การปลอมตนด้วยเสียงสำหรับเสียงพูดภาษาไทยโดยใช้ไซเกิลแกนบนสัทสัมพันธ์

นายชาตรี ชวนงูเหลือม

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

VOICE IMPERSONATION FOR THAI SPEECH USING CYCLEGAN OVER PROSODY

Mr. Chatri Chuannglueam

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2021

Thesis Title	VOICE IMPERSONATION FOR THAI SPEECH USING CYCLEGAN OVER PROSODY
By	Mr. Chatri Chuannguleam
Field of Study	Computer Engineering
Thesis Advisor	Professor Boonserm Kijsirikul, Ph.D.
Thesis Co-advisor	Assistant Professor Nuttakorn Thubthong, Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

. . . . . Dean of the Faculty of Engineering
(Professor Supot Teachavorasinskun, D.Eng.)

THESIS COMMITTEE

. . . . . Chairman
(Assistant Professor Sukree Sinthupinyo, Ph.D.)

. . . . . Thesis Advisor
(Professor Boonserm Kijsirikul, Ph.D.)

. . . . . Thesis Co-advisor
(Assistant Professor Nuttakorn Thubthong, Ph.D.)

. . . . . External Examiner
(Associate Professor Cholwich Nattee, Ph.D.)

ชาตรี ชวนงูเหลือม: การปลอมตนด้วยเสียงสำหรับเสียงพูดภาษาไทยโดยใช้ไซเกิลแกนบนสัทสัมพันธ์. (VOICE IMPERSONATION FOR THAI SPEECH USING CYCLEGAN OVER PROSODY) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ศ. ดร. บุญเสริม กิจศิริกุล, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : ผศ. ดร. ณัฐกร ทับทอง 0 หน้า.

การปลอมตนด้วยเสียงเป็นงานที่ท้าทายในเรื่องที่ต้องเลียนแบบในทุกแง่มุมของเสียงเป้าหมาย วิทยานิพนธ์เล่มนี้นำเสนอการแปลงสัทสัมพันธ์โดยใช้ไซเกิลแกน บนข้อมูลชุดเรียนรู้ของการแปลงเสียงแบบไม่ขนาน แบบจำลองไซเกิลแกนได้แสดงให้เห็นถึงประสิทธิภาพที่โดดเด่น ในการถ่ายโอนแบบอย่างของรูปภาพที่ไม่เข้าคู่กัน ในระเบียบวิธีนี้เสียงพูดที่เกิดการแปลงผันจะถูกสร้างจากการแปลงของลักษณะทางสัทสัมพันธ์และลักษณะเชิงสเปกตรัมของเสียงตั้งต้น ซึ่งจะสามารถแทนบุคลิกของเสียงของเป้าหมายได้ดีกว่า ผลการทดลองจากระเบียบวิธีการประเมินเสียงสังเคราะห์แสดงให้เห็นว่า การแปลงสัทสัมพันธ์ให้ผลดีกว่าระเบียบวิธีการแปลงเสียงแบบไม่ขนานแบบสัญนิยมอย่างมีนัยสำคัญ

ภาควิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต	.....
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่อ อ.ที่ปรึกษาหลัก	.....
ปีการศึกษา	2564	ลายมือชื่อ.ที่ปรึกษาร่วม	

6070162121: MAJOR COMPUTER ENGINEERING
KEYWORDS: VOICE IMPERSONATION / VOICE CONVERSION / GENERATIVE ADVERSARIAL NETWORK / PROSODY TRANSFORMATION

CHATRI CHUANNGULUEAM : VOICE IMPERSONATION FOR THAI SPEECH USING CYCLEGAN OVER PROSODY. ADVISOR : PROF. BOONSERM KIJSIRIKUL, Ph.D., THESIS COADVISOR : Asst. PROF. NUTTAKORN THUBTHONG, Ph.D., 0 pp.

