MINISTRY OF EDUCATION
AND TRAINING

VIETNAM ACADEMY OF
SIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SIENCE AND TECHNOLOGY

Nguyen Thi Thanh Mai

IMPROVING SPEAKER VERIFICATION ACCURACY USING DEEP LEARNING WITH LIMITED RESOURCES

SUMMARY OF DISSERTATION ON COMPUTER
Major: Computer science
Code: 9 48 01 01

Hanoi - 2024

The dissertation is completed at: Graduate University of Science and Technology, Vietnam Academy of Science and Technology

Supervisors:

Supervisor 1: Assoc. Prof. Dr. Nguyen Duc Dung, Institute of Information Technology

Supervisor 2: Assoc. Prof. Dr. Luong Chi Mai, Institute of Information Technology

Referee 1: Assoc. Prof. Dr. Phan Xuan Hieu

Referee 2: Assoc. Prof. Dr. Vu Hai

Referee 3: Dr. Do Van Hai

The dissertation will be examined by Examination Board of Graduate University of Science and Technology, Vietnam Academy of Science and Technology at…….. (time, date, year…)

This dissertation can be found at:

1) Graduate University of Science and Technology Library
2) National Library of Vietnam

INTRODUCTION

1. The necessity of the research

In recent years, speaker verification (SV) based on deep learning models has achieved superior results compared to traditional machine learning models. In the traditional approach, the process of extracting acoustic features was conducted manually and separately from the speaker feature modeling process. For instance, MFCCs (Mel-Frequency Cepstral Coefficients) have been widely used as input features for many speech processing systems, including SV systems. The strength of MFCCs lies in their ability to represent speech signals in a compressed form, capturing the essential phonetic content of the voice. However, most of the energy of MFCCs is concentrated in the lower-order coefficients, typically the first 13 to 39 coefficients. When MFCCs are used as input features for deep learning models like ResNets, the energy present in the higher-order MFCCs may reduce the performance of the SV system.

Thus, it is crucial to research and discover new acoustic features to enhance the performance of SV systems. Moreover, training deep learning models for SV often requires large and diverse datasets. Currently, English-language speaker data significantly outweighs Vietnamese-language data. Specifically, the VoxCeleb2 dataset [14] includes 6,112 speakers with 2,442 hours of recorded speech, whereas the Vietnamese VLSP2021-SV dataset [18] contains only 1,305 speakers with 41 hours of recordings (English data amounts to 60 times the hours of Vietnamese data).

Given the limited availability of Vietnamese-language speaker data, potential solutions include collecting additional Vietnamese speech data or reusing models trained on large English or Chinese datasets and fine-tuning them on Vietnamese speech data. However, collecting and expanding Vietnamese data can be costly and challenging. Consequently, models trained on limited datasets may overfit and fail to generalize to unseen data. Transfer learning offers the advantage of inheriting high-level features from large-scale English datasets, reducing the time required to train models on Vietnamese speaker data.

With the rapid development of deep learning models for SV tasks, selecting appropriate features and models for limited speech data becomes an important focus of this thesis, requiring thorough research, comparisons, experimentation, and evaluation.

In addition, research on SV systems is highly needed for integration into intelligent systems with widespread real-world applications such as:

- **Preventing unauthorized access:** SV systems help ensure that only authorized individuals can access systems, services, or sensitive information.
- **Protecting personal data:** In the context of increasing threats to personal information due to theft and fraud, SV systems provide an additional layer of security, ensuring data access is restricted to the rightful owner.
- **Reducing password management costs:** Traditional password management and recovery can be costly and complex, while voice verification can minimize these expenses.

- **Optimizing processes:** Voice verification systems can automate and simplify many authentication processes, reducing manual workload and staffing costs.

For such reasons, the thesis chooses the research topic ***"Improving speaker verification accuracy using deep learning with limited resources*** ". This is an urgent and topical issue with high applicability. The research results of the thesis help improve the accuracy of Vietnamese speaker verification.

2. **Research objectives**

The research objective of the thesis is to propose some solutions to improve speaker verification accuracy with limited resources. The specific objectives are:

- Research and select acoustic features for deep learning models to improve speaker verification accuracy;
- Research, analyze, compare, and evaluate deep learning models and transfer learning methods applied to limited data resources to improve speaker verification accuracy.

3. **Thesis layout**

Chapter 1: Overview of basic knowledge and deep learning applications for speaker verification task

Chapter 1 introduces an overview of the speaker verification task using a deep learning approach. Thereby, it describes the overall speaker verification system and research directions to improve speaker verification performance in accordance with current trends and practices.

Chapter 2: Improving the accuracy of Vietnamese speaker verification using Mel-Filter bank feature with ECAPA-TDNN model

Chapter 2 focuses on surveying, evaluating, and testing the input features for modern deep learning models, specifically the ECAPA-TDNN model. Experiments with the model proposed in the thesis show that the Mel-Filterbank Energys feature with the ECAPA-TDNN model gives better results than the MFCCs feature (for ECAPA-TDNN).

Chapter 3: Improving speaker verification accuracy using transfer learning with Rawnet3 model

Chapter 3 tests and evaluates the accuracy of the speaker verification system using transfer learning techniques. With pre-trained models on large datasets, using the Rawnet3 deep learning model with raw audio data as input, then fine-tuning and training on Vietnamese data gives better results than without transfer learning.

Conclusion . Present the main contributions of the thesis and point out the limitations and directions for further development.

4. **Contributions of the thesis**

- **Proposal to use MFBEs feature with ECAPA-TDNN model for Vietnamese speaker verification task; ([CT1] and [CT2])**
- **Proposal to use RawNet3 model in transfer learning for speaker verification task with limited resources. ([CT3] and [CT4])**

CHAPTER 1: OVERVIEW OF KNOWLEDGE AND APPLICATION OF DEEP LEARNING FOR SPEAKER VERIFICATION TASK

In Chapter 1, the first part introduces an overview of related research on speaker verification and the difficult tasks that need to be solved. Next, the researcher presents an overview of the research situation at home and abroad as well as approaches in speaker verification. Finally, the researcher presents an overview of the speaker verification system: characteristics, models, data, evaluation methods, and improvement methods to improve speaker verification accuracy.

1.1.Introduce

Speaker verification is one of the tasks in the field of voice-based biometric identification and authentication. The goal of this task is to check whether a person's voice matches a previously registered voice template.

Speaker verification is a task in the field of voice-based biometric identification and authentication. The goal of this task is to check whether a person's voice matches a previously registered sample voice.

Input : A voice signal segment of the user to be authenticated (called test voice), and a voice sample stored in the system (called registered voice/sample voice).

Output : An authentication decision to answer the question: "Is the test speaker the same as the registered voice sample?" Based on the comparison with the threshold, the system will return "true" (accept) or "false" (reject).

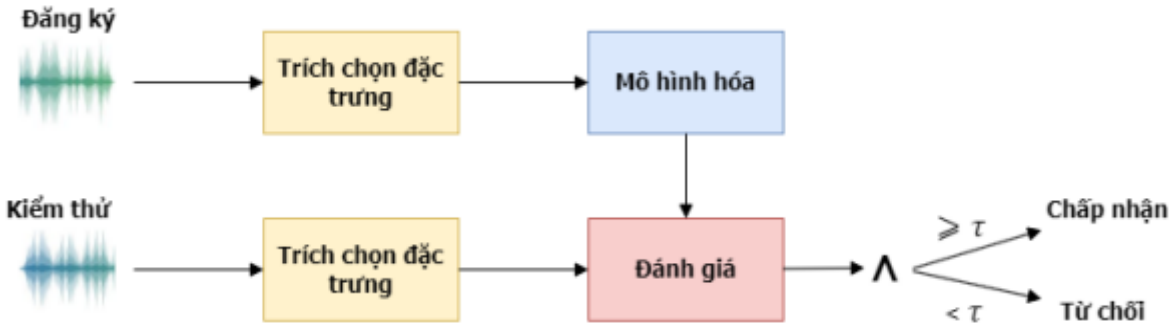

Figure 1.1: Overview diagram of speaker verification system.

