MINISTRY OF EDUCATION
AND TRAINING

VIETNAM ACADEMY OF
SIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SIENCE AND TECHNOLOGY

PHAM NGOC PHUONG

RESEARCH AND DEVELOP A VOICE ADAPTATION SYSTEM FOR VIETNAMESE SYNTHESIS AND ITS APPLICATIONS

SUMMARY OF DISSERTATION ON INFORMATION SYSTEM
Code: 9 48 01 04

Ha Noi – 2023

The thesis has been completed at: Graduate University of Science and Technology- Vietnam Academy of Science and Technology

Supervisor: Assoc. Prof. Dr . Luong Chi Mai

Reviewer 1: …
Reviewer 2: …
Reviewer 3: ….

The thesis shall be defended in front of the Thesis Committee at Vietnam Academy Of Science And Technology - Graduate University Of Science And Technology, at …..… hour……, date…… month…..…year 2023

This thesis could be found at:

- The National Library of Vietnam
- The Library of Graduate University of Science and Technology

LIST OF THE PUBLICATIONS RELATED TO THE DISSERTATION

1. Pham Ngoc Phuong, Tran Quang Chung, Luong Chi Mai: “Adapt-TTS: High-quality zero-shot multi-speaker text-to-speech adaptive-based for Vietnamese”. *Journal of Computer Science and Cybernetics*, V.39, N.2 (2023), pp. 159-173. 1-DOI: 10.15625/1813-9663/18136, VietNam.
2. Pham Ngoc Phuong, Tran Quang Chung, Luong Chi Mai: “Improving few-shot multi-speaker text-to-speech adaptive-based with Extracting Mel-vector (EMV) for Vietnamese”. *International Journal of Asian Language Processing,* 2023, Vol. 32, No. 02n03, 2350004, pp. 1-15, Singapore.
3. Pham Ngoc Phuong, Tran Quang Chung, Do Quoc Truong, Luong Chi Mai: “A study on neural-network-based Text-to-Speech adaptation techniques for Vietnamese”, *International Conference on Speech Database and Assessments (Oriental COCOSDA)* 2021, pp. 199-205. IEEE, Singapore.
4. Pham Ngoc Phuong, Tran Quang Chung, Nguyen Quang Minh, Do Quoc Truong, Luong Chi Mai: “Improving prosodic phrasing of Vietnamese text-to-speech systems”, *Association for Computational Linguistics, 7th International Workshop on Vietnamese Language and Speech Processing*, 12/2020, pp. 19-23, VietNam.
5. Nguyen Thai Binh, Nguyen Vu Bao Hung, Nguyen Thi Thu Hien, Pham Ngoc Phuong, Nguyen The Loc, Do Quoc Truong, Luong Chi Mai: “Fast and Accurate Capitalization and Punctuation for Automatic Speech Recognition Using Transformer and Chunk Merging”, *International Conference on Speech Database and Assessments (Oriental COCOSDA)* 2019, IEEE, pp. 1-5, Philippines.
6. Pham Ngoc Phuong, Do Quoc Truong, Luong Chi Mai: "A high quality and phonetic balanced speech corpus for Vietnamese" *International Conference on Speech Database and Assessments (Oriental COCOSDA)* 2018, pp. 1-5 Japan.
7. Certificate of copyright registration "Adapt-TTS text-to-speech conversion software" No. 7590/QTG issued on September 26, 2022 at the *Copyright Office*

PREAMBLE

When researching speech synthesis, one of the most exciting topics today is controlling and adapting speech characteristics to create synthetic speech according to arbitrary style and intonation. Typically, to build a synthetic voice with the characteristics of a specific speaker, it is necessary to record a large amount of data (about 10 hours in a standard studio environment) of that voice for training. That makes creating synthetic voices on demand costly, time-consuming, and complex in resource-poor languages like Vietnamese. Furthermore, currently, voice synthesis has higher requirements than just using existing voices, such as the need to build a unique voice, a personalized voice, or the need to restore the humanized voice and voice cloning. Research on adjusting and transforming voice characteristic parameters and adapting to speakers has mainly been applied in research by foreign authors on popular languages such as English and Japanese. , Central. In Vietnam, these studies still approach the adaptive synthesis method based on HMM and show low synthesis quality.

Research question: What method helps synthesize speech to ensure quality for a resource-poor language like Vietnamese while only having a few minutes of adaptive samples? How much is adaptive data (trained with the system) needed at least to ensure the high quality and similarity of the synthesized voice? If adapting with data samples takes only a few seconds and does not require retraining the model, can the system do it, and what is the minimum amount of adaptive samples needed?

The main goal of the thesis: research and build a Vietnamese speech synthesis system using adaptive training techniques for the speaker's acoustic characteristics based on DNN to 1) Improve the synthesis quality of Adaptive-based voice with naturalness enhancements; 2) New speech synthesis carries the acoustic characteristics of the target voice with high quality and similarity while using only a small amount of sample data; 3) Instant speech synthesis with tiny samples without retraining costs.

Thesis contributions: 1) Proposing two speaker-dependent adaptive synthesis models based on DNN with the condition of little training sample data but creating the best possible new voice *(from now on, the thesis is called summarize this concept with the term Few-shot TTS)*: i) Speaker-dependent adaptive synthesis model based on transfer-learning; ii) Speaker-dependent adaptive synthesis model based on feature representation vector; 2) Proposing a speaker-independent adaptive synthesis model based on DNN with the condition that only a few sample sentences are needed without retraining the model but still creates an acceptable new voice *(from now on the thesis abbreviate this concept with the term Zero-shot TTS)*; 3) Build a Vietnamese language database (database) that ensures quality as a base data set for the task of training synthetic and adaptive models. Low-cost database construction method and labeling

improvements to increase naturalness; 4) Build a multi-speaker adaptive application for multi-platform devices.

Subject and scope of research of the thesis: *The Vietnamese speech synthesis system can be personalized using the adaptive method under a limited number of adaptive samples with training and without retraining*. The research will develop voice imitation or restoration applications integrated or run on cross-platform computing platforms. Training data and sample data (target voice) are selected to be limited to Northern and Southern accents with the style of reading news information on political and social topics.

The thesis structure includes the following parts:

Chapter 1: Overview of speech synthesis and speech synthesis with the ability to adjust output characteristics. The general structure of a basic adaptive speech synthesis system. Overview of research on speech synthesis based on adaptation in general and Vietnamese adaptation in particular. Introduce the main goals and research scope of the thesis.

Chapter 2: Building a Vietnamese database for synthesis and adaptation systems and accompanying processes to improve quality and reduce costs when building multi-speaker databases for systems Vietnamese synthesis. Besides, adding label information, such as inserting breath stops and transcribing borrowed words, helps increase the naturalness of the synthesis model. This language database and information label enhancement technique are also the basis for building adaptive models in the following chapters.

Chapter 3: Presents a method to improve the quality of the speech synthesis model based on adaptation through two proposals: 1) Improve the adaptive synthesis model (Few-shot TTS) using Muli-pass fine-tune based on speaker and language transfer learning techniques (transfer-learning) with much fewer samples to learn than training the base model and 2) Improved adaptive synthesis model (Few-shot TTS) using the EMV vector to represent speaker characteristics with just a few sentences. Both adaptive techniques require sample data to be present in the training set, and the proposed models aim to use less and less adaptive data.

Chapter 4: Proposing a method to improve the performance of a low-cost adaptive synthesis model with the least possible sample conditions without retraining the model (Zero-shot TTS) through two techniques: 1) Apply effective speaker feature vector representation; 2) The Mel-spectrogram denoiser model allows for higher quality sound synthesis compared to baseline models. The Zero-shot TTS adaptation-based synthesis model does not require adaptation data to be in the training set and only uses a single sample sentence of the speaker for adaptation. This approach simplifies the synthesis of new voices and expands the applicability of adaptive synthesis models.

Conclude: Present the main contributions of the thesis and point out limitations and future development directions.