Voice impersonation can be a challenging task for mimicking all aspect of the target speaker. This Thesis proposes a prosody conversion using a cycle-consistent adversarial network (CycleGAN) on non-parallel training data in voice conversion (VC). A CycleGAN model had demonstrated an outstanding performance in style transfer of the unpaired images. In this method, the converted voice was generated from the transformation of both prosodic and spectral features of the source voice. In this way it can better represent the persona of the target speaker. Experimental results from standard evaluation procedures for evaluating synthesized voice demonstrated that our prosodic CycleGAN significantly outperformed the conventional non-parallel VC method.

Department:	Computer Engineering	Student's Signature	.....
Field of Study:	Computer Engineering	Advisor's Signature	.....
Academic Year:	2021	Co-advisor's signature	.....

Acknowledgements

I would like to thank Assistant Professor Nuttakorn Thubthong, my co-advisor, for driving me into the speech processing field and suggesting me to enter graduate school. I thanked Professor Boonserm Kijsirikul, my advisor, for introducing me to Machine Learning field. Even when I find myself stuck with the difficulty of understanding the theory behind AI process or stuck with my research ideas, he always knows what to say to cheer people up and give motivations. His ideas on research are extremely intriguing and his guidance on how to do research in broad range of field with novelty is exceptional. Talking with him can spanned for hours but what we students get from his casual talk is not only about research and academic, but it comprised of how to lead one's life, how to stay unique, and how to change thinking perspective to fit each environment which are all valuable life lessons.

I thanked my beloved mother and father who support me into higher education financially and always encourage me to stick with the research and for always listening to every complaint and problem on my research.

I also thanked my MIND lab mates in helping with the exchanging ideas, and constant words of encouragements.

CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

Chapter I

INTRODUCTION

1.1 Motivation

Speech is one of the most common tools for human communications. Everyone has a unique personality in speech. Such that, it is possible to impersonate the others by mimicking their voice. In a computer system, there is an essential element that can manipulate the characteristic of human speech called voice conversion (VC). VC is a technique that transforms a speech signal of a source speaker to sound like as it uttered by a target speaker, while keeping its linguistic contents (**?**). VC has been used in various applications such as speaking aid for impaired speech individual (**?**), transformation of emotion in speech (**?**), singing voice modification (**?**) and accent conversion for foreign language learning (**?**). More importantly, VC is known to be an advanced presentation method for identity theft and forgery. Due to the availability in resources and computational performance, VC has become more capable to fool the automatic speaker verification (ASV) systems (**?**). In recent years, various types of generative adversarial networks (GANs) had been used in VC which outperformed the conventional techniques such as Gaussian mixture model (GMM)-based (**?**) and deep neural networks (DNN)-based methods (**?**). Furthermore, the unsupervised learning nature in GAN can comfortably surpass the major limitation of the conventional VC for using with a non-parallel data that the linguistic contents of the source and the target are different. One of the most promising GAN-based model is the cycle-consistent generative adversarial network (CycleGAN) due to its simplicity and its outstanding performance in style transfer for the unpaired images transformation (**?**). It has shown the favorable results in a non-parallel VC even in a cross-language condition (**?**). However, the crucial goal for voice impersonation is to strongly convince a listener to believe that the converted voice had been naturally uttered by the target speaker. Most of the re-

searchers had used GANs to solely transform some speech features such as spectral features which are flexible and robust to noise while linearly transforming the other features such as prosody, although the features are also important for personality in speech. Therefore, the converted voice might not totally represent the persona of the target speaker which could be distinguished apart from the genuine voice by human perception. In addition, various languages have distinct characteristics which depend on different speech features. Thai language depends on the tonal characteristics in speech (**?**) which relies mostly on prosody. A prosodic feature such as fundamental frequency (F0) should be a possible option to carry the personality of the target in VC. We have proposed a method that transforms both prosodic and spectral features using a CycleGAN to improve the performance of non-parallel VC methods. This method would transfer better characteristics of the speaker in the conversion process especially in Thai language.