1.2.Related works

Foreign research situation

Speaker recognition and authentication research is still considered a pursuit towards improving recognition accuracy. For example, early research was limited to text-dependent constrained tasks and focused on dealing with variations caused by random pronunciation, where the Hidden Markov Model (HMM) [87] was the most popular model in text-independent speaker verification methods and had to deal with phonetic variations, which gave rise to the Gaussian mixture modelling with a universal background (GMM-UBM) [92]. Further research has attempted to deal with the inter-session variation due to channel and speaking style, where the i-vector/PLDA (Probabilistic Linear

Discriminant Analysis) architecture is the most popular and successful [19]. Recently, researchers have focused on dealing with complex variations in natural situations and deep learning methods have proven to be very powerful [106][108][114].

Deep learning methods for speaker recognition have attracted much attention due to advances in computational power and the availability of large datasets in the wild [72]. A large number of studies using DNN models for speaker embedding extraction have been carried out in the past few years. Most of the prominent studies have used Convolutional Neural Network (CNN) architectures such as ResNet [121] which have shown good results in the past few years. On the other hand, other successful models such as x-vectors [108] have used TDNN to extract feature embeddings from MFCCs. Most of the DNN models used in speaker recognition take a single utterance as input and provide a fixed-size vector as the speech embedding for an utterance. Another process then computes the similarity between the two embedding vectors (registered utterance and test utterance) to identify the speaker.

Recurrent Neural Network (RNN) is also used in some studies. Recently, RNN model [80][100][126] was developed using MFCCs coefficients.

In [118] the LSTM (Long Short-Term Memory) architecture is also applied on MFCC, the embedding results use speaker identification calculated by the average cosine distance. On the other hand, these models also use the LSTM architecture as an i-vectors extraction tool. Some other studies use a combination of CNN and RNN based on convolutional layers between input MFCCs and RNN.

In the publication [14], the research group used a CNN model trained on data of about 6,000 English voices. With this approach, each 3-second speech segment will be transformed into a spectrogram. These images will be the input for the CNN network and the system gives quite good results with an error rate of 3.95% on the test data [60].

A research direction of speaker recognition on short speech data of less than 2 seconds [126] has also attracted the attention of the research community. This research group also used x-vectors [108] as the basic model and then developed and extended the TDNN architecture [82].

In the published paper [73] experimenting on Voxceleb2 data [14], the US research team gave the error rate evaluation result of 3.82%, the Chinese AI company gave the result of 3.81%, the IDLab group in Belgium gave the best result of 3.73%.

From 2019 to 2023[15][40][73], several competitions focused on speaker recognition techniques were held. These competitions aim to promote research in the field of speaker recognition, while providing baseline recognition systems, training data, and evaluation criteria. The tasks in the competition include: speaker verification, speaker identification, and speaker separation.

Research situation in the country

In Vietnam, research and application of speaker recognition has also been a field that has attracted the attention of researchers and developers in recent years. The following groups and research directions can be mentioned: At the Zalo AI Challenge 2020, the speaker recognition task achieved an error rate of 5%. The model was trained on 400 Vietnamese voices and evaluated on the

data of the Organizing Committee. In addition, another research group also used the multi-task 11 learning model [84] combined with the Triplet loss function [23] for the voice authentication task. The model was trained on English data, then fine-tuned on a small amount of data for Vietnamese. The evaluation results on 65 voices of the VIVOS Vietnamese Database [67] had an error rate of 4.3%. In the natural language processing research community, the speaker recognition task is also a task of interest.

The VLSP 2021 Conference [18] also included a Vietnamese speaker recognition competition with a published database of more than 1,300 voices. The competition also attracted the research community and participating teams, and the best test result from the Organizing Committee had an error rate of 1.9%. One of the models that the participating teams tested was the ECAPA-TDNN model. The ECAPA-TDNN model [22] is also widely applied in tasks such as language recognition, emotion recognition, ...

The research team [111] used the log Mel-filterbanks feature as input for the ResNet deep learning network. The experimental results were evaluated on data collected by the team on YouTube channels with 580 speakers and 5,000 sentences. The experimental results showed that using the available training model on the English data set, then fine-tuning on Vietnamese data gave better results than training only on Vietnamese data.

The group of authors from the Academy of Posts and Telecommunications [76] also tested the comparison between the MFCCs feature and the GFCCs feature [112] on a limited Vietnamese dataset with 20 self-recorded speakers' training data.

The authors experimentally compared the error rates of two GMMs models and the ResNet model. The results showed that the ResNet model using GFCCs as input features gave a lower error rate than the traditional GMMs model.

The research team at Hanoi University of Science and Technology has also built the Vietnamese speaker recognition database Vietnam-Celeb [83] with 1,000 speakers. This is the latest and largest dataset used for the Vietnamese speaker recognition task. The researcher will present this database in detail in the following section.

1.3.Speaker verification with limited resource data

In today's digital age, speaker verification has become an important part of security and identification applications, such as in electronic payment systems, secure access to sensitive information, and voice recognition in virtual assistants. However, one of the biggest challenges in developing effective speaker verification systems is the lack of data resources, especially labeled data. Collecting labeled voice data (e.g., speaker identity) is often expensive and time-consuming. When the number of labeled samples is limited, it becomes difficult to train machine learning models, resulting in poor performance in speaker verification. Each speaker has unique voice characteristics, and the variation between individuals can be very large. When labeled data is insufficient, models cannot learn accurate features to distinguish between different speakers. Models trained on a small dataset may not be able to generalize well when applied to real-world situations where there is diversity in sounds, environmental conditions, and speaker accents.

Some major research directions and works related to solving the limited data task in speaker verification:

- Models such as Wav2vec [8] and HuBERT [37] have exploited self-supervised learning on unlabeled audio data, allowing the model to learn rich speech features without direct labeling. These models have been shown to be effective in speaker verification and recognition when data is limited.
- The CSSL (Contrastive Self-Supervised Learning) method in self-supervised learning [55] uses comparison between different audio segments of the same speaker to build features, thereby minimizing the need for labeled data in speaker verification.
- SpecAugment [79] is a popular data augmentation technique used on spectrograms by transforming audio segments with various levels, such as frequency and time shifts. This technique is applied to improve the generalization ability of speaker verification models.
- Research on Prototypical networks [103] shows the ability to recognize speakers with only a small number of samples. This method learns representative feature representations for each speaker class, thus allowing accurate classification even with a limited number of training samples.
- Siamese models [54] have also been applied to speaker recognition and authentication tasks with limited sample size, significantly improving the accuracy of the system.
- Transfer learning: Studies using speaker embeddings from models such as x-vector [108] and ResNet [74] allow the transfer of features learned from other tasks or more general data to the task of speaker verification with limited data. The Res2Net [94] and ECAPA-TDNN [22] models have achieved success in speaker recognition and speaker verification by leveraging skeleton layers that are capable of learning deep and scalable features from limited data samples.
- Handcrafted feature learning and joint learning: Traditional handcrafted features such as MFCCs and spectrograms are combined with features learned from deep learning models to take advantage of both types of features and improve model performance when data is limited [12]. With the currently published Vietnamese speaker datasets such as VLSP2021-SV [18], Vietnam-Celeb [83], thesis focuses on handcrafted feature selection methods combined with deep learning networks and transfer learning methods to improve speaker verification accuracy.

1.4.Deep learning approaches to speaker verification task

There are two approaches in speaker verification: statistical based approach and deep learning based approach. In this thesis, the researcher focuses on deep learning based approaches.

Deep neural networks have been very successful in feature extraction to learn discriminative embeddings in both computer vision and speech. The methods typically combine classifiers and train them independently. While concatenated methods are highly effective, as DNNs do not train end-to-end and still require feature extraction techniques. In contrast, CNN architectures can be used directly

from raw spectrograms and trained end-to-end. Deep learning systems from input to output models for speaker recognition typically use three stages:

- Feature extraction using DNN
- Frame level feature aggregation
- Optimizing the loss function for the classification objective.