CHAPTER 1. RELATED RESEARCH AND BASIC KNOWLEDGE OF SPEECH SYNTHESIS AND ADAPTATION

In Chapter 1, the first part introduces an overview of related research on speech synthesis systems and the complex problems that need to be solved. Next, the need for speech synthesis with the ability to adjust output characteristics and related research on adaptive speech synthesis and applications are presented. Then, the basic knowledge and principal components of an adaptation-based synthesis system are described as assessments of the quality of adaptation-based synthesis, an overview of the domestic and foreign research situation, and finally, the thesis's main research directions and scope.

1.1. Speech synthesis

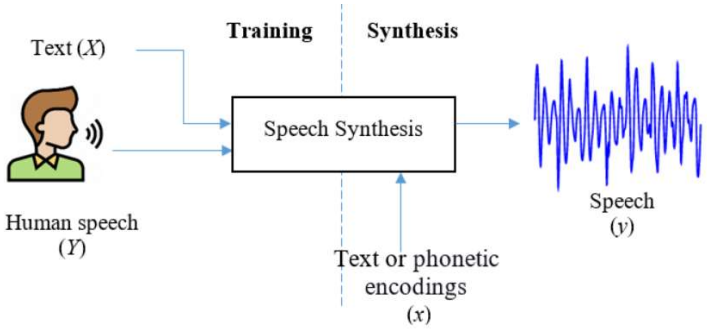


Figure 1: Artificial speech synthesis model

The concept of speech synthesis

Speech synthesis is the artificial generation of human speech from text input or phonetic encodings. Speech synthesis is part of the field of natural language processing.

Speech synthesis from text (Text to speech - abbreviated as TTS) is an essential technology in speech synthesis; this technology arbitrarily creates output speech sound waves from input text.

The TTS system can be described using a model that calculates the prediction distribution probability:

$$p(y|x, Y, X)$$

where Y is the speech sound used for training, X is the corresponding labeled text, x is the input text, and y is the speech to be synthesized.

1.1.1. Classification of speech synthesis methods

Two popular TTS architectures. The most advanced today are: *1) Autoregressive architecture and 2) Non-autoregressive architecture*. Each architecture has different advantages and disadvantages.

1.1.2. Speech synthesis with the ability to adjust output characteristics

DNN techniques have completely replaced HMM models in building acoustic and duration models. DNN models require only a single computation for feature prediction, making it more suitable for real-time synthesis. However, current

research on DNN modeling focuses mainly on speaker-dependent modeling, which requires a significant amount of data from a single speaker to create a stable acoustic model. Therefore, DNN-based multi-speaker adaptive methods have been proposed.

1.2. Adaptation in speech synthesis

1.2.1. Concept

Adaptive speech synthesis (or *'adaptive TTS'*) is the ability to synthesize arbitrary speech from any person with a small amount of real sample data (*reference speech*). , the synthesized voice will have the characteristics of the target speaker with voice characteristics and intonation features (*prosodic features*). TTS adaptation is referred to by different terms in academia and industry, such as *voice adaptation*, *voice cloning*, and *custom voice*.

The source and destination speech signals as X and Y represent the source and destination speech characteristics as x and y; the conversion function can be built as follows: y= f(x) in That f(.) is also known as the frame mapping function.

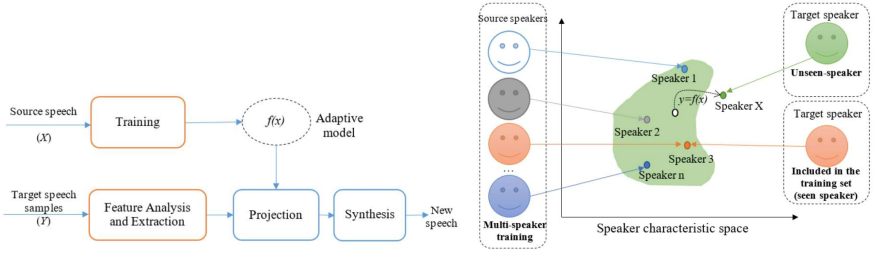


Figure 2: General model and characteristic space of adaptive-based TTS system

Adaptive TTS is considered from two perspectives: *1) General adaptation setup*, including improvements in generalization of the source TTS model to support new voices; *2) Effective adaptation*, including reduction of adaptation data and adaptation parameters.

1.2.2 General adaptation

1.2.3. Efficient Adaptation

The goal of adaptation is to use as little data and parameters as possible while still achieving good quality. Research in this category can be divided into several groups: 1) Few-data adaptation; 2) *Few-parameters adaptation; 3) Untranscribed data adaptation; 4) Zero-shot TTS* (Adaptation without retraining the model).

1.3. Current research status on adaptive synthesis

1.3.1. Some recent studies on some other languages

There are two main approaches in this direction: 1) Adaptation through fine-tuning part of the model or the entire model and 2) Adaptation based on speaker feature vector representation. Some studies have constructed few-shot adaptations by using only a few pairs of text and voice data, varying from a few minutes to a few seconds. Chien and colleagues explore several different types

of speaker embedding for few-shot adaptation. Yue and colleagues leveraged speech sequences for few-shot adaptation. Chen et al. and Arık et al. compared speech quality with different amounts of adaptation data and found that speech quality improves rapidly with the increase of adaptation data when the data size is small (less than 20 sentences) and improves slowly with dozens of adapted sentences.

- Adaptation with few parameters Some works propose reducing the adaptation parameters to as few as possible while maintaining the adaptation quality. AdaSpeech proposes conditional layer normalization to generate scale and offset parameters in layer normalization from embedding speakers based on contextual parameter generation and fine-tune only parameters related to layer normalization that have conditions and speaker embedding to achieve adaptive quality. Moss et al. proposed a fine-tuning method that selects different model hyperparameters for various voices based on Bayesian optimization.

- Zero-shot adaptation. Some studies conduct zero-shot adaptation, using speaker embedding to extract the speaker embedding of sample sounds. This scenario is quite attractive because there is no need for data and adaptation parameters. However, the adaptation quality needs to improve, especially when the target speaker is very different from the source speaker.

1.3.2. End-to-end studies for Vietnamese synthesis

According to VLSP statistics, in the four years of evaluation, in 2018, DNN models gained dominance in VLSP; in 2019-2020, research groups focused on using Tacotron2 models to dominate and developing Vietnamese synthesis systems (with acoustic modeling using Tacotron2 combined with popular Vocoders such as Waveglow or HifiGAN). Also, according to research from this organization, from 2021 onwards, Vietnamese synthesis research groups have focused on using the Fastspeech2 model to dominate, and some groups proposing to use VITS have also achieved outstanding results. End-to-end research for Vietnamese synthesis has been entirely up to date with international research; however, there are challenging issues in research on speech synthesis with Vietnamese characteristics, such as building a specialized database for synthesizing Vietnamese.

1.3.3. Some current research on adaptive synthesis for Vietnamese

Adaptive technique based on HMM: SonPT and NinhDK use the HMM hybrid model for Vietnamese synthesis and to adjust adaptive parameters when synthesizing; the study proposed to apply a likelihood linear regression algorithm (MLLR) combined with maximum a posteriori algorithm (MAP) or combined MAP with vector field smoothing (VFS) algorithm. SonPT has shown that with only a limited amount of "target" data (100 sentences), combined with a standard HMM set of many voices, many different characteristics, and speaking styles, it is possible to synthesize speech. have improved quality. However, the adaptive synthesis approach based on HMM statistical parameters has inherent disadvantages:

1) Models based on HMM give much lower synthesis quality than DNN; 2) Current studies show that HMM-based adaptive synthesis techniques provide lower quality than DNN techniques; 3) Adaptive models based on HMM cannot perform adaptive synthesis tasks with minimal data (only a few sentences) or adapt without retraining.

