IMPROVING PERFORMANCE OF VIETNAMESE SPEAKER RECOGNITION USING TRANSFER LEARNING AND ENSEMBLE EMBEDDING

Tan Hoang Ho[1], *Cao Truong Tran*[1,*]

Abstract

Speaker recognition technology is crucial for identifying or verifying individuals based on their unique vocal characteristics, such as pitch, tone, and speaking style. This technology is widely used to enhance security, improve customer service, support law enforcement, and personalize interactions with smart devices. In recent years, thanks to the application of deep learning techniques, speaker recognition has made significant progress. However, Vietnamese speaker recognition still faces many challenges. This paper presents new strategies that combine transfer learning and ensemble learning to improve the accuracy of Vietnamese speaker recognition. Experimental results on Vietnamese datasets show significant improvements in recognition accuracy. These findings highlight the potential of tailored approaches to advance speaker recognition technology for Vietnamese speakers and expand its applications in this field.

Index terms

Speaker recognition; speaker verification; speaker identification; transfer learning; representation learning.

1. Introduction

Natural speech not only conveys the semantic information the speaker intends to communicate (which can be recorded in written form), but also includes information about the speaker's emotional state and unique voice characteristics. These unique characteristics are divided into two levels: high-level and low-level. High-level information includes elements such as dialect, context, and speaking style, while low-level information consists of features like formants, formant bandwidths, pitch, and duration. These characteristics are not completely fixed from the time a person

[1]Institute of Information and Communication Technology, Le Quy Don Technical University
[*]Corresponding author, email: truongct@lqdtu.edu.vn DOI: 10.56651/lqdtu.jst.v13.n02.926.ict

learns to speak until they are old, but they remain relatively stable throughout a person's life. Once a person reaches adulthood, their speech habits and distinctive voice features become well-established and highly stable. This scientific basis underlies the development of speaker recognition systems [1]–[3].

Speaker recognition is a technology that identifies or verifies a person based on their voice. It does this by analyzing vocal characteristics such as pitch, tone, and speaking style to distinguish between individuals. Its applications are vast. These include enhancing security systems through voice-based authentication for accessing secure areas, devices, or accounts. It also improves customer service by verifying the identity of callers in call centers, thus streamlining processes and enhancing user experience. Speaker recognition assists law enforcement agencies in identifying suspects or verifying identities in criminal investigations. It enables personalized user experiences in smart home devices, such as virtual assistants that recognize individual users and tailor responses accordingly. Additionally, it enhances the security and convenience of telecommunications services by integrating voice recognition for authentication and authorization purposes [1]–[4].

Speaker recognition methods are typically divided into two groups: traditional machine learning-based methods (pre-deep learning) and deep learning-based methods. Traditional machine learning methods involve feature extraction techniques and classifiers such as Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and Support Vector Machines (SVMs). These methods rely on manually crafted features and statistical models to distinguish between speakers [5]–[8]. On the other hand, deep learning-based methods utilize neural networks, particularly deep neural networks (DNNs) and convolutional neural networks (CNNs), to automatically learn and extract features from raw audio data. These methods have shown significant improvements in accuracy and robustness, making them the current state-of-the-art in speaker recognition technology [9]–[15].

Embedding learning using deep learning for speaker recognition involves training neural networks to map voice samples into a high-dimensional space where similar voices are closer together [12]. Using architectures like CNNs and loss functions such as triplet loss, the network produces fixed-length embeddings that capture each speaker's unique voice characteristics. These embeddings enable tasks like speaker verification and identification with high accuracy and robustness, even in noisy environments. Advantages include better performance, handling large datasets, and less need for manual feature engineering, making it a powerful and flexible solution for modern speaker recognition systems [12], [13], [16], [17].

Vietnamese speaker recognition has also been a subject of research interest. Among the methods for Vietnamese speaker recognition, embedding learning has been proposed and proven to be better than traditional machine learning methods and other deep learning methods [15], [18]–[22]. However, it has not performed as well as embedding learning methods for other common languages like English. This is mainly due to the limited amount of available data and the less complex neural network

architectures used in these models. Consequently, the models have not achieved the same level of accuracy and robustness in recognizing Vietnamese speakers compared to their performance with languages that have more abundant data and more sophisticated neural network structures.

In this paper, we propose enhancements to embedding learning for Vietnamese speaker recognition. Specifically, we advocate for the use of transfer learning by leveraging pre-trained models trained on large-scale English datasets, followed by fine-tuning on Vietnamese datasets. Additionally, we introduce ensemble embedding techniques for speaker recognition, moving beyond pairwise comparisons to incorporate multiple comparisons simultaneously. Experimental results on Vietnamese datasets demonstrate that our proposed method significantly improves the accuracy of Vietnamese speaker recognition.