DNN-based architectures typically use 2D CNN with convolution for both time and frequency domains [44] or 1D CNN with convolution applied to time domain [31]. Some studies also use LSTM-based end-to-end architectures [98]. The output of the feature extractor depends on the input pronunciation length. The pooling layer uses and aggregates the frame-level feature vectors to obtain fixed-length embedded features that guide the method expansion in standard deviation as the mean. This method is called statistical pooling. Unlike the methods where information from all frames is weighted equally, attention models are developed to weight the discriminative frames. Here attention models and statistical models are combined for statistical attention pooling. This final pooling stage is of interest as LDE. This method is close to the NetVLAD layer [5] designed for image retrieval.

Such systems are trained end-to-end for classification using a softmax function or one of its modifications such as Angular softmax [39]. In some cases, the network is trained for validation using a Contrastive loss function[116] or a triplet loss function [23]. Similarity measures such as cosine [30] or PLDA [42] are often used to generate the final pairwise comparison scores.

1.4.Overview diagram of speaker verification system

The system consists of the following main components (Figure 1.1): feature extraction, speaker modeling, and evaluation. Feature extraction transforms the audio signal into a set of features that distinguish each individual speaker, also known as speaker embeddings. During the enrollment phase, the speaker model uses the input features to build a statistical model that represents the unique characteristics of each specific speaker. This model is often called a speaker model or a voice model, which is used to infer during the authentication process whether a given voice sample belongs to a registered speaker or not. The authentication decision is based on the scoring module, which evaluates the new speaker's features with the registered voice features. If the scoring score is greater than or equal to a predefined threshold τ , then the authentication process is successful and the user is authenticated.

Otherwise, the process will fail i.e. the given voice sample does not belong to the registered voice.

The modules mentioned above are the basic modules of a speaker verification system and directly affect the overall performance of the speaker verification system. The basic diagram in Figure 4 can be applied to both traditional and deep learning-based methods. In this section, we will analyze the feature extraction, speaker modeling, and evaluation modules in three state-of-the-art deep learning models for the speaker verification task: VGGVox, ECAPA-TDNN, and RawNet.

1.4.1. Feature extraction

Deep learning has been proven to be a powerful technique for extracting high-level features from low-level information. The features extracted from hidden layers of various deep learning

models are called deep features. Deep features can be extracted from any deep learning model such as convolutional neural networks (CNN), deep neural networks (DNN), recurrent neural networks (RNN), one-way long short-term memory networks (LSTM), bidirectional long short-term memory (BLSTM) and other similar models.

Deep features are extracted from deep neural networks (DNNs). MFCCs or any other relevant audio features are provided as input to the DNN. The deep features depend on the depth of the neural network. If we have a shallow neural network, the deep features provided by the lower layers can be considered as speaker-adaptive features. And from the upper layers, class-based discriminative features can be extracted. Deep features can also be extracted from the bottleneck layer of the DNN.

1.4.2. Speaker modeling

In this thesis, the researcher will present some of the most popular deep learning networks for speaker verification such as i-vector, d-vector, x-vector, resnets, ECAPA-TDNN, SincNet, RawNets, ...

The basic structure of a deep learning network typically includes the following components: Input Layer: This layer receives input data and forwards it to the network. The number of neurons in this layer depends on the size of the input data.

- Hidden Layers: These layers contain neurons and perform transformations and calculations on the input data. Each hidden layer can have multiple neurons and is usually determined by the number and type of neurons, as well as how they are connected to each other.
- Weights and Biases: Each connection between neurons in successive layers has a weight, which represents the strength of that connection. Additionally, each neuron has a bias to adjust and adapt to the input data.
- Activation Functions: Each neuron in the network typically applies an activation function to produce a non-linear output. Common activation functions include ReLU (Rectified Linear Unit), Sigmoid, Tanh, and Leaky ReLU.
- Output Layer: This layer produces the predicted output of the network. The number of neurons in this layer depends on the type of task, for example, one neuron per class in a binary classification task, or one neuron per class in a multi-class classification task.
- Loss Function: This function calculates the loss between the network's prediction and the actual value. It measures the model's performance and is used during training to tune the network's parameters.
- Optimizer: The optimizer is used to update the weights and biases of the network based on the loss values computed from the training data. Common optimization methods include Gradient Descent and its variants.

1.4.3 . Evaluation

Cosine distance is used to compare the similarity between two vectors.

Cosine distance is calculated by cosine similarity. Cosine similarity is defined as the similarity between two non-zero vectors. It calculates the cosine of the angle between two vectors in a multidimensional space. The relationship between cosine similarity and cosine distance is not

symmetric. Cosine similarity increases while the distance between vectors decreases and vice versa. The following equation calculates cosine similarity and cosine distance respectively. The functions are represented by the following formula:

$$\cos(\theta) = \frac{A.B}{\|A\|.\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1.11)$$

$$cosine_{distance} = 1 - \cos(\theta) \quad (1.12)$$

Where A, B are non-zero vectors and $\cos(\theta)$is the cosine similarity.

1.5. Test datasets for speaker verification task

VoxCeleb1

The VoxCeleb1 dataset [72] is a large dataset containing speech samples from celebrities, collected from YouTube videos. VoxCeleb1 contains more than 100,000 utterances from 1,251 speakers. This database was published in 2017, and is a large dataset for speaker recognition.

VoxCeleb2

The VoxCeleb2 database [14] is a large open dataset containing speech samples from many famous people , collected from YouTube videos. VoxCeleb2 is an extended version of VoxCeleb1 and was released later with significant improvements and extensions. VoxCeleb2 contains over 1 million utterances from over 6,000 famous people, extracted from videos uploaded to YouTube. The dataset is fairly gender-balanced, with 61% of the speakers being male.

VLSP2021-SV

Recently, the VLSP 2021 workshop [18] published a Vietnamese speaker verification and recognition dataset in a noisy environment containing 50 hours of speech from more than 1,300 speakers (the thesis calls this dataset VLSP2021-SV). The data is collected from many different sources, including from the ZaloAI competition, VLSP2020-SV, VIVOS, and data collected from TV shows and YouTube channels in environments with diverse background noise such as small talk, laughter, street noise, school, and music.

Vietnam-Celeb

The Vietnam-Celeb dataset [83] includes 1,000 speakers and more than 87,000 utterances. The total duration of the dataset is 187 hours, and the utterances are sampled at 16,000 Hz. The data covers all scenarios such as interviews, game shows, talk shows, and other types of entertainment videos.

Table 1.15: Statistics of subsets of Vietnam-Celeb .

Subset	Number of speakers	Number of sentences	Number of pairs
Vietnam-Celeb-T	880	82,907	-
Vietnam-Celeb-E	120	4,207	55,015
Vietnam-Celeb-H	120	4,217	55,015

Methods to improve the accuracy of speaker verification systems

Data Augmentation

Training in large-scale conditions is an effective way to improve speaker verification in noisy environments. In particular, the performance of deep learning-based speaker verification systems depends heavily on the amount of training data. One method to prepare a large amount of noisy data is data augmentation. In [108], the authors used additive noise and reverberation on the original training data to effectively augment the x-vector data. In [136], a combined learning strategy was applied to improve the x-vector extractor.

Feature selection

In speaker verification, hand-crafted feature selection as input to deep neural networks is a popular method to combine traditional acoustic features with the deep learning capabilities of neural networks. This helps to take advantage of both the available acoustic information from hand-crafted features and the complex pattern analysis capabilities of deep neural networks. Some popular hand-crafted features as input include: MFCCs, spectrograms, raw audio, FBank. Hand-crafted feature selection as input to deep neural networks not only improves the performance of the model but also increases the ability to exploit important features from the speech signal, especially in cases with limited data. Hand-crafted features can be combined with features learned from layers of deep neural networks, such as CNNs [14], [16], to create hybrid features, thereby optimizing the accuracy in speaker verification. Table 1.1 shows that the Mel-filter Bank input feature gives the best results on VoxCeleb1 and VoxCeleb2 data with EER 0.66%. If we only consider the audio wave features, the RawNet3 model gives the best results with an error rate EER 0.89%.