1.2 Problem Statement

In Thailand, public personalities such as politicians, news reporters, and actors are the most potential spoofing attack representation scenario due to a large amount of speech data in publicly available sources and recognizable by the general public. Unlike in English, Thai language depends on the tonal characteristics in speech (**?**) which relies mostly on prosody. A prosodic feature such as fundamental frequency (F0) may be a possible choice to carry the personality of the target in VC. In contrast to the spoofing attack, voice impersonation also benefits the speech synthesis process such as creating a voice template for announcing in public places (e.g., train, airport, shopping complex). Furthermore, it can greatly affect the language localization in media such as movie, conference and video game by producing the translated speech in foreign languages to have the same personality as the original speaker.

1.3 Objective

1. To impersonate the speech of the Thai public personalities.
2. To apply the CycleGAN technique for voice conversion on Thai language.
3. To demonstrate that prosody is crucial to the speech personality in Thai language.

1.4 Scope of Work

1. Using Thai language speech of the three recognizable persons as target speakers.
2. Using a non-parallel VC which is the procedure where the speech contents such as word and sentence from source and target speaker are different.

1.5 Benefits

1. To personalize and create a Thai speech according to the desired voice template.
2. To help improve the perceived quality and naturalness for VC.
3. To present a potential threat on speech identity in Thai language for the further countermeasure.

1.6 Research Implication

Giving an example on the importance of prosody in speaker's personality and Thai speech, such feature should be properly transformed in VC. For the representation of prosody using in this thesis, F0 is an excellence candidate due to its sim-

plicity in extraction and its success in VC (**?**) and the main component in speech synthesis method. Instead of transform F0 linearly, a CycleGAN model is train to transform such prosodic feature which can better pass on personality of the target to the converted voice. For the evaluation procedures, to measure the similarity between target and converted voice, a Dynamic Time Warping (DTW) method was selected in this thesis on the strength of its effectiveness and its possibility to measure prosodic features beyond F0 (**?**). This thesis can demonstrate that a prosodic feature such as F0 has a crucial role in Thai speech and VC.

Chapter II

BACKGROUND

2.1 Speech Features

Speech features are elements that characterize events in speech signal **?**. There are several ways to represent the phonemes in speech signal which can be categorized into two major groups: the segmental and suprasegmental (also known as prosody) features **?**. The segmental features can discriminate vowels and consonants which can be represented by spectral features. By splitting speech signal into short-time frames or segments, it can be extracted into the spectrogram which is a three-dimensional representation of the speech intensity in different frequency bands over time **?** using Fourier transformation. The result of the inverse Fourier transformation of the logarithmic power of the spectrogram is called cepstrum **?**. The cepstrum is often calculated with a nonlinear Mel frequency band mapping called Mel-cepstral which is used in several speech analysis tasks **?**. On the other hand, the prosody is the aspects of speech which span groups of syllables or words **?**. The principal interests of prosody are pitch (or fundamental frequency: F0) and stress which depend on the persona of the speaker. Intonation which is the pitch variation over sentence can give shape to the sentence and indicate its structure. Furthermore, pitch also can be used to help indicate the meaning of words in some languages **?**. Additionally, one of the most used parameters in VC and speech synthesis is aperiodicity band which is a spectral parameter associated with mixed excitation **?**.

2.2 Voice Conversion

Voice conversion (VC) is the transformations that use acoustic parameters from the speech of two speakers: source and target, to convert the source voice

to sound as it was naturally uttered by the target **?**. There are two major categories for VC: parallel and non-parallel VC. A parallel VC requires the same linguistic contents and only varies in the aspect of mapping model **?**. In the contrary, a non-parallel VC can be independent to linguistic contents. It can be categorized into two simple types: the feature-pair searching and individually replacement **?**. The feature pair searching methods focus on the aspect of aligning the similar features of the source voice and the targets. They require the linguistic information to label the events in speech. On the other hand, the individually replacement methods are interested in the aspect of the splitting the speech into two components: linguistic and speaker identity. They focus on the replacing of the source identity with the targets while keeping the linguistic component.