2. Related work

2.1. Speaker recognition

Speaker recognition methods are typically divided into two groups: traditional machine learning-based speaker recognition (pre-deep learning) and deep learning-based speaker recognition [1], [3]. Traditional machine learning-based speaker recognition methods consist of two main components: speaker feature extraction and speaker model building. The purpose of the speaker feature extraction step is to transform the raw speech signal into parameters called features, which serve the recognition process. After feature extraction, each speaker provides a set of feature vectors, usually comprising a large number of vectors. During the training phase, the goal is to develop a speaker classification model. This process essentially involves describing the set of feature vectors in the database in a way that reduces storage space while still accurately representing the speaker's unique characteristics [1].

Traditional speaker recognition methods rely on explicit feature extraction followed by probabilistic speaker modeling. These methods have achieved some success and have been applied in systems such as voice login and criminal voice identification. However, traditional feature extraction techniques, including widely used methods like Mel-Frequency Cepstral Coefficients (MFCC), are highly susceptible to noise. When noise is present, speaker recognition systems using these traditional methods often perform poorly, with recognition accuracy dropping to around 50%. This limitation makes these systems impractical for real-world applications, where natural speech is frequently noisy. Additionally, when constructing speaker models using statistical methods like GMMs, there is an assumption that the data follows a predefined distribution, such as a normal distribution. In practice, real-world data often do not conform to these assumed distributions, especially when the speech data contains noise. Consequently, speaker recognition models based on statistical methods also fail to achieve good performance, particularly under noisy conditions [5]–[8].

Deep learning methods for speaker recognition can be approached in two ways. The first approach uses deep learning techniques such as CNNs to build classifiers with speaker features that have been independently extracted, such as MFCC [2]. Although this approach has achieved significantly higher accuracy than traditional methods using statistical models, it still relies on explicit feature extraction like MFCC, which tends to perform poorly when the data is noisy. Moreover, these methods build classifiers with a predefined number of classes, requiring retraining of the models when new speakers need to be recognized. The second approach applies deep learning to learn speaker features directly from raw speech input or from minimally processed spectral data without extensive user intervention. Experiments have demonstrated that this approach allows for feature extraction that is not dependent on human biases, making the features more suitable for specific languages and less affected by noise. Additionally, this method does not construct classifiers with a predefined, finite number of speakers, allowing it to adapt when the number of speakers changes [9]–[15].

2.2. Vietnamese speaker recognition

Vietnamese speaker recognition has also been researched by domestic scientists. Similar to global research methods in speaker recognition, the methods studied by Vietnamese scientists can be divided into two groups: traditional machine learning-based speaker recognition and deep learning-based speaker recognition [21-28].

In studies [15], [18], MFCC is employed for speaker feature extraction, followed by vector quantization methods to construct speaker models. Study [18] introduces a new and robust method for text-independent speaker identification, leveraging discrete wavelet transform (DWT), mel-frequency discrete wavelet coefficients, wavelet-based sub-band weighting, and the likelihood combination Gaussian mixture model. Study in [15] introduces a new method for Vietnamese text-dependent speaker recognition. The system models each speaker using a Gaussian Mixture Model, while the phonemes in the keywords are represented by HMMs. By combining the prior and posterior probabilities for both keywords and speakers, the system effectively identifies speakers. The results demonstrate that the approach improves speaker identification performance, particularly when the speaker does not utter a sufficiently long phrase.

In recent years, deep learning research has increasingly focused on its application and effectiveness in Vietnamese speaker recognition [20], [21], [23]. Studies [12], [13] highlight that employing deep learning for extracting speaker features from raw speech data yields superior results compared to other speaker recognition methods. This approach also demonstrates robustness against noise and remains effective across varying speaker populations. Additionally, research [12] demonstrates that the effectiveness of deep learning in learning speaker features from raw data is influenced by the speaker's language. Researchers conducted two experimental setups: the first trained the model on Chinese data and tested it on English datasets, while the second

Table 1. Speaker recognition datasets statistics

Name	Language	Data Source	# of Spks	# of Utters	# of Hours
VoxCeleb1	English	YouTube	1,251	153,516	352
VoxCeleb2	English	YouTube	6,122	1,128,246	2,794
ZaloAI + Vietnam-Celeb	Vietnamese	YouTube	980	20,944	-

involved training the model first on Chinese data and then augmenting it with English data before testing on English datasets. Experimental findings showed a 10% increase in accuracy for the second configuration compared to the first, highlighting the significant impact of language characteristics on deep learning-based speaker feature extraction from raw data.

In study [21], transfer learning was utilized to learn embeddings for Vietnamese speaker recognition. English speaker recognition models were enhanced for Vietnamese. Experimental results showed that enhanced training could improve accuracy by up to 10%. However, the accuracy of these models still did not reach the levels achieved by models applied to data-rich languages like English. The primary reasons are the limited data and the insufficient depth of neural network architectures. This highlights the need for further research to improve the accuracy of these models.