1.4. Conclusion of Chapter 1 and main research contents of the thesis

Adaptation-based speech synthesis is a problem in speech conversion and adaptation to transform the new speech synthesis result with the characteristics of the sample voice. Vietnamese is a low-resource and complex language because it contains intonation components, so improving the quality of speech synthesis and adaptation is still a complex problem that more people need to solve. Therefore, there still exist difficulties such as speech synthesis and adaptive conversion problems in terms of synthesis quality and training sample cost for Vietnamese, which can be mentioned as follows: 1) Studies on Adaptive synthesis for Vietnamese are still minimal; there is a need for studies to evaluate the impact of limited or no target speech samples in the training process; 2) Research on adaptation for Vietnamese has only used the HMM model for training on limited sample data and there is no adaptive model for Vietnamese using DNN; 3) There have been no studies applying the End-to-End adaptive model to Vietnamese with small sample data with or without training; More research is needed on improving the quality of synthesis and adaptation of the Vietnamese language; 4) There is a need for applications to evaluate the feasibility of Vietnamese speech adaptation models.

From the above practical issues, the thesis will focus on researching some main contents as follows: 1) Building a database to serve synthesis and adaptation; 2) Research adaptive Few-shot TTS technique based on DNN for Vietnamese and evaluate; 3) Research and evaluate the zero-shot TTS adaptive technique based on DNN for Vietnamese.

Scope of research: The research object is the Vietnamese language. The research model is an adaptive synthesis of Vietnamese to personalize the synthetic voice, specifically, voice cloning with single-speaker and multi-speaker data. A voice cloning application that evaluates feasibility is a voice cloning application.

CHAPTER 2. BUILDING A LOW-COST VIETNAMESE DATABASE FOR SPEECH SYNTHESIS AND ADAPTATION

One of the main tasks of the thesis when researching adaptive-based speech synthesis is to study the Vietnamese speech synthesis model. However, the most significant limitation in synthetic research on Vietnamese speech is the need for an extensive database to ensure quality and low cost for synthetic research. In addition, one of the remaining problems of the Vietnamese synthesis system is the naturalness of long sentences and reading borrowed words. Chapter 2 analyzes techniques for building a language database for adaptive synthesis and additional labeling and transcription methods to improve intonation for the Vietnamese speech synthesis system; the contents include: 1) The presentation presents an analysis of current databases for speech synthesis; 2) Present the process of building a quality-assured language database for synthesis and adaptation CT6] [CT4]; 3) Some methods of adding label information to increase the naturalness of the Vietnamese TTS system through techniques such as adding punctuation, inserting breath stops, and transcribing borrowed words [CT5] [CT4]; 4) Results of building a database.

2.1. Build a comprehensive and adaptive database

End-to-end research for Vietnamese synthesis has been quite up-to-date with international research. However, studies also highlight challenges in speech synthesis research with Vietnamese characteristics, such as building a specialized database for Vietnamese synthesis, problems synthesizing long sentences, or reading borrowed words.

2.1.1. Statistics of database sets for current synthesis

*** Some open Vietnamese databases for trainees have been announced:**

Corpus	Time (hours)	Speakers	Utterance	Reading style
VLSP 2020	9.5	1	7,770	Read stories with different intonations
VAIS1000	0.2	1	1,000	VOV announcer
INFORE	25	1	14,935	Read stories using TTS
VietTTS-v1.1	35.9	1	22,884	Read stories using TTS

In addition, it can be seen that in Vietnam, there are many large corporations and research units, but the databases serving TTS research are very lacking and limited in both quality and diversity. From that reality, the thesis determines the urgency of building a Vietnamese voice database that ensures quality, low cost with diverse voices (gender, age, living area), and Clear strategy, serving research synthesis and voice adaptation.

When studying speech synthesis or adaptation, creating a model for a specific speaker with limited data is challenging, even for resource-rich languages. For Vietnamese, it is difficult to conduct such a study due to data design requirements:

1) The sound must be recorded in a soundproof, noise-free studio to ensure quality recording; 2) Speakers must be chosen to cover a wide range of ages and

areas of residence; 3) It is necessary to select to ensure acoustic balance (enough phonemes and vocabulary in the voice); 4) Thus, a database good enough to train a speech synthesis and adaptation system must be a selective combination of collecting a database from available sound sources and building a self-recorded database.

2.1.2. Process of building a database for synthesis and adaptation

To ensure that the database for synthesizing sounds is large enough and diverse, it is necessary to build a database from two sources: One is self-recording; Second is to assign audio labels from available sources. The process described in Figure 3 can be described as follows:

Process of building a self-recorded database: Building recorded documents (Developing content selection tools, selecting text to ensure sound balance)→Selecting speaker (Determining voice type, checking speaker voice) → Prepare to record (Pronunciation rules, recording equipment, recording environment, recording software) → Record (record on software, check and monitor the recording process) →Set Standard recording database (reviewed, selected and synthesized).

Process of building a database from available audio sources: Select audio topic → Collect audio/video sources → Convert audio and standardize format → Preliminary identification of text → Select Vietnamese audio based on text → Splitting preliminary identified audio →Developing labeling tools →Transcription rules →Labeling → Cross-checking → Acceptance.

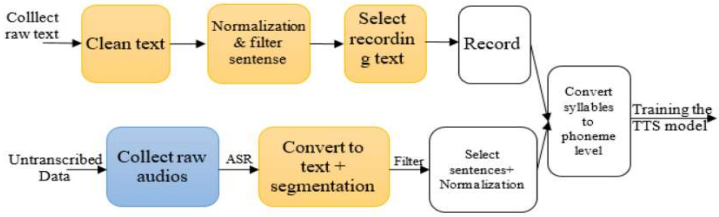


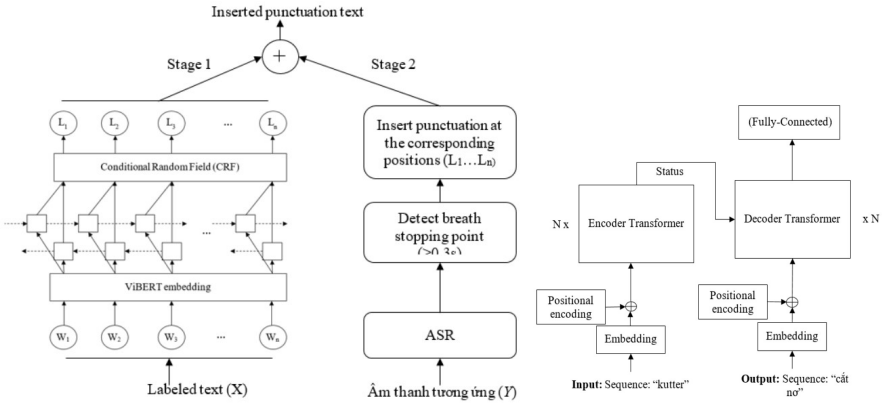
Figure 3: Process of building a comprehensive and adaptive database

2.1.3. Add information labels to the database

2.1.3.1. Insert breath stops and punctuation for label text

Text labels are enhanced with information about where to insert punctuation in the appropriate position according to two strategies, according to Vietnamese standards and according to the reader's pause to take a breath. To solve this problem, the thesis used two solutions: First, using the BERT model to restore punctuation in labeled text sentences; Second, mimic the reader's pause time, using the measured time stamp from the available ASR system, if the silence time is more than 0.3 seconds, place a comma in this silence position.

Besides, the system also adds a period at the end of the labeled text to represent the end of the sentence. The processed text will be taken to the next transcription post-processing step before being used to train the TTS model. The architecture is depicted in Figure 4a.



a) b)

Figure 4: Method of adding information labels and transcribing borrowed words to the database

2.1.3.2. Transliteration dictionary of borrowed words

To handle and solve this problem, Vietnamese pronunciation has been used to transcribe these English words, for example, "kuttner" will be pronounced with "cắt no". To convert from foreign transcription to Vietnamese transcription, the thesis used the basic Transformer translation model with the architecture described in Figure 3b. To train the translation model, many English-Vietnamese word pairs must be created. The result is the construction of a phonetic dictionary of borrowed words to serve Vietnamese trainees.