2.3 Generative Adversarial Network

Cycle-consistent Generative Adversarial Network

The cycle-consistent Generative Adversarial Network (CycleGAN) is based on the GAN technique proposed by Goodfellow et al. **?**. It had demonstrated an outstanding performance in style transfer of the unpaired images **?**. It consists of two discriminators (DA and DB) and two generators (GA and GB), as shown in Figure **??**. A generator serves as the mapping function from the distribution of input A to B. On the other hand, discriminators aim to classify between the real and generated distributions. The goal of this model is to learn the mapping function from given training samples to convince discriminators to classify the generated object as the real object.

The source and target are represented with A and B respectively. The input A as source is added to discriminator D_A and generator G_{A2B} to be classified as real A and to create generated B respectively. Discriminator D_B then classifies the generated B whether it is real B or generated B. Accordingly, the generated B is transformed back to cyclic A to begin the new cycle. By mapping from source $a \in A$

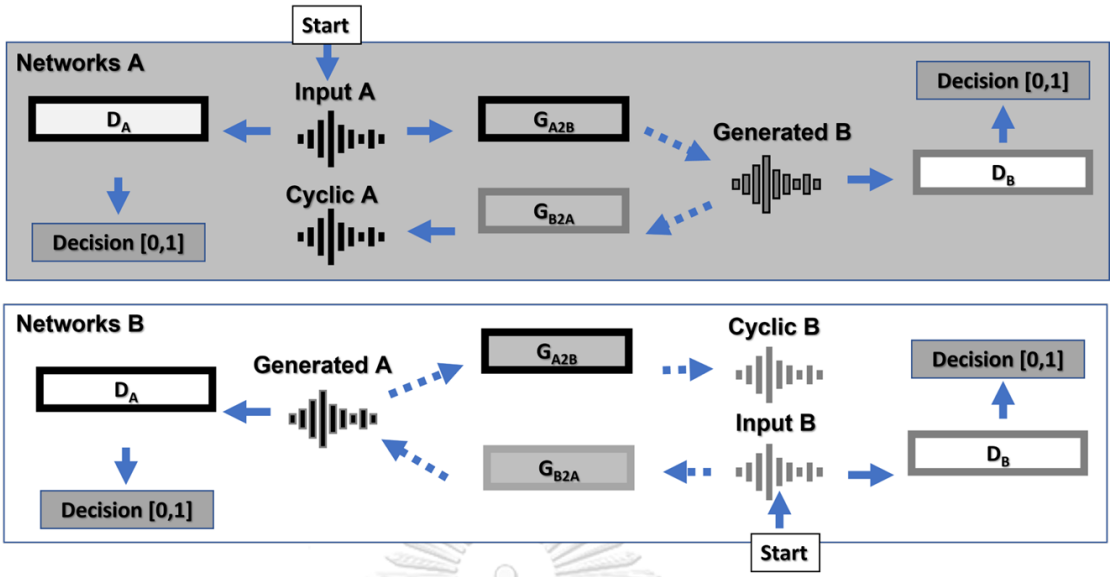

Figure 2.1: Diagram of CycleGAN. DA and DB are discriminators. GA and GB are generators. A and B are the real distribution.

to target $b \in B$ with CycleGAN, two losses: adversarial loss and cycle-consistency loss were used to learn the model.