2.3. Speaker recognition datasets

The effectiveness of speaker recognition systems is heavily dependent on the quality and size of the data sets used for training. Most of the available datasets are for widely spoken languages like English, such as VoxCeleb [24], VoxCeleb 2 [25]. For Vietnamese, a low-resource language, the situation is more challenging. The two primary Vietnamese datasets, ZaloAI and VLSP 2021, have demonstrated reasonable success in speaker verification tasks. However, these datasets are limited and have reliability issues due to insufficient data-building processing, lacking advanced techniques like pre-processing, which can lead to data mislabeling. Table 1 summarizes key statistics of popular speaker recognition datasets.

VoxCeleb1 [24] comprises over 100,000 utterances from 1,251 celebrities, sourced from YouTube videos. The datasets maintains a gender balance, with 55% of the speakers being male. It encompasses a diverse array of ethnicities, accents, professions, and age groups among the speakers. Furthermore, the dataset includes detailed information on the nationality and gender of each speaker, obtained from Wikipedia.

VoxCeleb2 [25] includes over 1 million utterances from over 6,000 celebrities sources from YouTube videos. It is gender balanced with 61% male speakers and features a diverse range of ethnicities, accents, professions, and ages. The videos are captured in various challenging visual and auditory environments, such as red carpets, outdoor stadiums, indoor studios, public speeches, professional multimedia excerpts, and hand-held device recordings. The audio is often degraded with background noise, laughter,

overlapping speech, and varying room acoustics.

The ZaloAI dataset [22] is used for the voice verification challenge in a ZaloAI competition 2020, the dataset is to build a speaker text-independent verification model for Vietnamese voices regardless of gender and regional accent of Vietnamese speakers using a diverse speech dataset. The dataset consists of 400 speakers short speech signals recorded in an uncontrolled environment, each utterance's length is from 0.8 s - 11 s.

VietCeleb Dataset [21] is a corpus of approximately 580 Vietnamese speakers of 16 kHz, which the minimum length of each utterances is 1.8 s, prepared by Le Quy Don Technical University. The data is derived from interview speech from the YouTube Channel, and has been carefully segmented and aligned. The size of dataset is about 1.6 GB, all utterances are normalised for speaker recognition tasks.

2.4. Transfer learning

According to [26], transfer learning is a machine learning technique where a model developed for a specific task is reused or fine-tuned for a different but related task. Instead of training a model from scratch, transfer learning allows a model to leverage the knowledge learned from a large dataset or a different domain to improve performance on a smaller or target dataset. This approach is particularly useful in scenarios where data for the target task is limited, as the pre-trained model already possesses a strong foundational understanding of features from its initial training.

The process of transfer learning typically involves two steps. First, a model is pre-trained on a large, generic dataset to learn general features such as shapes, structures, or patterns in the data. In the second step, the pre-trained model is fine-tuned on the target dataset, adjusting its weights to specialize in features relevant to the new task. For instance, in natural language processing, models like BERT or GPT are pre-trained on vast text corpora and then fine-tuned on domain-specific datasets for tasks like sentiment analysis, machine translation, or question-answering.

Transfer learning offers several benefits, such as reduced training time and improved model performance. By starting with a model that already has a foundation of learned features, training becomes faster and requires less computational power. Additionally, transfer learning helps achieve higher accuracy in the target task since the pre-trained model provides a strong starting point. This technique has been widely adopted in various fields, including computer vision, natural language processing, and speech recognition, demonstrating its effectiveness in machine learning applications.

2.5. Ensemble learning

According to [27], ensemble learning is a powerful machine learning technique that combines multiple models to improve overall performance. Instead of relying on a single model's predictions, ensemble methods aggregate the outputs of several models, aiming to reduce errors and enhance the robustness of the final decision. This approach leverages the diversity and strengths of individual models to achieve better

accuracy and generalization, making it particularly effective in complex tasks such as classification, regression, and anomaly detection. There are various strategies in ensemble learning, with two of the most common being bagging and boosting. Bagging, or Bootstrap Aggregating, involves training multiple models independently on different subsets of the training data, often created through sampling with replacement. The outputs of these models are then averaged (for regression) or voted upon (for classification) to produce the final result, as seen in algorithms like Random Forest. Boosting, on the other hand, focuses on training models sequentially, where each subsequent model attempts to correct the errors of the previous ones. This iterative approach, exemplified by algorithms like AdaBoost and Gradient Boosting, often yields highly accurate predictions. The advantages of ensemble learning are evident in its ability to improve accuracy and reduce overfitting. By combining multiple models, ensemble techniques mitigate the weaknesses of individual models, leading to more stable and reliable predictions. While ensemble methods can increase computational complexity and prediction time due to the involvement of multiple models, their benefits often outweigh these drawbacks, particularly in high-stakes applications such as finance, healthcare, and autonomous systems. As a result, ensemble learning has become a cornerstone of modern machine learning practices.