2.1.4. Result

The result has been to build a database of 54 multi-speakers, including 26 male voices and 28 female voices with North-South dialects, with the length of each voice varying from a few tens of minutes to several hours (including self-recorded voices) audio collected from available audio sources).

No	Data set	Number of utterance	Number of syllables	Average syllables/ utterance	Unique syllables
I	**Recording database**				
1.1	250 common sentences spoken by multiple speakers	250	3,268	13.06	1.205
1.2	9,600 individual sentences spoken by multiple speakers	9,600	105,232	10.96	5.516
II	**Database collect from open sources**				
2.1	Single female speakers	5,074	100,284	19.76	2.278
2.2	Single male speakers	13,125	280,130	21.34	2.893

2.2. Conclusion of Chapter 2

Chapter 2 listed published datasets for speech synthesis of resource-rich languages and the Vietnamese language. Analysis shows that there is a serious lack of databases for Vietnamese language research (including synthesis and adaptation). Chapter 2 also presents research results on building a low-cost database from two sources of self-recorded data and available data according to strict procedures to achieve a single-speaker and multi-speaker database with guaranteed quality for synthesis and adaptation [CT6] [CT3]. In addition, the method of adding information labels to text information is presented through inserting accents combined with inserting breath stops and pronouncing borrowed words to increase naturalness when training Vietnamese TTS systems. [CT5][CT4].

This section presents the process and methods. The purpose of this process and methodology is to ensure low cost in building a database from unlabeled data available on the Internet using good quality ASR and an effective data collection and filtering strategy. As a result, we have built 02 sets of standard single-speaker databases (1 male and 1 female), 02 sets of multi-speaker databases (250 common sentences and 9,600 unique sentences), and a test database set applied for synthesis and adaptation. In particular, the multi-speaker database has 54 speakers, including 26 male voices and 28 female voices with Northern/Southern dialects, with the length of each voice varying from a few dozen minutes to several hours (including self-recorded voices) and collected from available audio sources). This is an essential foundation for building systems to synthesize and adapt Vietnamese in the following Chapters.

CHAPTER 3. TRAINED ADAPTIVE SYNTHESIS MODEL WITH SMALL SAMPLES (FEW-SHOT TTS)

Chapter 3 will answer the question: Which method helps synthesize speech well for resource-poor languages like Vietnamese while only having a few minutes of adaptive samples? How much is adaptive data (trained with the system) needed at least to ensure the high quality and similarity of the synthesized voice? Chapter 3 will present proposals to improve the synthesis model based on adaptive high-quality Vietnamese speech synthesis; the content includes: 1) Presenting the Vietnamese speech adaptation method using Multi-pass fine-tune [CT3]; 2) Presenting the Vietnamese speech adaptation method using speaker characteristic vector (EMV) [CT2]; 3) Testing and evaluation have demonstrated that just 1 to 4 minutes of adaptation data gives good synthesis quality and high similarity.

3.1. Adapted for voice synthesis and methods

3.1.1. Method

1) TTS Adaptation based on fine-tuning the model: This method uses data from multiple speakers to train an average model. The average model is tuned to a small amount of target data for a particular target speaker. Adaptation can be achieved by returning all or part of the model parameters.

2) TTS Adaptation based on speaker feature encoding: In this method, an embedding vector or network is used to condition the identity and speaking

style of the training voice. During the training of the average model, vector or embedding networks are used to distinguish acoustic features from different speaker identities and speaking styles.

3.2. Improve the quality of single-speaker adaptive TTS with Multi-pass fine-tune technology

With the traditional fine-tuning approach, creating a new voice in a new language different from the pre-trained model still requires much data (≥ 5 hours, which is not easy in low-resource languages). If too small of an amount of data is used, it is easy to cause overfitting due to direct adaptation on the end-to-end acoustic model (Figure 5).

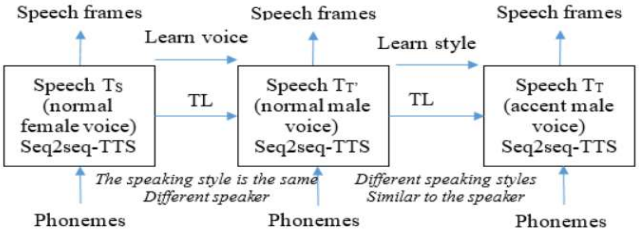


Figure 5 Flow diagram of speech adaptation using traditional fine-tuning

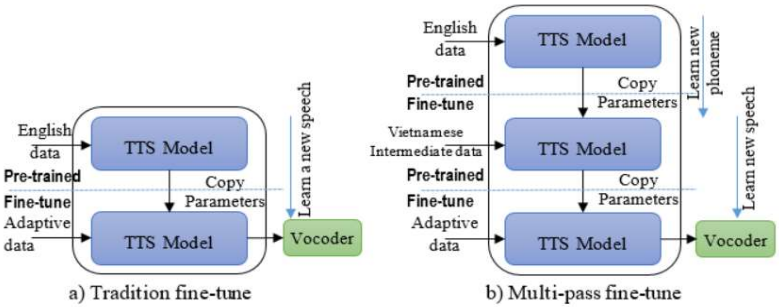


Figure 6: Voice adaptation to a new voice with Multi-pass fine-tune

To solve these problems, the thesis proposes a model that borrows a pre-trained model from English and refines it for the first time with a pre-trained model in Vietnamese as an intermediary, then refines it a second time with an adaptive speech model. The thesis calls this method "multi-pass fine-tune" and shows the diagram on the right side of Figure 6, which only requires a small sample to tune a new voice. Since the main acoustic features have been learned/transferred from English (large data set) and Vietnamese (medium data set), only a small amount of data is needed to adapt the model to create new speech to learn target voice characteristics..

Figure 7 describes the steps to implement Multi-pass fine-tune: 1) First train the network with a large amount of English data to generate the parameter set of the English model; 2). Then use this network to adaptively train the intermediate Vietnamese accent, the parameters of the English model is updated with the

parameters of the intermediate Vietnamese model; 3) Finally, use this network to adaptively train the target Vietnamese accent, the parameters of the intermediate Vietnamese model is updated by adaptation parameters of Vietnamese accent destination With the traditional fine-tuning method, the model only trains the English model and updates the parameter set by the adaptive parameter set.

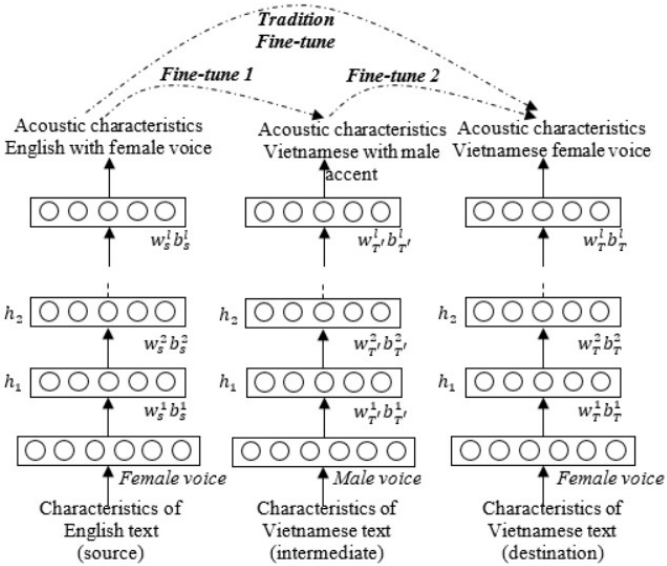


Figure 7: Updating adaptive parameters using Multi-pass fine-tune and traditional tuning

3.2.1. Evaluation testing

Data used: English dataset: Using LSpeech-1.1 corpus; Intermediate Vietnamese dataset: Male voice reading news with a total duration of 15 hours. Adaptive dataset: 4 adaptive datasets from 50 to 800 sentences (corresponding from 4 to 60 minutes). The baseline Tacotron2 + Waveglow was used to evaluate the adaptive TTS system.