Adversarial loss measures the discrepancy between the converted data $G_{A\to B}(a)$ and the target data b. Therefore, distribution of the converted data $P_{G_{A\to B}}(a)$ must have the distribution closer to the target data $P_{Data}(b)$ to gain smaller loss. This can be written as

$$L_adv\,(\,G_{A\to B},\,D_B\,) = L_{b\,P_{Data}(b)}[log\;\;D_B(b)\,] + L_{a\,P_{Data}(a)}[log\;\;(1 - D_b(G_{A\to B}(a)))\,]. \tag{2.1}$$

On the other hand, cycle-consistency helps to preserve the contextual information of a in the conversion, described as

$$L_cyc\,(\,G_{A\to B},\,G_{B\to A}\,) = L_{a\;P_{Data}\;(a)}\,[\,\|G_{B\to A}(G_{A\to B}(a)) - a\,\|1] + L_{b\;P_{Data}(b)}[\,\|G_{A\to B}(G_{B\to A}(b)) - b\,\|1]. \tag{2.2}$$

Hence, the two losses can be illustrated with trade-off parameter λ_{cyc}:

$$L_{full} = L_{adv}\,(\,G_{A\to B},\,D_B) + L_{adv}\,(\,G_{B\to A},\,D_A) + \lambda_{cyc}L_{cyc}(G_{A\to B},\,G_{B\to A}). \tag{2.3}$$

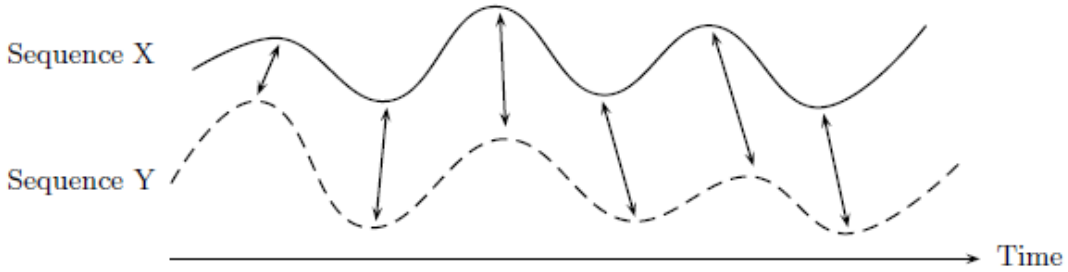

Figure 2.2: Time alignment of two time-dependent sequences. Aligned points are indicated by the arrow.

2.4 Dynamic Time Warping

Dynamic Time Warping (DTW) is the widely used technique that can determine an optimal alignment between two given (time dependent) sequences under certain restrictions (**?**). In speech processing, DTW can be used to compare different speech patterns in speech recognition. DTW has demonstrated the meaningful results with prosody when use with Hermes measurement to evaluate the similarity between sets of speech sentences (**?**). The main idea of DTW is to compare the distance between two aligned sequences as shown in Figure **??**.

Chapter III

REVIEW OF LITERATURE

3.1 Voice Conversion with Deep Neural Networks

Most early research had focused on developing a parallel VC. The widely used methods were based on the mapping model, in which the aligning of the similar pair features of the source and target speakers are performed. Toda et al. (**?**) used a Gaussian mixture model (GMM) as a mapping model on spectral features which yielded a good result in speech quality. To improve the capability of the GMM-based approach, Sone et al. (**?**) applied a deep relational model (DRM) with GMM to outperform a deep neural network (DNN)-based methods.

3.2 Voice Conversion with GANs

Recently, GAN has become an attractive tool in VC. Kaneko et al. (**?**) proposed the sequence-to-sequence VC using the GAN-based training framework that surpassed a deep autoencoder technique. Furthermore, various GAN-based techniques have exceptionally improved the previous non-parallel VC. Hsu et al. (**?**) and Chou et al. (**?**) incorporated VAE with GAN in a non-parallel VC which indicated an improvement over the conventional VC method. Additionally, StarGAN-VC (**?**) demonstrated superiority over the VAE-GAN-based method. One of the most favorable of GAN-based techniques in a non-parallel VC is CycleGAN due to its remarkable performance in style transfer of unpaired images (**?**) which is desirable for a non-parallel data condition. Fang et al. (**?**) proposed a transformation of the lower order of Mel-cepstrum (spectral features) using CycleGAN. Although it yielded a good voice quality, but prosody was transformed using linear trans-