3. Proposed method

The proposed method involves two main phases: training process and application process as shown in Algorithm 1 and Algorithm 2, respectively.

3.1. Training process

The training process is designed to train a speaker embedding model using datasets of English and Vietnamese speakers, employing a specified training algorithm. Initially, the algorithm initializes the embedding model using the chosen "*TrainEmbeddingAlgorithm*". This step sets up the model's structure and parameters. Subsequently, the algorithm proceeds to train the initial embedding model on the English speaker dataset ("*EnglishDataset*"). This phase allows the model to learn distinguishing features specific to English speakers. Following this training, the model undergoes a fine-tuning process using the Vietnamese speaker dataset ("*VietnameseDataset*"). This fine-tuning adjusts the model to capture nuances unique to Vietnamese speakers. Once both training phases are completed, the algorithm outputs a "*TrainedModel*", representing the optimized speaker embedding model capable of encoding and differentiating between English and Vietnamese speakers based on learned linguistic features.

In the training process, the proposed method involves learning feature vectors from the English dataset (in step 2) before training on the Vietnamese dataset. This approach leverages the concept of transfer learning, which allows models to transfer knowledge gained from one dataset (English) to another (Vietnamese). By doing so, it reduces the

Algorithm 1: Training Process

Input:

- *English_Dataset*: English speaker dataset
- *Vietnamese_Dataset*: Vietnamese speaker dataset
- *Train_Embedding_Algorithm*: Speaker embedding training algorithm

Output:

- *Embedding_Model*: Trained embedding model

Steps:

1) Initialize embedding model
   Embedding_Model = Train_Embedding_Algorithm.initialize()
2) Train initial embedding model on English dataset
   Embedding_Model = Train_Embedding_Algorithm.train(Embedding_Model, English_Dataset)
3) Fine-tune embedding model on Vietnamese dataset
   Embedding_Model = Train_Embedding_Algorithm.train(Embedding_Model, Vietnamese_Dataset)
4) Return the trained model
   return Embedding_Model

overall training time and improves the accuracy of the models, as the initial training on the English dataset provides a strong foundation of learned features that can be fine-tuned for the Vietnamese data.

3.2. Application process

The application process facilitates the comparison of audio embeddings to determine if they originate from the same person, using cosine similarity metrics. It takes as input a set of reference embedding vectors ("*m_vectors*"), an embedding model capable of extracting embeddings from audio files ("*embedding_model*"), a list of audio files to be evaluated ("*audio_files*"), and a threshold value (theta). The algorithm first extracts embeddings from each audio file using the "*embedding_model* ". It then calculates the cosine similarity between each reference vector ("*m_vector*") and each extracted vector ("*n_vector*") from the audio files. After computing the cosine similarities, it calculates the average similarity across all comparisons. If this average similarity exceeds the threshold "*theta*", the algorithm concludes that the audio files likely originate from the "*Same person*". Otherwise, it determines they are from "*Different people*". This process provides a mechanism for verifying speaker consistency across audio recordings using machine learning techniques.

Instead of relying on a single pair of vectors for decision-making, the algorithm utilizes a set of vector pairs in the application process. This approach is based on the concept of ensemble learning, which combines multiple models or decision sources to improve overall performance. By aggregating the outputs from multiple vector pairs,

Algorithm 2: Application Process

Input:
- *Embedding_Model*: Model to extract embeddings trained in the training process
- *m_vectors*: List of m embedding vectors of one person stored in the database
- *audio_files*: List of n audio files of one person need to be checked.
- *theta*: Threshold value

Output:
- Is the speaker in the n audio files the same as the m embeddings?

Steps:

1) Initialize an empty list to store the cosine similarities:
   `cosine_similarities = []`
2) For each audio file in *audio_files*, do
   a) Use *Embedding_Model* to extract an embedding from the audio file
   b) Store the extracted embedding in a list called *n_vectors*
3) For each vector `x` in *m_vectors*, do
   a) For each vector `y` in *n_vectors*, do
      i) Calculate the cosine similarity between vector *x* and vector *y*
         `similarity = cosin(x,y)`
      ii) Append the cosine similarity to the list of cosine similarities:
         `cosine_similarities.append(similarity)`
4) Calculate the average of all cosine similarities:
   avg_similarity = mean(cosine_similarities)
5) If *avg_similarity* > *theta*, then
   `return "Same person"`
6) Else
   `return "Different people"`

the method reduces the likelihood of errors from any single pair, leading to a more robust and accurate decision-making process. This strategy enhances the accuracy of the models by leveraging the collective strength of the ensemble.