Speaker verification System Evaluation Metrics

EER

EER is the point where the false acceptance rate (FAR) is equal to the false rejection rate (FRR) in the authentication system. The meaning of this index is explained as follows: The higher the FRR, the more secure the system. However, there are many rejections of legitimate users. Thus, users have to perform many authentications to get a successful message, leading to a decrease in user experience. Therefore, the sensitivity and convenience of the system are poor. On the contrary, if the false rejection rate (FRR) is too small, the FAR is often very high. As a result, the system accepts many invalid user authentications or users easily authenticate successfully. This affects the security of the system. The point where FAR = FRR is called the equal error rate (EER). At this point, the system balances between security and sensitivity and convenience. Therefore, EER is often used as a measure for authentication systems. The smaller the EER, the better the speaker verification quality of the system.

Chapter 1 Conclusion

In Chapter 1, thesis presented an overview of the speaker recognition and authentication system based on deep learning models. Like the traditional approach, the authentication system includes three main modules: feature extraction, feature modeling, and evaluation. Through this chapter, thesis also

surveyed the world's published data sets for the authentication task, the latest approaches today as well as the challenges for this task. Specifically:

- Datasets used for training and evaluating the validation model: VoxCeleb1, VoxCeleb2, Cn-Celeb2, VLSP2021-SV, Vietnam-Celeb.
- Modern deep learning models with diverse input features applied to speaker verification task: ResNets, x-vector, ECAPA-TDNN, RawNets.
- With deep learning models, training data plays a vital role in the accuracy of the model.

Thus, it is necessary to solve the task of limited training data in two ways: Applying transfer learning and selecting appropriate acoustic features for the most advanced speaker verification models to improve the efficiency of Vietnamese speaker verification. The overview studies in this chapter are the basis for the researcher to propose solutions to improve the speaker verification system with limited resources in the following chapters.

Some initial research results on the basic voice authentication system are published in the work **[CT4]** in the section "List of works of the author"

CHAPTER 2: IMPROVING THE ACCURACY OF AUTHENTICATION OF VIETNAMESE SPEAKING PEOPLE USING MEL-FILTERBANK ENERGYS FEATURES WITH ECAPA-TDNN MODEL

In Chapter 2, thesis focuses on selecting acoustic features as input to modern deep learning networks in the speaker verification system with Vietnamese data. Through this chapter, thesis presents the extraction of MFBEs and MFCCs features and analyzes the limitations of MFCCs in speaker verification. From there, there are comparisons of the differences between the two features MFBEs and MFCCs and experimental evaluation of these two features in the Vietnamese speaker verification system.

Vietnamese speaker verification task and speech features

Speaker verification is an important area in speech recognition technology, especially in the context of increasing demands for security and privacy of personal information. In Vietnam, the task of Vietnamese speaker verification is attracting great attention from both researchers and technology developers, due to the unique linguistic and cultural characteristics of Vietnamese. Vietnamese has a rich tone system, with six different tones, along with phonetic diversity between the Northern, Central, and Southern regions. This diversity creates great challenges for authentication systems, requiring the development of methods that can accurately recognize and distinguish between different voices, even in different environmental conditions. Vietnamese speaker verification is currently facing many challenges such as lack of training data, evaluation results on the latest Vietnam-Celeb dataset [83] have an EER error rate greater than 10%, so experimental studies are still needed to improve the quality of the authentication system. With modern approaches based on deep learning, choosing acoustic features and choosing training models are among the solutions to improve speaker verification accuracy on Vietnamese speaker data.

Limitations of MFCCs

Features for automatic speech recognition have been evaluated based on the fuzzy method including MFCC, DTW and FFT. The results show that MFCC improves the performance of the Fuzzy model compared to the FFT feature [102]. Al-Ali et al. improved forensic voice authentication based on the combination of MFCC and DWT features, in which their model was evaluated in a noisy environment [2]. Although MFCC is capable of capturing speaker characteristics, the performance of MFCC degrades on complex speech datasets and in noisy environments. For example, [4] study shows that speaker recognition using MFCC and k-NN is significantly reduced in noisy environments and concludes that cleaning the input signal can improve the results when using the highest MFCCs. To overcome this task, in [10], a multi-channel training framework in deep speaker embedding network was proposed for speaker verification in reverberant and noisy environment. This method takes temporal, frequency and spatial information from multi-channel input to improve the robust speaker embedding process. The work concluded that slightly increasing the parameters of the model can help this method significantly outperform the i-vector system with MFCC using signal enhancement at the input. In addition, Jahangir et al. proposed MFCC-based combined features and time-based features. These combined features are fed into DNN for speaker recognition. The results showed that the limitation of MFCC feature can be solved by this method [43].

Comparison of characteristics of MFCCs and MFBEs

The main difference between MFBEs and MFCCs is the use of discrete cosine transform (DCT) [109]. MFBEs can be either homogeneous or non-homogeneous DCT depending on the specific implementation, whereas MFCCs always involve DCT to compress information into a smaller set of coefficients. Both MFCCs and MFBEs have an impact on the representation of audio signals in speech and audio processing applications. MFCCs provide time series information of energy versus frequency from an audio source. The correction from the raw DFT-based energy series serves two purposes:

- Change the linear scale (of frequency and energy) from the raw DFT to a log scale. This matches the hearing of humans (and most animals) in perceiving sound.
- Compressing large amounts of data into smaller features while still being able to distinguish differences between sounds . This is especially useful in the high frequency domain for most speech recognition applications, detecting the difference between energy levels at 1001 Hz and 999 Hz.

The advantage of using the discrete cosine transform over the discrete Fourier transform is that it removes noise from the speech signal.

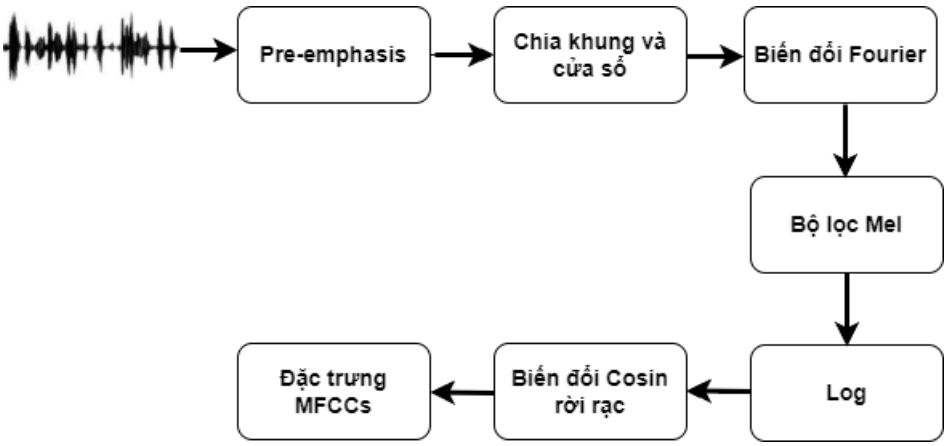

Figure 2.1: MFCCs feature extraction steps.

The analysis of different input feature representations aims to select suitable inputs for deep neural acoustic models. MFCCs are inferior to DCT transforms, MFCCs cause deep neural models to discard speaker information.