3.2.2. Result

***Traditional refined model quality**

Table 1: Adaptive quality statistics table (MOS) according to Multi-pass fine-tune model and other models

Time	Training from scratch (TV data)	English + adaptive pre-trained model	Intermediate Vietnamese pre-trained model + adaptation
16 minutes	1.29	1.33	**3.78**
60 minutes	1.31	2.68	3.87
5 hours	2.66	N/A	N/A

- If training from the beginning with the Vietnamese data set, with 5 hours of data, the speech quality is very poor (MOS = 2.66). Training with less than 1 hour of data will hear nothing.
- If the pre-trained model is in English with 1 hour of Vietnamese adaptive data, the quality will be the same as training from the beginning of 5 hours of Vietnamese data, but the synthesized sound quality is still poor (MOS = 2.68).

*** Fine-tune multi-pass modeling quality**

In column 4 of Table 1, based on the pre-trained model in English, if fine-tuned from the intermediate Vietnamese data set to the small adaptive data set, it only takes 16 minutes (200 sentences) to give good quality voice with a MOS score of 3.78/4.69 compared to that of the speaker's voice.

*** Similarity**

Table 2: Similarity assessment table of traditional and Multi-pass fine-tune models when compared to human speech with only 4 minutes of adaptation data

Model	MCD	SIM
Groundtruth	-	3,99
Tradition fine-tune	10.65	1.13
Multi-pass fine-tune	**7.94**	**2.87**

Table 2 shows that Multi-pass fine-tuning allows the creation of new voices with a much lower MCD than traditional fine-tuning (**2.74** reduction). With just ***4 minutes of adaptation data***, the Multi-pass fine-tune model produces a synthesized voice with much higher similarity than the synthesized voice from traditional fine-tuning (**2.87/3.99** of the real voice). Analyze and evaluate SIM according to Mirjam Wester and colleagues [65]. Through ANOVA analysis results, it shows that the traditional refined model has ($F=5,188 > F$ critical, $p<0.05$) and the proposed model has ($F=12,287 > F$ critical, $p < 0.05$) shows that the experimental results are different and statistically significant.

3.3. Improving the quality of adaptive synthesis using EMV feature vectors

3.3.1. Proposed feature extraction vector Extracting Mel-Vector (EMV)

Adaptation-based multi-speaker synthesis systems must use speaker features to train and adapt the adaptive model. Traditional methods often use an embedding module to extract representative order vectors. However, this primary method cannot capture the individual characteristics of each speaker, such as their identity, gender, age, and health, because it only relies on speaker identifiers as input. To solve this problem, some studies propose an alternative method involving a style vector representing the speaker's speaking style. Therefore, the thesis proposes a proposed Mel-Vector Extraction module (EMV module for short) based on the original architecture of a modified Mel-style speech style encoder that can extract a fixed vector from the Mel spectrum, as depicted in Figure 8.

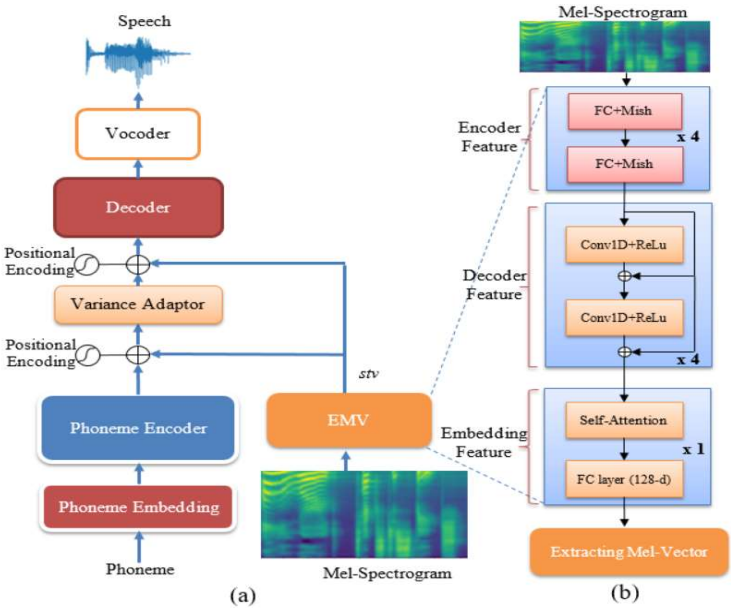


Figure 8: a) Architecture diagram of the model based on Vietnamese Multi-TTS adaptation with Extraction Mel-vector (EMV) module

Considering $\hat{y}$ to be the synthetic speech produced by the generative model G with input as text x and feature vector stv and trainable parameters θ, we have the following Mel spectrum output representation: $\hat{y} = G(x, stv; \theta)$

In which the speaker feature representation vector stv is generated by the EMV module through encoding the Mel spectrum of the original sound X as an adaptive sample as follows: $stv = EMV(Mel\_X)$

The overall architecture of the proposed model includes the following main components: The Phoneme Encoder module is used to transform the input phoneme sequence into a hidden one. Positional Encoding so that the model can determine information about the relative position of words in a sentence. The EMV module extracts speaker and speech style features from the Mel spectral input into a speaker feature vector. The Variance Adapter then adds duration, pitch, and intensity information to this hidden phoneme string. Moreover, the decoder will use this information to predict the Mel spectrum. Finally, the Vocoder block will convert these Mel spectra into speech signals. The overall architecture is depicted in Figure 8.

The detailed function and architecture of the proposed EMV modules include three main components: Feature Encoder Block, Feature Decoder Block, and Embedding Feature Vector Block. Feature). Accordingly, At the Encoder Feature block, the Mel-spectrogram input is first fed to the fully connected (FC) layer and the Mish activation functions (The Mish activation function was chosen due to its superior performance compared to ReLU and Swish) in many

experiments to convert each frame of the Mel-spectrogram into hidden sequences, which then pass through two FC layers to convert the input features into encoder features. Next, this vector will be passed through the Feature Decoder block. By using Conv1D +ReLu with residual connection to capture the sequence of information from the given speech sample, this module aims to convert the Decoder Feature into a decoder feature. In addition, skip connection is also integrated, which will use valuable features of previous blocks and solve the problem of gradient cancellation. Finally, the output of the Feature Decoder will be passed to the Embedding Feature module, which has a self-attention module with residual connection plus an affine layer to encode the information in a general way across the entire feature speaker and speaking style. Apply it at the frame level so that EMV can extract better speaking style information even with a short speech sample. Then, temporally weight the self-attention output of the adaptive samples and obtain a one-dimensional style vector stv. So, this module will create a vector representing the Mel-spectrogram, and this vector will be added to the text-to-speech model.

3.3.2. Training loss function

The model's reconstruction loss function is the sum of the L1 loss function that predicts the Mel spectrum from the EMV conditional vector and the L2 loss functions that predict pitch and additional loss functions that predict the acoustic features as the loss function predicts pitch and intensity. The general loss function of the model is as follows:

$$L_{final} = L_{mel} + L_{duration} + L_{pich} + L_{energy}$$

1. L_{mel}: The distance between the predicted Mel spectrum and the target Mel spectrum is described as follows: $L_{mel} = \mathbb{E}[\| \hat{y} - y \|_1]$
2. $L_{duration}, L_{pich}, L_{energy}$(variance reconstructs loss): Mean squared error between duration of syllables, pitch, and energy of prediction sample and target. $L_{duration} = ||d - \hat{d}||_2^2, L_{pich} = ||p - \hat{p}||_2^2, L_{energy} = ||\varepsilon - \hat{\varepsilon}||_2^2,$

Test evaluation and resultsa

a) Evaluation testing

To evaluate multi-speaker TTS systems, this section uses a modern multi-speaker speech synthesis model such as Fastspeech2 and the HifiGAN vocoder as the baseline model. In the proposed adaptive-based Multi-TTS model, as depicted in Figure 8, the base speaker embedding will be replaced with an EMV module to encode speaker features directly from the Mel-spectrogram. Data-distributing techniques are also used to keep speech characteristic parameters adaptive.