4. Experimental designs

4.1. Comparison strategy

Fig. 1 shows the comparison strategy. Strategy_1 is the proposed method, and to evaluate its effectiveness, experiments will compare Strategy 1 with other two strategies.

- "Strategy_1" involves a two-step training process for the speaker model: first, training on the English dataset VoxCeleb2, followed by additional training on the

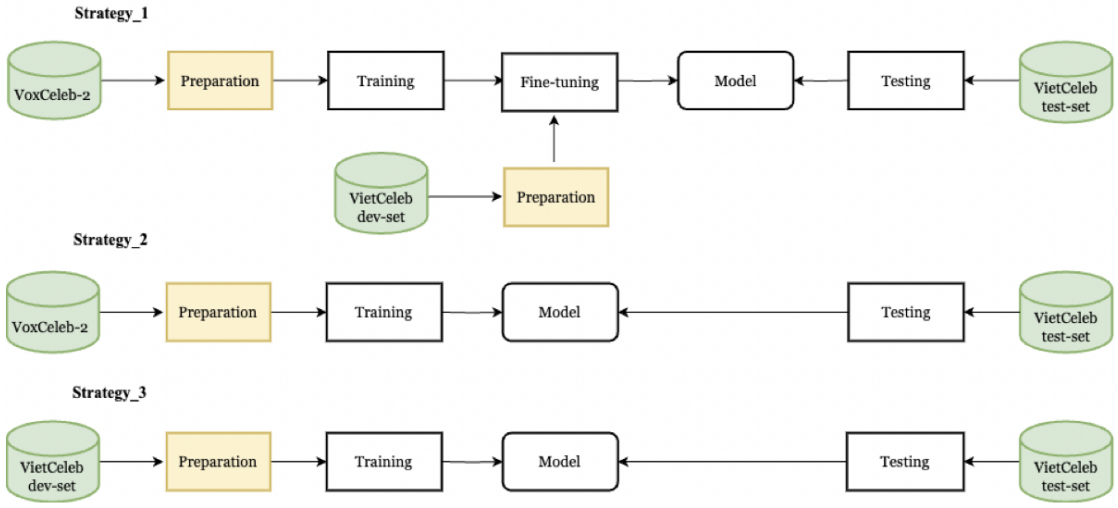


Fig. 1. Comparison strategies

Vietnamese dataset VietCeleb, and then testing with the Vietnamese dataset VietCeleb

- "Strategy_2" involves training exclusively on the English dataset VoxCeleb2 and then testing with the Vietnamese dataset VietCeleb.
- "Strategy_3" involves training exclusively on the Vietnamese dataset VietCeleb and then testing with the Vietnamese dataset VietCeleb.

4.2. Experimental setup

All training audio data are segmented into 200-frame chunks. The acoustic features used are 80-dimensional log Mel-filter banks (Fbank) with a 10 ms frameshift and a 25 ms frame window. All methods are trained 75 epochs with a consistent configuration. Additionally, the sample rate for all training is consistently set at 16 kHz. ResNet34, ResNet152, ResNet221, and ResNet293 are selected for backbone models. The $theta$ parameter is automatically selected to ensure that the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are as close to each other as possible (with minimal difference). Therefore, each configuration will have a different $theta$ value. Each strategy is evaluated on 4 scenarios which are (m = 1, n = 1), (m = 3, n = 1), (m = 5, n = 1) and (m = 5, n = 3). The models are trained on an NVIDIA Quadro RTX 5000 GPU with 16 GB of RAM.

4.3. Evaluation metric

False Acceptance Rate is calculated as (1), where FP represents false positives and TN represents true negatives. It measures the proportion of times the system incorrectly identifies an impostor as a legitimate user.

$$FAR = \frac{FP}{FP + TN} \tag{1}$$

False Rejection Rate is calculated as (2), where FN represents false negatives and TP represents true positives. This rate indicates how often the system fails to recognize a genuine user.

$$FRR = \frac{FN}{FN + TP} \tag{2}$$

An equal error rate (EER) measures the trade-off between security and usability, and it is the most used performance evaluation metric in biometric-based authentication scheme. The EER is utilized to determine the threshold where the FAR and FRR are equal. The EER is calculated as (3). A lower EER value indicates that the speaker recognition system is more accurate.

$$EER = \frac{FAR + FRR}{2} \tag{3}$$

We also use the F1 score to measure the system's accuracy. The formula for the F1 score is:

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

Here, Precision is the ratio of correctly predicted positive observations to the total predicted positives and Recall is the ratio of correctly predicted positive observations to all observations in the actual class.

5. Results and discussion

Tables 2 and 3 present equal error rates (EER) and F1 scores of three strategies, respectively. The configurations are denoted by pairs (m, n) with their corresponding EER and F1 scores.