2.2. Proposal to use MFBEs feature with ECAPA-TDNN model in Vietnamese speaker verification

Training phase

Unlike the VGGVox model [74], ECAPA-TDNN and x-vector do not use a two-dimensional (2D) spectrogram as input but use a one-dimensional (1D) array of MFCCs with a frame length of 25 ms (x-vector uses a 24-dimensional MFCC, mean-normalized over a 3-second window, with a 10-ms step; ECAPA-TDNN uses an 80-dimensional MFCC, mean-normalized over a 2-second window, with a 10-ms step). The x-vector model is based on a deep neural network with TDNN layers that extract frame-level features and then synthesize the information into a fixed-dimensional representation. Then, a fully connected layer generates the final speaker feature encoding. ECAPA-TDNN focuses on frame-level feature extraction and improves on the feature synthesis level on the original x-vector model and its variants. In the original x-vector model, the temporal context layer is limited to 15 frames. To improve the feature extraction efficiency and reduce the number of model parameters, ECAPA-TDNN integrates the Res2Net module [94] (dimensionality reduction) with the SE block [38] to become SE-Res2Block during frame-level feature extraction. The input features of the ECAPA-TDNN model are 80-dimensional MFBEs from a 25 ms window with a 10 ms frame offset. The two-second MFBEs feature vectors are normalized through cepstral mean subtraction. To increase the amount of training data, thesis uses data augmentation techniques during training such as adding environmental noise.

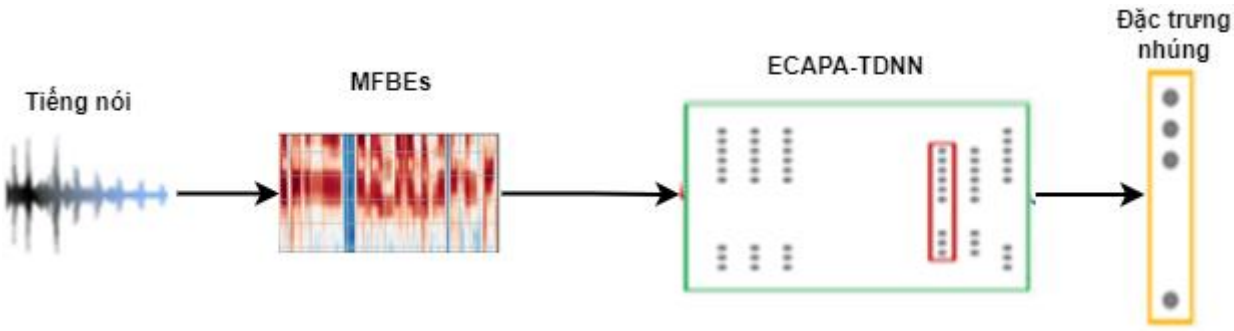

Figure 2.5: Proposed use of MFBEs features with ECAPA-TDNN model in Vietnamese speaker verification.

Verification phase

The verification phase will include the following steps:

- Extract features from input audio,
- Compare the embedded features of the audio to be authenticated with the registered audio,
- Based on the similarity between two embedded features to determine whether two audio clips come from the same person or not.

The evaluation scores are computed based on the Cosine distance [30] between the speaker embeddings (see Figure 2.5). All these scores are then normalized using the adaptive s-norm method [17], [51].

2.4. Experiment

Data set

VLSP2021-SV

In this chapter, thesis uses the VLSP2021-SV dataset by author Vi Thanh Dat et al. [18]. The dataset consists of short speech clips selected from interviews on YouTube. In these videos, the voices are all natural voices. The training dataset has more than 31,000 utterances from 1,305 speakers.

Vietnam-Celeb

In the experiment, thesis also used the Vietnam-Celeb dataset [83]. This dataset has more than 1,000 speakers with more than 87,000 utterances. The total recording time is about 187 hours, the recorded utterances are sampled at a frequency of 16,000 Hz.

Experimental environment

In the experimental part, the researcher evaluates the combination of two feature extraction methods with two deep learning models: The specific combination is as follows:

- ECAPA-TDNN model with MFBEs and MFCCs,
- ResnetSE-34 model with MFBEs and MFCCs.

Features: thesis uses MFCCs (80 dimensions) and MFBEs (80 dimensions) as input to the ECAPA-TDNN (or ResnetSE-34) model; the extracted features use a Hamming window of 25 ms

width, 10 ms step from the audio. The features in the training data are divided into 2 s lengths, then normalized. In the experiment, thesis uses data augmentation techniques.

Model architecture: in the ECAPA-TDNN model, the convolutional layers have 1024 channels. The nodes in the last connected layer have 192 dimensions, the total number of trained speakers is 1,305. In the ResnetSE-34 model, the nodes in the last layer are set to 512 dimensions, the total number of trained speakers is 1,305.

Thesis experiments are based on PyTorch. Models are trained on NVIDIA A100 GPU servers with 80GB of memory and Adam optimization. Thesis uses an initial learning rate of 0.001 and a 10% reduction after 2 epochs. The model is trained for 200 epochs with a mini-batch size of 100. Model training takes about 6 hours.

Experimental results and analysis

Comparison between ResnetSE-34 and ECAPA-TDNN models: Table 2.8 shows the error rates of different systems. The ECAPA-TDNN model has the smallest error rate of 11.37% (Vietnam-Celeb-E) and 12.74% (Vietnam-Celeb-H). That means the ECAPA-TDNN model has better performance than the ResnetSE-34 model under the same training and testing conditions.

Table 2.8: Experimental results of different features, evaluated on Vietnam-Celeb-E and VietnamCeleb-H datasets.

Model	Input data	Training data	Evaluation results (%EER)	
			Vietnam-Celeb-E	Vietnam-Celeb-H
ECAPA-TDNN [83]	80 MFCCs	VLSP2021-SV	11.58	14.3
ECAPA-TDNN	80 MFBEs	VLSP2021-SV	**11.37**	**12.74**
ResNetSE-34 [16]	80 MFCC	VLSP2021-SV	12.98	14.31
ResNetSE-34	80 MFBEs	VLSP2021-SV	**11.84**	**13.16**

Comparison between MFCC and MFBEs features: thesis analyzes the performance of different features and the results are listed in Table 21. In Table 21 , MFBEs feature gives better results than MFCC feature. The results shown in Table 21 are evaluated on Vietnam-Celeb-E and Vietnam-Celeb-H datasets. The systems in the evaluation condition learn speaker embedding features from MFBEs energy in short time and MFCC. Based on the experiment in Table 21, it can be seen that speaker embedding features modeled from MFCC give lower results than MFBEs under the same conditions. In particular, in the ECAPA-TDNN model (80 MFCCs), the error rates are 11.58% and 14.3%, respectively, while the ECAPA-TDNN model (80 MFBEs) is 11.37% and 12.74%. The DNN model learns speaker information from the MFBEs energy input data more effectively than MFCCs because it applies the discrete cosine transform to the MFBEs feature to create the MFCCs feature. Furthermore, with the challenges in speaker recognition, many researchers have also turned their attention to MFBEs instead of MFCCs. The main reason is that the MFCCs feature loses some information during the process of transforming MFBEs to MFCCs.

2.7. Conclusion of Chapter 2

In Chapter 2, the author has conducted research, analysis, and experimental comparison of the use of MFCCs and MFBEs features as input for two models ECAPA-TDNN and ResNetSE-34.
The main contributions of Chapter 2 include:

- Theory: Analyze, evaluate, and compare the differences between two feature extraction methods MFCCs and MFBEs using deep learning networks in the task of authenticating Vietnamese speakers;
- Experiment: The results show that the MFBEs feature with ECAPA-TDNN model gives good results for the MFCCs feature on two datasets Vietnam-Celeb-E and Vietnam-Celeb-H. This research result is extremely useful for further research in speaker verification and Vietnam-Celeb-H. This research result is extremely useful for further research in speaker recognition.

The results of the proposal to use MFBEs features as input for the ECAPA-TDNN model have been published in the works **[CT1], [CT2]** in the section "List of works by the author".

CHAPTER 3: IMPROVING THE ACCURACY OF VIETNAMESE SPEAKER VERIFICATION USING TRANSFER LEARNING WITH THE RAWNET3 MODEL

3.1. Introduction to transfer learning in speaker verification task

Transfer learning has been shown to be an effective approach in improving the overall performance as well as reducing the training cost of deep learning models. The efficiency of transfer learning is affected by a number of factors such as the amount of data available for the source and target tasks, and the choice of model.