Dataset: The adaptive data is divided into four sets (1 min, 2 min, 4 min, and 16 min, respectively) to train Few-shot models using EMV.

b) Results

Table 3 shows that, with just 1 minute of target speech data, the adaptive-based Multi-speaker TTS model can synthesize a sound with a MOS score of 3.81

compared to the speaker's score of 4.6. This score is much higher than the MOS generated from the base Multi-TTS model (using 16 minutes of target voice). The WER score also demonstrates that the Multi-TTS model based on adaptive speech synthesis is better than the baseline Multi-TTS model.

Table 3: Quality assessment table between base Multi-TTS model (using basic speaker embedding) and Adaptive-based Multi-TTS model (using EMV module)

Model/ Duration	Multi-TTS baseline		Adaptive-based Multi-TTS	
	MOS (↑)	WER(↓)	MOS(↑)	WER(↓)
Groundtruth	4.60	1.35	4.60	1.35
1 minute	3.39	8.40	**3.81**	**5.00**
2 minutes	3.52	7.28	3.87	2.75
4 minutes	3.59	6.16	4.00	2.00
16 minutes	3.61	5.60	-	1.25

Table 4 shows that, with only 1 minute of adaptive sample speech data, the adaptive-based Multi-TTS model has a SIM similarity score of 2.60 compared to 4.0 for human speech. This score is much higher than the 1.96 SIM score of the base Multi-TTS model (using 1 minute of adaptive samples). The MCD score of the adaptive-based Multi-TTS model also decreased by more than 10% compared to the baseline TTS model. Analyze and evaluate SIM according to Mirjam Wester and colleagues [65]. Through the results of ANOVA analysis, we see that the base model has ($F=4,636 > F_{critical}$, $p < 0.05$) and the proposed model has ($F = 4,608 > F_{critical}$, $p < 0.05$) for found that the experimental results were different and statistically significant.

Table 4: Similarity between the Baseline Multi-TTS Models and the Adaptation-Based Multi-TTS Models compared to the groundtruth audio with only 1 minute of adaptation data

Model	MCD	SIM
Groundtruth	-	4.0
Multi-TTS baseline	7.36	1.96
Adaptive-based Multi-TTS	**6.54**	**2.60**

Figure 8, illustrates the t-SNE projection of speech style vectors between the speaker voice and the corresponding synthetic voice. Using the voices of 10 speakers (5 men and 5 women), the system's speaker model using EMV represents the speaker's characteristics very well when showing similarities between human voices and synthesized speech through points of expression that are clearly and tightly clustered in each separate area.

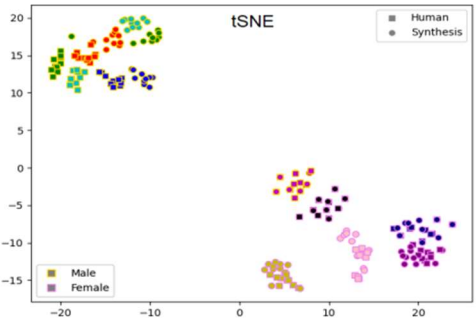


Figure 8: Distributed t-SNE of human voice and synthesized speech (using EMV)

3.2. Conclusion of Chapter 3

Chapter 3 presents some advanced adaptation techniques that are highly effective with small amounts of adaptation data in the world. From the analysis, the disadvantages of traditional fine-tuning methods have been pointed out. It is proposed to build an adaptive synthesis system for Vietnamese Multi-pas fine-tune using transfer learning techniques and testing to evaluate the adaptive system. Evaluation results have proven that: 1) With the Multi-pass fine-tune technique, only a small amount of data (4 minutes) allows the system to synthesize highly similar voices (with high scores). SIM scores 2.87/3.99) compared to 1.13/3.99 for the traditional fine-tuning model, and only 16 minutes of adaptive data allows the system to synthesize high-quality speech with a MOS score of 3.78/4.69 (equivalent to 4.03/5) compared to 2.68/3.99 for the traditional refined model [CT3].

In Chapter 3, an adaptive model is also proposed to improve the quality of the Vietnamese Multi-TTS system with EMV feature vector module architecture to overcome the disadvantages of embedded vectors representing speech features. The proposed model has superior performance to the baseline Multi-TTS model using traditional speech feature vector representation. Through experiments, in just **1 minute**, the proposed model achieved high similarity and good voice quality compared to the original voice. With only 1 minute of adaptive data, the above Multi-TTS model with adaptive capability gave a MOS quality of **3.8/4.6** and a SIM similarity score of **2.6/4** [CT2]. This MOS point is equivalent to using **16 minutes** of adaptive data based on the Multipass-fine-tune technique presented in 3.3.1. That proves that the EMV module effectively represents the speaker's features compared to the introductory speech feature representation vector (x-vector, d-vector, Thin ResNet), which is suitable for the training model. Few-shot TTS is capable of representing hidden speaker features seen during training. However, the need to retrain the model is a limitation of these techniques, and a few minutes of adaptive data needs to be more attractive. In the next Chapter, the thesis will evaluate the EMV module to enhance the performance of the adaptive-based synthesis system for the Zero-shot TTS model with data that has never appeared during the training

process. On that basis, we will research and propose solutions and improvement models in Chapter 4.

CHAPTER 4. UNTRAINED ADAPTIVE SYNTHESIS MODEL WITH MINIMUM SAMPLES (ZERO-SHOT TTS)

As presented in Chapter 3, current adaptation-based speech synthesis techniques rely on two main streams: one is to fine-tune the model using small-sized adaptive data, and the other is to train the entire model. Image through a vector representing the speaker characteristics of the target accent. However, both of these methods require adaptive data to appear during training, which makes training time to generate new voices quite expensive. Additionally, the traditional TTS model uses a simple loss function to reproduce acoustic features, but this optimization is based on incorrect distribution assumptions, leading to noisy synthesized audio results. Chapter 4 proposes an Adapt-TTS model that improves sound synthesis performance at an acceptable level from a small adaptive sample without training. Chapter 4 will present the following contents: Extracting Mel-vector (EMV) architecture allows for better representation of speaker characteristics and speaking style; The improved Zero-shot TTS model with the Mel-spectrogram denoiser diffusion component uses the properties of the back diffusion process integrated with the EMV feature vector to extract voices to generate new voices for the model. The zero-shot TTS model allows synthesizing new voices without training with better quality [CT1]; Testing and evaluation results for the proposed model [CT7];

4.1. Proposing the Adapt-TTS model to improve performance for Vietnamese adaptive synthesis

The architecture of Adapt-TTS includes the main components: EMV module to extract speaker features and speaking style into a feature vector, Phoneme Encoder module used to transform phoneme string into phoneme hidden sequence, then the Variance Adapter will add duration, pitch, and intensity information to this hidden sequence. The Mel-spectrogram denoiser will receive the hidden information in the previous steps to decode it into Mel spectra with high quality based on the diffusion model architecture kernel. Finally, the Vocoder block will convert these Mel-spectrogram into speech signals.

4.1.1. Encoding featured with EMV

Chapter 4 proposes a module similar to Chapter 3 called Extracting Mel vector (Extracting-Mel vector module or EMV) that can extract a fixed vector from the speaker's Mel spectrum graph to represent exact speaker characteristics such as speaker characteristics and speaking style. EMV will take reference voice X as input; the purpose of this block is to extract an embedding vector stv containing the style and characteristics of speaker *X*.