Comparing Strategy_1, Strategy_2, and Strategy_3 reveals that Strategy_1 consistently outperforms the others. For example, with the Resnet293_LM model at the (m, n) = (5, 3) configuration, Strategy_1 achieves an EER of 0.07 and an F1 score of 99.93%, while Strategy_2 achieves an EER of 0.16 and an F1 score of 99.66%. Strategy_3 lags significantly with an EER of 1.45 and an F1 score of 98.55%. The superior performance of Strategy_1 is due to its two-step training process, which leverages both the English VoxCeleb2 dataset for initial training and the Vietnamese VoxCeleb2 dataset for fine-tuning, ensuring the model is both broadly trained and specifically fine-tuned. In contrast, Strategy_2, which only uses the VietCeleb dataset, and Strategy_3, which only uses the VietCeleb dataset, lack this balanced approach, leading to lower performance. Thus, Strategy_1 proves to be the most effective approach for optimizing model performance. Moreover, the VoxCeleb2 dataset

Table 2. Equal error rate (EER) of different models and strategies under various conditions

Models	Strategies	(m, n) = (1, 1)	(m, n) = (3, 1)	(m, n) = (5, 1)	(m, n) = (5, 3)
		EER	EER	EER	EER
Resnet293_LM	Strategy_1	2.93	1.33	1.12	**0.07**
	Strategy_2	3.19	1.85	1.58	0.16
	Strategy_3	6.36	4.30	3.77	1.45
Resnet221_LM	Strategy_1	3.01	1.70	1.50	**0.09**
	Strategy_2	6.39	3.25	2.68	0.32
	Strategy_3	7.80	6.05	5.53	3.55
Resnet152_LM	Strategy_1	3.07	1.61	1.36	**0.11**
	Strategy_2	6.45	3.72	3.14	0.63
	Strategy_3	6.83	4.68	4.26	2.37
Resnet34_LM	Strategy_1	4.66	3.42	3.12	0.69
	Strategy_2	6.76	3.93	3.42	**0.36**
	Strategy_3	14.17	10.30	9.09	4.73

Table 3. F1 score of different models and strategies under various conditions

Models	Strategies	(m, n) = (1, 1)	(m, n) = (3, 1)	(m, n) = (5, 1)	(m, n) = (5, 3)
		F1	F1	F1	F1
Resnet293_LM	Strategy_1	97.16	98.79	99.04	**99.93**
	Strategy_2	96.82	98.18	98.51	99.66
	Strategy_3	93.65	95.71	96.25	98.55
Resnet221_LM	Strategy_1	97.10	98.30	98.60	**99.87**
	Strategy_2	93.65	96.87	97.50	99.69
	Strategy_3	92.32	94.20	94.82	96.58
Resnet152_LM	Strategy_1	96.98	98.53	98.77	**99.91**
	Strategy_2	93.56	96.34	96.89	99.37
	Strategy_3	93.18	95.35	95.79	97.63
Resnet34_LM	Strategy_1	95.37	96.69	97.02	99.32
	Strategy_2	93.29	96.14	96.72	**99.65**
	Strategy_3	86.06	89.74	90.92	95.31

contains over 1 million recordings, while the VietCeleb dataset has only around 21,000 recordings. Therefore, although both were trained and tested on the same data domain, Strategy_3 performed worse than Strategy_2.

Comparing the performance of different strategies at various (m, n) highlights as the values of (m, n) increase, indicating more training and testing data, the performance generally improves across all strategies. For instance, with the Resnet293_LM model, at (m, n) = (1, 1), Strategy_1 achieves an EER of 2.93 and an F1 score of 97.16%, while at (m, n) = (5, 3), it achieves an outstanding EER of 0.07 and an F1 score of 99.93%. Similarly, Strategy_2 achieves an EER of 3.19 and an F1 score of 96.82% at (m, n) = (1, 1), and an EER of 0.16 and an F1 score of 99.66% at (m, n) = (5, 3).

Strategy_3 achieves an EER of 6.36 and an F1 score of 93.65% at (m, n) = (1, 1), and an EER of 1.45 and an F1 score of 98.55% at (m, n) = (5, 3). We tested larger (m, n) configurations such as (7, 3) during the experiment. However, the accuracy of the models only matched that of the (5, 3) configuration. The reason is that with the (5, 3) configuration, the models already achieved high accuracy (99.93%), leaving little room for further improvement. Moreover, the (5, 3) configuration generated 15 voting pairs, which is a sufficiently large number for conducting voting in binary classification tasks.