Transfer learning has become a popular and effective technique for improving model generalization and reducing the cost of model training. In deep learning, transfer learning often involves using a pre-trained neural network, which is trained on a large dataset of a specific task and serves as a basis for training a new model for a related task. The transfer process is usually performed by using the weights learned from the pre-trained model as initial values and then fine-tuning these weights on new data. Transfer learning reduces the training time and improves the generalization ability of deep learning models. A typical example is the popular deep CNN model in image recognition such as VGG [101], ResNet [33], Inception [110]. These models are trained on the ImageNet dataset, a large image recognition dataset consisting of tens of millions of labeled images organized according to the semantic hierarchy of WordNet [20]. These pre-trained models are not only used in object detection [28], semantic segmentation, medical image classification but also used62 for speaker recognition [74], [122]. But that module and its variants are trained and implemented in popular libraries such as TensorFlow and PyTorch [81].

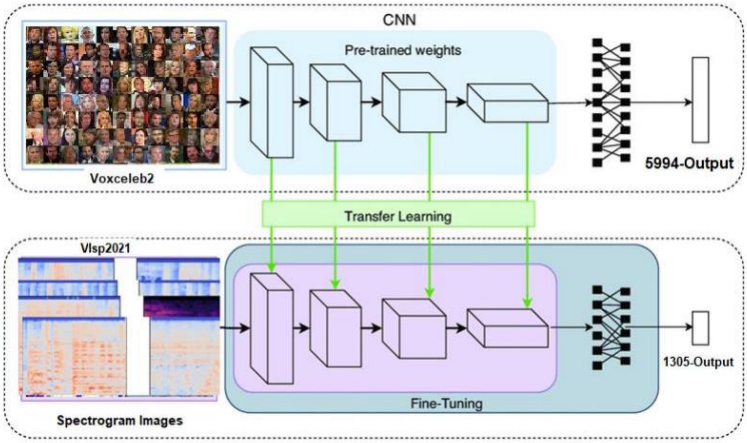

Figure 3.1: Transfer learning from pre-trained model on VoxCeleb2 dataset, fine-tuning the model on VLSP2021-SV dataset.

The diversity of pre-trained models stems from the fact that, first, every deep learning model has its strengths and weaknesses, and because no one knows in advance about the suitability of a particular model for a particular task. For example, VGG is famous for its simple structure and high performance, but it also has many parameters that make the model expensive to train and deploy. ResNet and Inception, along with their variants, are important architectures that help address different challenges in the field of deep learning. The advantage of the ResNet model is the ability to train deep networks efficiently, while Inception emphasizes learning diverse and optimal features. The limitations of these models revolve around computational complexity, architectural complexity, and potential challenges in model optimization. The second fact is that different neural network architectures are designed for different purposes or to solve tasks of a specific task. For example, MobileNet [36] combines depthwise convolution and pointwise convolution to reduce the number of computations and parameters while retaining essential features from the input data. The MobileNet architecture enables the deployment of deep learning models on devices with limited computational resources.

Choosing the right pre-training model is one of the key factors to ensure effective transfer learning.

Transfer learning can be particularly beneficial in situations where labeled data for speaker recognition is limited, as it allows knowledge to be leveraged from larger and more diverse datasets, potentially leading to improved performance and faster convergence during training.

Transfer learning with limited data resources

In the speaker verification task, pre-trained models are usually divided into two categories: supervised training models and self-supervised learning models [131]. Supervised training models are trained with large datasets of labeled speakers such as VoxCeleb2 [14], CnCeleb2 [64].

These models are often used to initialize speaker verification models ([130], [127], [128]) or apply domain adaptation techniques ([129], [93]) to transfer their robustness and generalizability to speaker verification situations with limited data resources. Speaker verification with limited data

resources involves developing speaker verification systems when data resources are limited, such as in the case of less common languages or remote speaker verification applications. Data resource limitations can be due to a small number of speakers, a lack of diversity in speech content, or an insufficient number of different acoustic environments in the dataset. To overcome the data shortage task, a simple approach is to augment the audio data from low-resource sources. Common methods include noise injection [104], echo [29], rate change [123], and SpecAug [79]. Some generative methods have also been applied to perform augmentation for speaker verification with limited data [120], [128]. In this chapter, we investigate the potential of large pre-trained models ResNetSE-34, ECAPA-TDNN, RawNet3 (on VoxCeleb2 data [14]) to improve speaker verification accuracy with small amount of data.

3.2. Data selection for speaker verification task

VoxCeleb2 [14] is one of the largest datasets commonly used for speaker recognition and speaker verification tasks. It is an extension of the original VoxCeleb1 [72] dataset and is designed for research in speaker recognition, speaker verification and related fields. VoxCeleb2 has been widely used in developing and evaluating various speaker recognition models and algorithms. This dataset has been described in detail in Chapter 1 of the thesis.

Table 3.1: Summary of three public datasets for the speaker verification task. VoxCeleb2 is used to pre-train three deep learning models and the VLSP2021-SV dataset is used to fine-tune and evaluate the models.

Dataset	Language	Recording hours	Training set # of utterances (# of speakers)	Test set # of test pairs (# of speakers)
VoxCeleb2 [14]	Mostly English	2,300	1,128,246 (6112)	37,720 (40) 581,480 (1251) 552,536 (1190)
CN-CELEB2 [64]	China	1,090	529,485 (2000)	3,586,776 (-)
VLSP2021-SV [18]	Vietnam	41	31,600 (1305)	20,000 (114)

Besides English, several datasets have been built for other languages such as Chinese and Vietnamese. For example, CN-CELEB1 [25] is a large speaker recognition dataset collected 'in the wild' containing more than 130,000 utterances of 1,000 Chinese celebrities. The published results of the research group [25] show that the performance of the latest speaker recognition methods is far inferior to that of the VoxCeleb2 research group [14]. Recently, the VLSP 2021 workshop has published a validation and recognition dataset, this dataset is VLSP2021-SV. Some statistics of the three datasets are reported in Table 3.1. We can see that VoxCeleb2 is currently much larger than the two Vietnamese and Chinese speaker recognition datasets. It enables the development and evaluation of advanced algorithms, and contributes to the development of this field by providing a large and diverse collection of audio recordings. In particular, VoxCeleb2 is used as a pre-training data source for transfer learning. Researchers and developers can fine-tune pre-trained models on VoxCeleb2 with a small amount of labeled data from their own application, effectively adapting the new model to the specific task.

3.3. Pre-trained models for transfer learning

ECAPA- TDNN uses frame-level acoustic feature arrays similar to traditional i-vectors, while VGGVox uses 2D spectrograms and RawNet uses raw speech signals as input. Different models use different frame-level feature extraction methods to represent individual speaker features or fix the number of dimensions of speaker embedding vectors in speaker recognition and authentication tasks .

VGGVox

VGGVox uses 2D short intensity spectra as input to the deep CNN network. For example, the spectra of size (512,300) are generated from a 3-s speech signal segment with a sliding window using a Hamming window of 25 ms width and 10 ms stride.

In the paper (Nagrani et al., 2019), the author extracts speech segment features with four deep CNN backbones: VGG -M, ResNet-34, ResNet-50 and Thin-ResNet. In the training phase, VGGVox randomly takes 3s segments in each training data sentence. In the testing phase, different strategies are used to evaluate the performance of the speaker verification task.

RawNet3

RawNet3 is an improved model from RawNet2 [47] and ECAPA-TDNN [22]. As illustrated in Figure 3.5, RawNet3 consists of a parametric analysis filter bank layer (ParamFbank) [78] and a Res2Dilated block with α feature map expansion and max pooling (AFMS-Res2MP). When a unidirectional wave signal is fed into RawNet3, it is transformed into a bidirectional feature map by the ParamFbank layer. The ParamFbank layer learns real-valued parameterized filters, which are extended from the SincNet layer [89]. The feature map, after being processed through the ParamFbank layer, will be passed through three AFMS-Res2MP blocks to capture speaker information at different levels.

