0.024
5	0.720	0.024	0.896	0.033	0.677	0.026	0.722	0.023
6	0.724	0.024	0.892	0.033	0.681	0.025	0.724	0.022
7	0.727	0.023	0.887	0.033	0.684	0.024	0.726	0.021
8	0.730	0.023	0.881	0.035	0.687	0.024	0.727	0.020
9	0.732	0.023	0.877	0.036	0.689	0.024	0.729	0.019
10	0.734	0.022	0.873	0.036	0.691	0.023	0.729	0.019

The Random Case configuration, representing typical real-world conditions without specific enrollment criteria, achieves moderate improvements over the Bad Case. Accuracy increases from 84.88% with one sample to 89.38% with 10 samples; however, these values remain consistently lower than those in the Optimal Case. F1-score and other metrics similarly improve as sample quantity increases, but relatively high standard deviations indicate limited robustness compared to the Optimal Case configuration. This baseline scenario illustrates that random sample selection, while beneficial, lacks the consistency and quality necessary for optimal performance.

Table 2: Performance Metrics for Different Enrollment Sample Sizes in Random Case.

Size	Accuracy		Precision		Recall		F1-score	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	0.849	0.022	0.943	0.019	0.840	0.026	0.868	0.026
2	0.863	0.024	0.954	0.019	0.855	0.027	0.882	0.026
3	0.872	0.024	0.959	0.018	0.864	0.026	0.891	0.026
4	0.878	0.024	0.962	0.017	0.871	0.026	0.897	0.025
5	0.882	0.024	0.965	0.016	0.875	0.026	0.900	0.024
6	0.886	0.023	0.966	0.016	0.878	0.025	0.903	0.024
7	0.888	0.023	0.967	0.015	0.881	0.024	0.906	0.023
8	0.890	0.022	0.969	0.015	0.883	0.024	0.908	0.022
9	0.892	0.022	0.969	0.014	0.885	0.024	0.909	0.022
10	0.894	0.021	0.970	0.014	0.887	0.023	0.911	0.022

The Optimal Case configuration, where enrollment samples meet or exceed the positive threshold, consistently demonstrates the highest results across

all metrics. With only one sample, the model attains an Accuracy of 92.05%, which further improves to 93.62% with 10 samples. The F1-score shows a similar progression, increasing from 94.15% to 95.48%, with low standard deviations reflecting high reliability and robustness. These findings confirm that prioritizing high pairwise similarity in enrollment data produces a stable and accurate speaker profile, effectively mitigating the impact of noise and variability.

Table 3: Performance Metrics for Different Enrollment Sample Sizes in Optimal Case.

Size	Accuracy		Precision		Recall		F1-score	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	0.921	0.008	0.966	0.005	0.927	0.008	0.942	0.006
2	0.926	0.009	0.970	0.006	0.932	0.008	0.947	0.007
3	0.930	0.009	0.973	0.006	0.936	0.009	0.950	0.008
4	0.932	0.009	0.975	0.005	0.937	0.008	0.951	0.007
5	0.933	0.008	0.974	0.005	0.939	0.008	0.952	0.007
6	0.934	0.008	0.975	0.005	0.940	0.008	0.953	0.007
7	0.935	0.009	0.976	0.005	0.940	0.007	0.954	0.006
8	0.935	0.008	0.976	0.005	0.941	0.007	0.954	0.006
9	0.936	0.007	0.976	0.005	0.941	0.007	0.955	0.006
10	0.936	0.007	0.977	0.005	0.941	0.006	0.955	0.006

Across all the configurations, increasing the number of enrollment samples generally enhances performance; however, the rate of improvement diminishes beyond 5 samples, particularly in the Optimal Case. This suggests that, above a certain threshold, additional samples contribute minimally to performance, especially when data quality is already high, confirming that sample quality is more critical than quantity in enrollment selection.

Although a large number of utterances were used for evaluations in our experiments, in practice, we can iteratively choose enrolling utterances that maximize pairwise similarity instead of random utterances. This process can be repeated over several iterations until we get a suitable number of qualified utterances, e.g., five utterances.

4.4 Summary of Experimental Findings

Our experimental findings underscore the critical role of data quality in enrollment selection, with clear evidence that high pairwise similarity among enrollment samples significantly boosts speaker identification performance. The **Optimal Case**, which emphasizes selecting samples that meet a predefined positive similarity threshold, consistently outperformed the Bad Case and Random Case. Specifically, with 5 enrollment samples, the Optimal Case achieved an F1-score of 95.21%, compared to 72.20% in the Bad Case and 90.02% in the Random Case. These results demonstrate that carefully curated, high-quality samples are essential for creating robust speaker profiles.

In addition to quality, our findings indicate that increasing the number of enrollment samples can improve performance, but only up to a certain point. The rate of improvement diminishes beyond 5 samples, particularly in the Optimal Case, suggesting that a modest number of high-quality samples is sufficient for reliable identification. This observation confirms that sample quality is more critical than quantity, as adding more samples beyond a certain threshold yields minimal benefits.