Comparing the backbone models (Resnet293_LM, Resnet221_LM, Resnet152_LM, and Resnet34_LM) reveals that models with greater depth generally achieve higher accuracy. Resnet293_LM, the deepest model, consistently demonstrates superior performance across various strategies and configurations. For example, at the (m, n) = (5, 3) configuration with Strategy_1, Resnet293_LM achieves an exceptionally low EER of 0.07 and a high F1 score of 99.93%. In contrast, Resnet34_LM, the shallowest model, performs significantly worse, with an EER of 0.69 and an F1 score of 99.32% under the same conditions. Resnet221_LM and Resnet152_LM, with intermediate depths, also show better accuracy than Resnet34_LM but do not quite match the performance of Resnet293_LM. For instance, with Strategy_1 at (m, n) = (5, 3), Resnet221_LM achieves an EER of 0.09 and an F1 score of 99.87%, while Resnet152_LM achieves an EER of 0.11 and an F1 score of 99.91%. These results illustrate that deeper models, by capturing more complex features, lead to higher accuracy in speaker recognition tasks. Thus, the depth of the backbone model plays a crucial role in determining its accuracy, with deeper models like Resnet293_LM outperforming shallower ones.

In conclusion, Strategy_1, which employs a two-step training process using both the VoxCeleb2 and VietCeleb datasets, consistently outperforms the other strategies across all backbone models. Additionally, the depth of the backbone models significantly impacts performance, with deeper models such as Resnet293_LM achieving the highest accuracy and lowest error rates. Furthermore, using higher (m, n) configurations for speaker embedding results in better performance, as evidenced by lower EER and higher F1 scores. This highlights the importance of both comprehensive training strategies and model depth, along with more extensive speaker embedding configurations, in enhancing speaker recognition accuracy.

6. Conclusion

This paper proposes innovative strategies including transfer learning and ensemble embedding techniques to improve Vietnamese speaker recognition. Experimental results demonstrate that the proposed method consistently outperforms other strategies across various backbone models. The depth of the backbone models significantly enhances accuracy and reduces error rates. Moreover, employing higher configurations for speaker embedding improves performance, highlighting the overall effectiveness of comprehensive training strategies and model depth in advancing speaker recognition

capabilities. These insights underline the robustness of our approach in optimizing Vietnamese speaker recognition systems. These outcomes underscore the promise of tailored approaches in advancing speaker recognition technology for Vietnamese speakers, promising broader applications and continued progress in this dynamic field.

Acknowledgement

This research is funded by LQDTU Research Programs (Grant No. 24.KGM.04).

References

[1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015, DOI: 10.1109/MSP.2015.2462851

[2] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021, DOI: 10.1016/j.neunet.2021.03.004

[3] K. Radha, M. Bansal, and R. B. Pachori, "Speech and speaker recognition using raw waveform modeling for adult and children's speech: A comprehensive review," *Engineering Applications of Artificial Intelligence*, vol. 131, p. 107661, 2024, DOI: 10.1016/j.engappai.2023.107661

[4] N. Singh, R. Khan, and R. Shree, "Applications of speaker recognition," *Procedia Engineering*, vol. 38, pp. 3122–3126, 2012, DOI: 10.1016/j.proeng.2012.06.363

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000, DOI: 10.1006/dspr.1999.0361

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007, DOI: 10.1109/TASL.2006.881693

[7] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008, DOI: 10.1109/TASL.2008.925147

[8] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010, DOI: 10.1109/TASL.2010.2064307

[9] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP),* Czech Republic*, 2021*, 2014, pp. 4052–4056, DOI: 10.1109/ICASSP.2014.6854363

[10] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP),* Floren*,* Italy, 2014, pp. 1695–1699, DOI: 10.1109/ICASSP.2014.6853887

[11] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, vol. 2017, 2017, pp. 999–1003, DOI: 10.48550/arXiv.1804.10080

[12] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017, DOI: 10.48550/arXiv.1705.02304

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE in international conference on acoustics, speech and signal processing (ICASSP), Canada*, 2018, pp. 5329–5333, DOI: 10.1109/ICASSP.2018.8461375

[14] J. S. Chung, J. Huh, and S. Mun, "Delving into voxceleb: environment invariant speaker recognition," *arXiv preprint arXiv:1910.11238*, 2019, DOI: 10.48550/arXiv.1910.11238

[15] D. D. T. Thu, L. T. Van, Q. N. Hong, and H. P. Ngoc, "Text-dependent speaker recognition for vietnamese," in *2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR).* HaNoi, Vietnam, 2013, pp. 196–200, DOI: 10.1109/SOCPAR.2013.7054126

[16] M. Jakubec, R. Jarina, E. Lieskovska, and P. Kasak, "Deep speaker embeddings for speaker verification: Review and experimental comparison," vol. 127, p. 107232, 2024, DOI: 10.1016/j.engappai.2023.107232

[17] W. Lin and M.-W. Mak, "Mixture representation learning for deep speaker embedding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 968–978, 2022, DOI: 10.1109/TASLP.2022.3153270

[18] P. T. Nghia, P. V. Binh, N. H. Thai, N. T. Ha, and P. Kumsawat, "A robust wavelet-based text-independent speaker identification," in *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007),* India, vol. 2. IEEE, 2007, pp. 219–223, DOI: 10.1109/ICCIMA.2007.149