The main idea behind RawNet is to use a convolutional neural network (CNN) directly on raw audio, similar to how CNNs are applied to images. One of the main advantages of RawNet is the ability to learn discriminative features directly from the raw audio signal, which can help improve performance over traditional feature extraction.

In the improved versions, RawNet2 [42] and RawNet3 [50] use a series of advanced frame-level feature extraction techniques such as using the Res2Net frame module and scaling the feature map with important weights or focusing on important features in the frame-level feature map. In addition, the first convolution layer is replaced by a sinc-conv layer for better results.

All three models were trained on the VoxCeleb2 dataset (37,720 pairs of utterances from 40 speakers ; 1,092,009 utterances from 5,994 speakers).

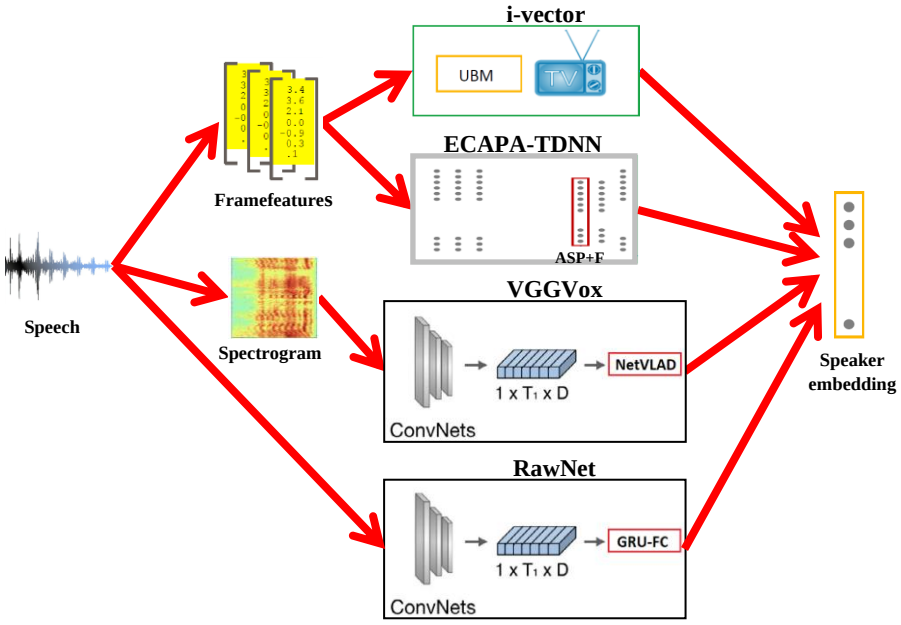

Figure 3.2: Illustration of the difference between the traditional i-vector method and three deep learning models for speaker modeling.

Table 3.4: Comparison of differences between models.

Model	Number of parameters (M)	Input data	Number of dimensions of embedded features	Error Rate (%)
VGGVox (ResNet-34) [16]	22.0	2D spectrum image	512	2.22
ECAPA-TDNN [22] (C=512)	6.2	1D array of MFCCs	192	1.01
ECAPA-TDNN [22] (C=1024)	14.9	1D array of MFCCs	192	0.87
RawNet2 [47]	13.2	59,049 samples	1024	3.00
RawNet3 [50]	16.3	59,049 samples	256	0.89

Table 3.4 shows that with the same raw input data, RawNet3 achieves an EER of 0.89%, which is better than RawNet2's 3.00%. This improvement is due to the integration of the strong features of the ECAPA-TDNN model. These features include redundantly connected backbones, multi-layer feature pooling, replacing GRUs with channels and context-dependent statistical pooling, and using AAM-softmax objective functions during training. The number of speaker embedding features of RawNet3 is 256, much smaller than RawNet2's 1024. The number of embedding features of ECAPA-TDNN is 192 and that of ResNet-34 is 512. ECAPA-TDNN achieves an EER of 0.96%, which is slightly worse than RawNet3 but much better than RawNet2. This result shows the effectiveness of the ECAPA-TDNN model with 80-dimensional MFBEs input features. VGGVox (ResNet-34) also has an EER error rate of 1.18%, which is not as good as RawNet3. It may be due to the conversion of 1D speech signals into 2D spectral images, then applying the feature extraction backbone that has been proven effective in the image field. The application of modern machine learning techniques on

VoxCeleb uses NetVLAD to synthesize frame-level features, combining Cosine similarity comparison and classification for the authentication process.

3.4. Experiment

Data set

VLSP2021-SV

In this experiment, thesis uses the VLSP2021-SV dataset (detailed in Chapter 1) by Vi Thanh Dat et al. [18]. The dataset consists of short speech segments selected from interviews on YouTube. In these videos, the voices are all natural voices. The training dataset has 31,600 utterances from 1,305 speakers.

Fine-tuning on pre-trained model

The general goal of transfer learning is to leverage existing knowledge or data to improve the performance of a speaker verification system. The simplest way is to reuse a pre-trained model and extract high-level features from the input audio, creating an embedded feature. These features can be used as input to a speaker verification model or other tasks. In this section, thesis fine-tunes the pre-trained models with the same VLSP2021-SV development dataset, replacing the last classifier layer with the number of speakers in the VLSP2021-SV dataset. Specifically, the following settings are applied in the experiments :

- ResNetSE-34: thesis uses MFBEs with 64 dimensions, window length 25ms, step 10ms, FFT size 512, frequency domain limit 20-7600 Hz;
- ECAPA-TDNN: thesis uses 14.85 million parameters without classification, the model input is 80-dimensional MFCCs feature. Other parameters are the same as the original ECAPA-TDNN model [22].
- Rawnet3: model has 16.28 million parameters, using a sliding size of 10 for the parameterized analysis filter layer. The model uses raw audio as input data.

models are trained on NVIDIA Telsa P100 GPUs, 16GB of memory, using Adam optimization, initial learning rate is 0.0001. Each model is trained for 500 epochs, learning rate is reduced by 5% after 10 epochs. Thesis uses online data augmentation techniques during training. Thesis uses MUSAN [104] and RIR[59] datasets to supplement training data.

Experimental results

In Table 3.6, thesis presents the experimental results of three deep learning models for the speaker verification task with limited resources: without transfer learning and with transfer learning. It can be seen that the RawNet3 model, although inherited from the two ECAPA-TDNN models and the RawNet2 model, has an error rate of 4.07%, which is not better than the ECAPA-TDNN model with an error rate of 3.92% (models trained from scratch). With the transfer learning technique, RawNet3 achieves an error rate of 1.61% on the VLSP 2021 evaluation set. This is a large difference when compared to the error rate of 2.21% of the ECAPA-TDNN model. Transfer learning also helps improve the performance of VGGVox (ResnetSE-34) when the error rate decreases from 5.98% to 2.22%, very close to the 2.21% of ECAPA-TDNN. From the methodological comparison and analysis in Section 3.3 and the experimental results in this section, we can argue that converting raw speech