Based on these insights, the most effective approach for enrollment selection is to maintain a modest sample size of approximately **5 high-quality** utterances per user, with each sample meeting the positive threshold. This solution optimizes the balance between performance and resource efficiency, delivering a robust speaker identification system that maximizes accuracy with minimal data. Such a strategy is especially valuable in low-resource scenarios, where data quality is prioritized over quantity to achieve optimal results.

5 CONCLUSION

In this paper, we presented a data-centric approach for optimizing the enrollment selection process in speaker identification systems, with a particular focus on low-resource languages such as Vietnamese. Our proposed method emphasizes the importance of selecting high-quality and representative samples during the enrollment phase to mitigate the challenges posed by variability in voice characteristics and environmental factors. This approach is especially effective in low-resource settings, where the availability of large and diverse datasets is limited, making it difficult to capture all aspects of a speaker's vocal characters.

Through a series of experiments, we demonstrated that by filtering out low-quality and inconsistent samples, we can create robust speaker profiles that enhance the accuracy and reliability of the speaker identification system. Our method consistently outperformed random and bad enrollment selection strategies, showing significant improvements in key performance metrics such as accuracy and F1-score. The results validate the effectiveness of leveraging a small set of high-quality samples to achieve comparable performance to systems that require much larger datasets in high-resource scenarios.

Additionally, we acknowledged a trade-off in our approach, where the system requires consistency in the user's voice tone between enrollment and authentication. While this constraint may limit flexibility, it

significantly enhances the robustness of the system, which is particularly important in low-resource languages.

Overall, our findings highlight the potential of a data-centric approach to overcome the challenges inherent in low-resource speaker recognition, paving the way for more efficient and effective systems in this domain. Future work may explore further optimizations in the enrollment process and investigate how additional techniques, such as data augmentation, can be applied to further improve performance in low-resource settings.

REFERENCES

Aronowitz, H. (2014). Inter dataset variability compensation for speaker recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4002–4006.

Chung, J. S., Huh, J., and Mun, S. (2019). Delving into voxceleb: environment invariant speaker recognition. *arXiv preprint arXiv:1910.11238*.

Chung, J. S., Huh, J., Mun, S., Lee, M., Heo, H. S., Choe, S., Ham, C., Jung, S., Lee, B.-J., and Han, I. (2020). In defence of metric learning for speaker recognition. *arXiv preprint arXiv:2003.11982*.

Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., and Na, H. (2021). Ecapa-tdnn embeddings for speaker diarization. *arXiv preprint arXiv:2104.01466*.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Glembek, O., Ma, J., Matějka, P., Zhang, B., Plchot, O., Bürget, L., and Matsoukas, S. (2014). Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4032–4036.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jung, J.-w., Kim, Y. J., Heo, H.-S., Lee, B.-J., Kwon, Y., and Chung, J. S. (2022). Pushing the limits of raw waveform speaker recognition. *arXiv preprint arXiv:2203.08488*.

Kimball, O., Schmidt, M., Gish, H., and Waterman, J. (1997). Speaker verification with limited enrollment data. In *Eurospeech*, pages 967–970.

Kurita, T. (2019). Principal component analysis (pca). *Computer vision: a reference guide*, pages 1–4.

Li, J., Zhang, K., Wang, S., Li, H., Mak, M.-W., and Lee, K. A. (2024). On the effectiveness of enrollment speech augmentation for target speaker extraction.

Li, L., Wang, D., Kang, J., Wang, R., Wu, J., Gao, Z., and Chen, X. (2022). A principle solution for enroll-test mismatch in speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:443–455.

Mak, M.-W., Hsiao, R., and Mak, B. (2006). A comparison of various adaptation methods for speaker verification with limited enrollment data. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I.

Mingote, V., Miguel, A., Giménez, A. O., and Lleida, E. (2020). Training speaker enrollment models by network optimization. In *INTERSPEECH*, pages 3810–3814.

Mohd Hanifa, R., Isa, K., and Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, 90:107005.

Ngo, M.-N. and Le, L.-Q. (2024). Evaluation of command and speaker recognition on vietnamese voice dataset to enhance security. In *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech*, pages 3214–3218.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE.

Thanh, P. V., Hoa, N. X. T., Vu, H. L., and Trang, N. T. T. (2023). Vietnam-celeb: a large-scale dataset for vietnamese speaker recognition.

Wang, Q., Rao, W., Sun, S., Xie, L., Chng, E. S., and Li, H. (2018). Unsupervised domain adaptation via domain adversarial training for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4889–4893.

Zeinali, H., Wang, S., Silnova, A., Matějka, P., and Plchot, O. (2019). But system description to voxceleb speaker recognition challenge 2019. *arXiv preprint arXiv:1910.12592*.