[19] D. Tran and M. Wagner, "A fuzzy approach to speaker verification," *International journal of Pattern Recognition and Artificial Intelligence*, vol. 16, no. 07, pp. 913–925, 2002, DOI: 10.1142/S0218001402002064

[20] S. T. Nguyen, V. D. Lai, Q. Dam-Ba, A. Nguyen-Xuan, and C. Pham, "Vietnamese speaker authentication using deep models," in *Proceedings of the 9th International Symposium on Information and Communication Technology*, 2018, pp. 177–184, DOI: 10.1145/3287921.3287954

[21] C. T. Tran, D. T. Nguyen, and H. T. Hoang, "Deep representation learning for vietnamese speaker recognition," in *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*. Bangkok, Thailand, 2021, pp. 1–4, DOI: 10.1109/KSE53942.2021.9648808

[22] P. V. Thanh, N. X. T. Hoa, H. L. Vu, and N. T. T. Trang, "Vietnam-celeb: a large-scale dataset for vietnamese speaker recognition," 2023, DOI: 10.21437/Interspeech.2023-1989

[23] D. V. Thanh, T. P. Viet, and T. N. T. Thu, "Deep speaker verification model for low-resource languages and vietnamese dataset," in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation China*, 2021, pp. 442–451.

[24] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, pp. 1–40, 2020, DOI: 10.1016/j.csl.2019.101027

[25] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018, DOI: 10.48550/arXiv.1806.05622

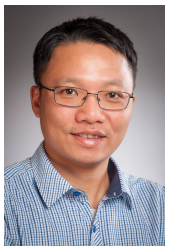
[26] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009, DOI: 10.1109/TKDE.2009.191

[27] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, pp. 241–258, 2020, DOI: 10.1007/s11704-019-8208-z

Manuscript received: 20-08-2024; Accepted: 27-12-2024.

■

Tan Hoang Ho graduated with a degree in Information Technology from Le Quy Don University in 2023. Currently, he is working as a researcher at the Faculty of Computer Science, Le Quy Don Technical University. His primary research interests focus on advanced topics in speaker recognition, natural language processing, large language model, and their practical applications in cutting-edge technologies. His work often involves exploring innovative methods to enhance the accuracy and efficiency of machine learning models for these fields, contributing to the development of intelligent systems and AI-driven solutions.
Email: hotanhoang2501hh@gmail.com



Cao Truong Tran received the PhD degree in computer science from Victoria University of Wellington, New Zealand. He also did postdoc at Victoria University of Wellington. He is researching in the field of machine learning and evolutionary computation, specialized with evolutionary machine learning for data mining with missing data. He serves as a reviewer of international journals, including IEEE Transactions on Evolutionary Computation, IEEE Transactions on Cybernetics, Pattern Recognition, Knowledge-Based Systems, Applied Soft Computing and Engineering Application of Artificial Intelligence. He is also a PC member of international conferences, including IEEE Congress on Evolutionary Computation, IEEE Symposium Series on Computational Intelligence, the Australasian Joint Conference on Artificial Intelligence, and the AAAI Conference on Artificial Intelligence.
Email: truongct@lqdtu.edu.vn

CẢI TIẾN CHẤT LƯỢNG CỦA NHẬN DẠNG NGƯỜI NÓI TIẾNG VIỆT SỬ DỤNG HỌC CHUYỂN ĐỔI VÀ HỌC BIỂU DIỄN CỘNG ĐỒNG

Hồ Tấn Hoàng, Trần Cao Trưởng

Tóm tắt

Công nghệ nhận dạng người nói giúp nhận diện hoặc xác minh danh tính dựa trên các đặc điểm giọng nói riêng của từng người, chẳng hạn như cao độ, tông giọng và cách nói chuyện. Công nghệ này được sử dụng rộng rãi để tăng cường an ninh, cải thiện dịch vụ khách hàng, hỗ trợ điều tra và tạo ra các tương tác cá nhân hóa với thiết bị thông minh. Trong những năm gần đây, nhờ ứng dụng các kỹ thuật học sâu, nhận dạng người nói đã đạt được nhiều tiến bộ. Tuy nhiên, việc nhận dạng người nói tiếng Việt vẫn còn gặp nhiều khó khăn. Bài viết này đề xuất những chiến lược mới, kết hợp giữa học chuyển giao và học cộng đồng, nhằm cải thiện độ chính xác trong nhận dạng người nói tiếng Việt. Các kết quả thử nghiệm trên dữ liệu tiếng Việt cho thấy độ chính xác đã được cải thiện rõ rệt.

Từ khóa

Nhận dạng người nói; học chuyển đổi; học cộng đồng; học biểu diễn; học sâu.