formation which result in the absence of the persona of the target speaker such as intonation and rhythm. Inspired by a PPGs-based model, Yeh et al. (**?**) proposed the non-parallel VC using CycleGAN over PPGs which removed the length constraint of the converted voice to capture to rhythm of the target voice. However, the alignment of features using PPGs may introduce new errors.

3.3 Voice Conversion with Prosody

In contrast to the spectral features, most research focus on F0 to represent prosody in VC. F0 contour was widely used in tonal language such as Mandarin (**?**) which proved in conveying the speaker emotion in the converted voice. Moreover, the using prosodic conversion together with the spectral features has the higher quality of the converted voice than just the spectral conversion (**?**). In term of style transfer in speech, the F0 transformation can generate a significantly correlated voice to the target speaker in VC (**?**).

In conclusion, the use of CycleGAN in a non-parallel VC to transform both spectral features and prosody might transfer the persona of the target to the converted voice. Therefore, it might have a better performance in the impersonation of the target voice.

Chapter IV

PROPOSED METHOD

4.1 Speech Dataset

Our dataset was built from publicly available video contents (news, documentary). All voices gathered are in Thai language. We had chosen 2 recognizable persons (male and female) as target speaker. After a total of 600 utterances (300 utterances for each gender) was gathered, two native Thai speakers then read the transcription from the target speaker by his own style as source voice. The recording conditions are in standard studio.

4.2 Training and Conversion Setup

We adopted the technique of CycleGAN for a non-parallel VC based on Fang et al **?** as baseline method due to its performance that outperformed a conventional VC. Our proposed method aimed to create converter modules which transform prosodic features and spectral features separately by training CycleGAN models as shown Figure **??**. The architecture for generator and discriminator in CycleGAN were 1D gated convolutional neural networks (Gated CNN) and 2D Gated CNN respectively.

As shown in Figure **??**, the audio data form target and source was extracted using WORLD **?** to the three main speech features: Mel-cepstrum, F0, and aperiodicity bands. As a prosody representation, F0 was encoded due to its dimensionality conditions. Then MCEP and encoded F0 were fed to the CycleGAN model separately to create converter 1 and 2 respectively.

Similarly to the training process, the source voice was extracted by WORLD **?** to the same features. Then they converted to the characteristic of the target using our converter 1 and 2 created by prosodic CycleGAN as shown in Figure **??**. The

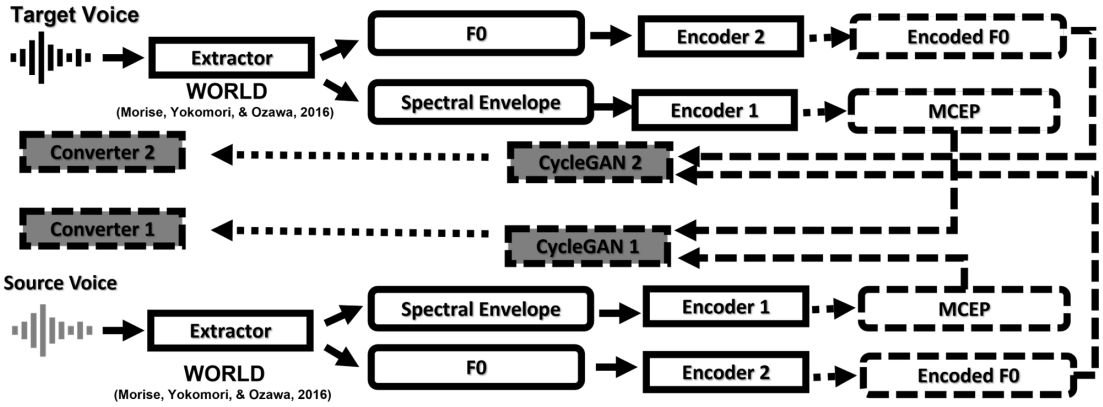

Figure 4.1: Diagram of training process for the proposed method. Both the encoded F0 as Prosodic feature and MCEP as Spectral feature are transform using CycleGAN to create 2 separated converters.