signals to 2D spectral images and then applying conventional deep CNN models for image recognition may not be an ideal approach for the speaker verification task. VGGVox has experimented with different backbone networks such as VGG, ResNets, frame-level feature synthesis methods, recognition and authentication enhancement techniques. However, VGGVox still lags behind ECAPA-TDNN and RawNet on the VoxCeleb2 and VLSP2021-SV datasets. The reason may be that the conventional 2D convolution operator does not design the projection of variable-length utterances to fixed-length embedded features like TDNN in ECAPA-TDNN or the context-dependent statistical synthesis in both ECAPA-TDNN and RawNet3. In Table 3.6 MFBEs(2D) means the MFBE feature is used as input to the 2D convolutional neural network. The MFBE coefficients are arranged in a two-dimensional matrix, where one dimension is time and the other dimension is the MFBE coefficients. This matrix can be considered as an image with different color channels, and is used as input to the Conv2D layer. MFBEs(1D) means the MFBE feature is used as input to the Conv1D convolutional neural network. The MFBE coefficients are arranged in a two-dimensional matrix, where one dimension is time and the other dimension is the MFBE coefficients. This matrix is used as input to the Conv1D layer. The Conv1D will apply 1D filters along the time axis of the MFBE matrix. The one-dimensional acoustic vector array input of ECAPA-TDNN shows its superiority for training speaker verification models with small datasets. ECAPA-TDNN achieves an error rate of 3.92% when trained on only 41 hours of audio (VLSP2021-SV) while RawNet3 only achieves an error rate of 4.07%. On the larger VoxCeleb2 dataset (2,300 hours of audio), both RawNet3 and ECAPA-TDNN perform well with error rates of 0.87% and 0.89%. This comparison means that RawNets requires more training data and it performs better than ECAPA-TDNN. On the other hand, transfer learning in the raw speech signal domain is proven to be more effective than the acoustic feature domain. Transfer learning helps ECAPA-TDNN reduce the error rate from 3.92% to 2.21% while RawNet3 reduces the error rate from 4.07% to 1.61%. The reason may be that the acoustic feature extraction used in ECAPA-TDNN depends on general parameters such as the dimensionality of MFCCs, window size, and stride width. Optimizing these parameters on one dataset (multiple nationalities and dialects in VoxCeleb2) is not a good choice for another dataset (e.g. Vietnam). The RawNet model does not require hyperparameters, it automatically learns frame-level features from raw signal data.

Table 3.6: Performance comparison of three deep learning models for the Vietnamese speaker verification task. (No transfer learning and transfer learning).

Model	**Input characteristics**	**No transfer learning**	**Transfer learning**	**Number of parameters (M)**	**GFLOPs**
		SV EER (%)	SV EER (%)		
ResnetSE-34	MFBEs 64 (2D)	5.98	2.22	22.0	3.82
ECAPA-TDNN	MFBEs 80 (1D)	3.92	2.21	14.9	3.96
RawNet3	Raw waveform	4.07	1.61	16.3	8.13

ECAPA-TDNN and RawNet3 have higher computational efficiency than ResnetSE-34 due to their special design for tasks such as speech recognition and raw audio processing. The FLOPs value

depends on the specific network architecture including the number of layers and layer sizes, kernel size, and input data size. The input data of ECAPA-TDNN and RawNet3 models is raw audio which is usually higher in dimension than the input of ResnetSE-34 (2D array – 64 MFBEs). Processing high-dimensional data requires more computational resources, which also contributes to higher FLOPs. Table 3.6 shows that RawNet3 has the highest FLOPs of 8.13G. This is because the input of RawNet3 model has the largest dimension (59,049 samples) while the other two models are ResnetSE-34 (64 MFBEs) and ECAPA-TDNN (80 MFBEs).

ECAPA-TDNN and RawNet3 have higher computational performance than ResnetSE-34 due to their special design for tasks such as speech recognition and raw audio processing. The FLOPs value depends on the specific network architecture including the number of layers and their sizes, kernel size, and input data size. The input data of ECAPA-TDNN and RawNet3 models is raw audio which is usually higher in dimension than the input of ResnetSE-34 (2D array – 64 MFBEs). Processing high-dimensional data requires more computational resources, which also contributes to higher FLOPs. Table 25 shows that RawNet3 has the highest FLOPs of 8.13G. This is because the input of RawNet3 model has the largest size (59,049 samples) while the other two models are ResnetSE-34 (64 MFBEs) and ECAPA-TDNN (80 MFBEs).

3.5. Conclusion of chapter 3

The researchers studied the effectiveness of three deep learning models for the Vietnamese speaker verification task. Experimental results showed that ECAPA-TDNN performed well with a small training dataset. The model achieved an error rate of 3.92% when trained on a dataset of 31,600 utterances (VLSP2021-SV dataset), compared to 4.07% for RawNet3 and 5.98% for VoxCeleb. However, when the models were trained on a much larger dataset of 1,128,246 utterances (VoxCeleb2), fine-tuning RawNet3 reduced the error rate to 1.61%. Transfer learning also helped the system using the ResnetSE-34 model achieve an error rate of 2.22%, close to the 2.21% error rate of ECAPA-TDNN. From the comparison and analysis of the three models, thesis concluded that the strength of RawNet3 lies in inheriting the characteristics of ECAPA-TDNN, such as frame-level feature synthesis and the effective combination with different objective functions. The raw input data also makes Rawnet less dependent on acoustic feature extraction and takes advantage of the diversity of training data. Several directions can also be tested in the future. First, the acoustic properties of different spoken languages have not been thoroughly studied. For example, Vietnamese is a tonal language and has six tone types, while English does not. English uses pitch and intonation, not tones like Vietnamese. Fine-tuning a pre-trained model on the same tone language can benefit the task of Vietnamese speaker verification. The second direction is to apply the architecture that has been successfully demonstrated in other fields such as transformer [115] in natural language processing to synthesize frame-level features to better represent speakers. The application of other preprocessing techniques such as speech segment detection and acoustic feature selection also helps to improve the performance of any speaker verification system. The results of the Proposal to use RawNet3 model in transfer learning for speaker verification task with limited resources are published in [**CT3], [CT4]** in the section “List of works of the author”.

CONCLUSION AND DEVELOPMENT DIRECTION

Main results of the thesis

The thesis presents a systematic study on the task of speaker verification for speech with limited resources, which is an important task in the field of natural language processing. The thesis focuses on researching and proposing features with deep learning models to improve the accuracy of the Vietnamese speaker verification system. In addition, the thesis also studies and proposes a modern deep learning model in transfer learning for the task of speaker verification with limited resources. The research results of the thesis can be summarized as follows:

1. **Proposing the use of MFBEs features with ECAPA-TDNN model for Vietnamese speaker verification task. ECAPA-TDNN model (80 MFBEs) has reduced EER error rate by 11.37% (compared to 11.58%) and 12.47% (compared to 14.3%) respectively ([CT1], [CT2])**
2. **Proposed use of RawNet3 model in transfer learning for speaker verification task with limited resources. RawNet3 model gives the best result with EER error rate of 1.61%. ([CT3], [CT4])**

Limitations of the thesis

1. The database for Chinese and the training model on this data are also widely published. Currently, thesis has no experiments or evaluations when applying it to Vietnamese data.
2. The researcher also needs to test and evaluate the combination of MFCCs and MFBEs features for the speech authentication task in both English and Vietnamese. Development direction: Test MFBEs/LPCCs features or combine MFCCs to improve SV accuracy; Use the available training model from Chinese data, then fine-tune on Vietnam-Celeb-T.

Further research directions

From the results achieved in the thesis, the following issues need to be researched in the coming time:

1. Testing MFBEs/LPCCs or MFCCs combination improves speaker verification system accuracy;
2. Use the training model on Chinese data, then fine-tune on the Vietnam-Celeb-T dataset.

LIST OF THE PUBLICATIONS RELATED TO THE DISSERTATION

1. [CT1] T. -T. -M. Nguyen, D. -D. Nguyen and C. -M. Luong, "Vietnamese Speaker Verification with Mel-scale Filter Bank Energies and Deep Learning," *in IEEE Access*, doi: 10.1109/ACCESS.2024.3479092. **(SCIE, Q1)**
2. [CT2] Nguyễn Thị Thanh Mai, Nguyễn Đức Dũng, "Kết hợp đặc trưng MFCCs và Mel-Filter Bank Energies trong xác thực người nói tiếng Việt", *VNICT-2024*, Tr. 288-293.
3. [CT3] Thi-Thanh-Mai Nguyen, Duc-Dung Nguyen, Chi-Mai Luong (2024). "Transfer Learning for Vietnamese Speaker Verification." *Vietnam Journal of Science and Technology*. (**Accepted**) **(SCOPUS, Q4)**
4. [CT4] Mai Nguyen Thi Thanh, Dung Nguyen Duc, "Vietnamese Speaker Verification based on ResNet model", *VNICT-2023*, pp 377-381.