4.1.2. Mel-spectrogram denoiser

The decoder block takes input from the hidden phoneme sequence through the variance adapter to add variance information (e.g., duration, pitch, and energy) and then combines it with the EMV vector (representing human features). Then, Mel-spectrogram-denoiser module will take as input sequence x_t, variable c is the output of the variance adapter, and time step t to perform high-quality audio denoising and synthesis based on the diffusion model. The inference process of the diffusion model for multi-speaker TTS will optimize the objective function $f_\theta(x_t|t,c)$ to convert the noise distributions into a mel-spectrogram distribution corresponding to the given text and the model. It includes two main processes:

Diffusion process: First, the mel-spectrogram is gradually corrupted with Gaussian noise and transformed into latent variables. This process is called the diffusion process. Assuming a sequence of variables $x_1 \ldots . x_T$ with equal dimensions, where *t= 0,1 ....,T* is the index for diffusion time steps, the diffusion process transforms the mel-spectrogram x_0 into Gaussian noise x_T through a chain of Markov transition.

Reverse process: The reverse process for generating a mel-spectrogram is the opposite of the diffusion process. Rather than introducing noise, the goal of the reverse process is to recover a mel-spectrogram from Gaussian noise. This process is defined by the conditional distribution $p_\theta(x_{0:T}|x_T,c)$, and can be decomposed into multiple transitions based on the Markov chain. Using the reverse transitions $p_\theta(x_{t-1}|x_t,c)$, the latent variables gradually reconstruct a mel-spectrogram corresponding to the diffusion time-step with the text condition. Mel-spectrogram denoiser thus learns a model distribution $p_\theta(x_0|c)$ via the reverse. Set $\alpha = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ The training objective of the mel-spectrogram denoiser is as follows :

$$minL_\theta = E_{x_0,\epsilon,t}\left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, c)||_1\right]$$

4.1.3 Conditioned sound generation

With the task of generating conditional sound based on multiple input information, considering y as additional condition information labels, then convert all the above diffusion formulas with the same conditions as follows:

$$p_\theta(x_{t-1}|x_t) \rightarrow p_\theta(x_{t-1}|x_t,y) \; vs \; \epsilon_\theta(x_t,t) \rightarrow \epsilon_\theta(x_t,t,y)$$

where y is the conditions such as variable c is the output of the variance adaptor, speaker feature vector stv generated by EMV, then represent the above formulas as follows:

$$p_\theta(x_{t-1}|x_t) \rightarrow p_\theta(x_{t-1}|x_t,c,stv) \; vs \; \epsilon_\theta(x_t,t) \rightarrow \epsilon_\theta(x_t,t,c,stv)$$

Brief of training and inference

Training. The final loss value during Mel-spectrogram denoiser training includes the following parts: $L_{final} = L_\theta + L_{SSIM} + L_{duration} + L_{pich} + L_{energy}$

where L_θ (sample reconstruction loss) is MSE mean square error between predicted and target mel-spectrogram sample; L_{SSIM}(structural similarity index

measure loss - SSIM) is 1- the SSIM index between the predicted and target mel-spectrogram sample; $L_{duration}$, L_{pich}, L_{energy} (variance reconstructs loss) is Mean squared error between duration of syllables, pitch, and energy of prediction sample and target.

During inference, the mel-spectrogram denoiser predicts the input x_0 without noise and then re-adds the noise using the posterior distribution, thereby generating mel-spectrogram planes with increasing details. Specifically, the denoising model $f_\theta(x_t, t, c)$ first predicts x_t, then x_{t-1} is sampled using the posterior distribution $q(x_{t-1}|x_t, x_0)$ given by x_t and predicts x_{t-1}. Finally, a pre-trained vocoder converts the spectrogram plane generated from x_0 to a waveform.

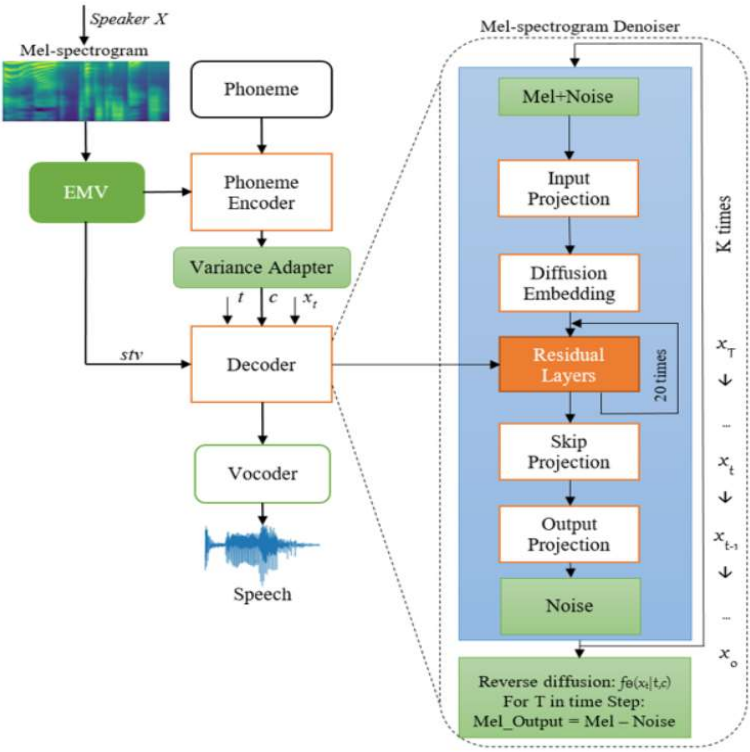


Figure 9: Detailed architecture of the Mel-spectrogram Denoiser

4.2. Test evaluation and results

4.2.1. Evaluation testing

To evaluate the quality of synthesized sounds created from the proposed models, 5 data sets were prepared, of which 4 sets were synthesized from reference audios with time intervals of 1 second, 3 seconds, respectively. seconds, 5 seconds and 1 set of original voices for comparison. Use two models for

synthesis: one is the base model proposed by Fastspeech2 and the other is the Adapt-TTS model. Use 30 listeners for evaluation. Evaluate the integrated system by combining both objective evaluation (WER) and subjective evaluation (MOS/SIM).

4.2.2. Result

a) Overall quality

Table5 shows that with only 3 seconds of adaptive sound from the new speaker, even without retraining, the Adapt-TTS model was able to synthesize the sound with an MOS score of 3.29 compared to 4.53 for the speaker. This score is higher than the base model's score of 2.16. The WER score also shows that in just 1 second the adaptive sound of the new speaker (reference speaker) can synthesize a sound reaching WER 3.38.

Table 5: Results of evaluating the quality of MOS/WER synthesis of the base models and proposed models for voices not yet in the training set (unseen-speaker) with 95% confidence level.

Model/ Duration	Base-line		Adapt-TTS	
	MOS (↑)	WER(↓)	MOS(↑)	WER(↓)
Groundtruth	4.53	1.35	4.53	1.35
1 second	2.05	8.78	2,89	**3.38**
3 seconds	2.16	7.77	**3.29**	3.14
5 seconds	2.18	6.76	3.31	3.04

b) Similarity

Table 6: Results of evaluating the similarity SIM of basic models and proposed models with 95% confidence level

Model/ Duration	Base-line	Adapt-TTS
	SIM	SIM
Groundtruth	3.90	3.90
1 second	1.16	1.71
3 seconds	1.24	**2.22**
5 seconds	1.31	2.6

Table 6 shows that with only 3 seconds of adapting sound from the new speaker, the Adapt-TTS model achieved similarity with a SIM index of 2.2/3.9 of the speaker's voice. Meanwhile, the base model only achieves a SIM index of 1.24/3.9. Conduct SIM evaluation analysis according to Mirjam Wester et al. [65], summarizing listener similarity scores for all evaluated sentence pairs (between synthetic voice and original sound) shown in Figure 10, in which symbols S1, S2, .. represent the order of reviewers. The performance shows confidence in the "definitely similar" and "possibly similar" abilities of the two proposed Adapt-TTS models, and the base model is prominent. Through the results of ANOVA analysis, we see that the basic model has ($F=1,675 > F$ critical, $p < 0.05$) and the proposed Adapt-TTS model has ($F = 2,099 > F$ critical,

p < 0, 05) shows that the experimental results are different and statistically significant.

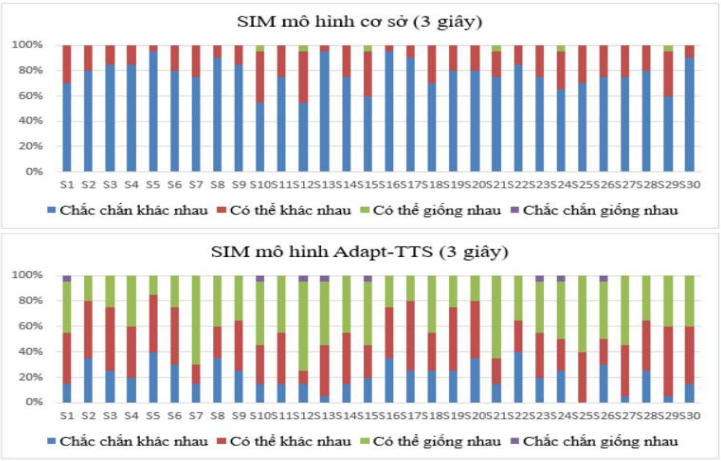
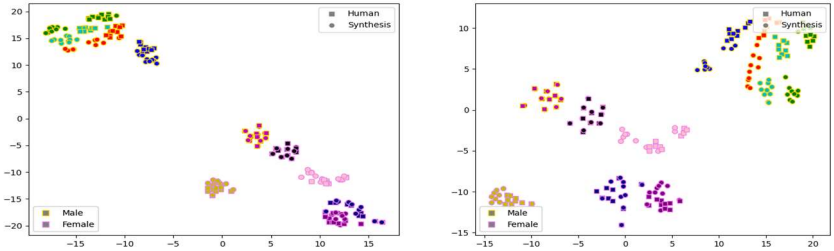


Figure 10: Comparison of the similarity of the base model (top) and the proposed Adapt-TTS model (bottom) across all pairs of evaluation sentences

t-SNE projection of EMV vectors obtained from invisible speakers in the Vietnamese multi-speaker dataset, specifically, selecting 10 speakers (5 male and 5 female). Adapt-TTS shows an explicit decomposition of speaker representation vectors and is closer to the original audio when compared to the baseline model. The t-SNE chart of the Adapt-TTS model (Figure 11a) shows that the synthetic and natural sounds of the same speaker are clustered close together, showing similarity. Gender characteristics were also clearly clustered in two different regions.



a) b)

Figure 7. Modeling the spatial distribution of t-SNE between the synthesized voice of the proposed model on the human voice by 10 speakers by a) Adapt-TTS model and b) Baseline model.

Thus, in the proposed Adapt-TTS model, which does not initially provide any information about the speaker's identity to the encoder model, the distribution predicted by the encoder is forced to be independent of the speaker's identity. Therefore, Adapt-TTS can convert speech using only the encoder model. The encoder characterizes the speaker with EMV and allows the adaptation of voice

characteristics. The Mel-spectrograms are predicted more accurately thanks to the Decoder with the Mel-spectrogram denoiser component following a diffusion model to create high-quality sound with little noise.

4.3. Conclusion of Chapter 4

Chapter 4 proposed an Adapt-TTS architecture that allows synthesizing a new voice using the Zero-shot TTS voice adaptation method with just a single sentence of the new speaker's audio sample without training. Model again. The proposed use of EMV in combination with the Mel spectral diffusion denoising model allows for humanized speech synthesis with acceptable quality. Experiments have proven that a 1-3 second sample of the sample voice can synthesize a voice with MOS quality of 3.3/4.5 and SIM similarity of 2.2/3.9 [CT1]. Although the sound quality created from the Zero-shot TTS adaptive model cannot achieve the quality or replace the adaptive synthesis models with retraining, such as Few-shot TTS, the proposed model compensates and allows one to quickly learn new voices without retraining. The synthesized sound quality is still guaranteed at an acceptable level and achieves good similarity with the target voice. The proposed Adapt-TTS model allows the synthesis of new speech based on a single adaptive sentence sample without having to retrain the model, allowing the application of the synthesis model to be expanded and applied to multiple applications form in life [CT7].

CONCLUSION

Chapter 1 presented detailed surveys and analyses of current Research and related knowledge on speech synthesis and adaptation. Chapter 2 presents the results of building a database using effective and low-cost methods as a foundation for building synthetic and adaptive models in the following chapters. Chapters 3 and 4 presented the most important proposals and experiments of the thesis ' Research and development of a speech adaptation system for Vietnamese language synthesis and applications' with main contributions. The content of Chapters 3 and 4 presented three speech synthesis methods that ensure quality for resource-poor languages like Vietnamese while only taking a few minutes of adaptive samples, which are DNN-based adaptive techniques for the model: Speaker-dependent (Few-shot TTS) and speaker-independent (Zero-shot TTS). The detailed proposed techniques of these Chapters also answer the research questions about the minimum number of adaptive samples (trained on the same system and not trained on the same system) accompanied by experiments. Specific testing and evaluation:

1) **Proposing a Multi-pass fine-tune model** to adaptively synthesize Few-shot TTS for high-quality Vietnamese using transfer learning techniques. The proposed speaker-dependent model can clone a newly trained voice to solve the problem of requiring less data from the cloned voice than the traditional method. Only a 4-minute speech sample allows the system to synthesize highly similar speech (with a SIM score of **2.87/3.99**), and only 16 minutes of adaptation data allow the system to synthesize speech with high quality with a MOS score of **3.78/4.69** compared to 2.68/3.99 of the traditional fine-tuning model [CT3];

2) **Proposing an EMV (Extracting-Mel vector)** architecture capable of effectively extracting features and representing speakers and a Few-shot TTS adaptation model for Vietnamese to help enhance the adaptation quality. The proposed speaker-dependent model is capable of cloning a new accent, requiring less data than fine-tuning techniques. With only 1 minute of adaptive data, the above Multi-TTS model with adaptive capability has achieved a MOS quality of **3.8/4.6** and a SIM similarity score of **2.6/4** [CT2]. In addition, the Variance adapter architecture can adjust voice (controlling voice characteristics such as pitch and intensity).

3) **Proposing the Adapt-TTS model** to solve the problem of voice cloning without retraining (Zero-shot TTS). The proposed speaker independence model solves the problem of cloning a new voice with very little data and no retraining and can be applied in practice. The proposed model can be replicated with a single sample sentence (1-3 seconds) through the EMV feature vector and Mel-spectrogram denoiser architecture without training. Re-modeling, the MOS synthesis quality reached **3.3/4.5**, and the SIM similarity reached **2.2/3.9** [CT1];

4) **Building a voice database that ensures quality and low cost** for synthesis and adaptation tasks [CT6] [CT3]; Technique for adding label information to increase the naturalness of Vietnamese speech synthesis systems through (inserting punctuation, inserting breath stops, and transcribing borrowed words) [CT5][CT4]. The result of this section is an essential database for synthesis and adaptation for use throughout Chapters 3 and 4 of the thesis.

5) **Develop a voice cloning application** that can be used on multi-platform devices to imitate and synthesize any voice to demonstrate the feasibility and performance of the proposed models with demonstrations [CT7]. Each proposed adaptive model will have its advantages and disadvantages and, therefore, different practical applications: The Few-shot TTS model will provide good synthesis quality with only a few minutes to tens of minutes of data. Adaptive data allows duplicating voices or creating exclusive voices for broadcasting and reading automatic reports; a Zero-shot TTS adaptive model with only one sentence of data and no training is suitable for instant voice learning of the user, applied to smart speakers.

Development

1) Research solutions to enhance adaptive quality for emotional or voice samples with little data.

2) Experiment with the models proposed in this study with published data sets in English, Chinese, etc., to compare the model's effectiveness.

3) Apply the proposed model for multi-lingual adaptation techniques

4) Continue to improve the Adapt-TTS model and compression algorithms for the corresponding training/synthesis model to reduce computational costs and be able to run on devices with small resources.