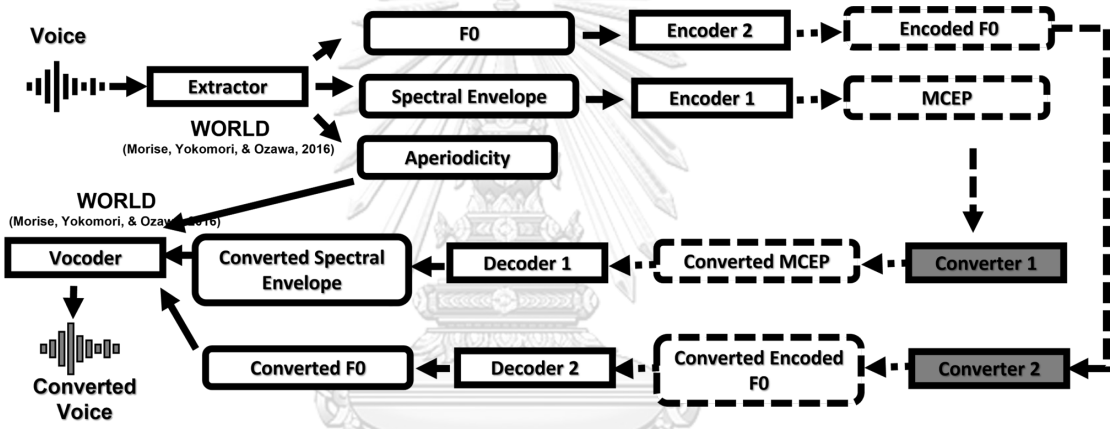

Figure 4.2: Diagram of voice conversion process for the proposed method. Converted voice is created by Vocoder from transformed features of source using converters from training process.

converted voice that has the personality as the target then was created by WORLD **?** as a vocoder in the speech synthesis process.

Chapter V

RESULTS

A total of 200 utterance (100 utterances for each gender) converted form each method was reserved as test set for the evaluation procedures.

To demonstrate the improvement in the conversion of prosody (F0), the proposed method (Prosodic CycleGAN) and the baseline method (CycleGAN) need to be compared. As show in Table 1 , a dynamic time warping (DTW) approach **?** was used to compute the prosodic similarity in the test set between the converted and the target voice.

The proposed Prosodic CycleGAN method had significantly outperformed the baseline CycleGAN. The possible reason is due to the superiority of the non-linear prosody conversion in carrying extra characteristics of the speaker compared to the linear transformation. Thus, it might help reduce the DTW alignment cost. We noticed that the Male to Female conversions (SM→SF) for all methods had inferior results compared to the Female to Male conversions (SF→SM). One possible reason is the lower variance in time in the male voice might have the insufficient components in the conversion.

Method	Con-version	r	DTW	DTW +r	Perc
CycleGAN (baseline)	$SF \rightarrow SM$	0.70	27681	0.75	0.86
CycleGAN (baseline)	$SM \rightarrow SF$	0.68	32471	0.74	0.84
CycleGAN (baseline)	AVG	0.69	30076	0.75	0.85
Prosodic CycleGAN (proposed)	$SF \rightarrow SM$	0.75	26574	0.82	0.90
Prosodic CycleGAN (proposed)	$SM \rightarrow SF$	0.70	28152	0.80	0.87
Prosodic CycleGAN (proposed)	AVG	0.73	27363	0.81	0.89

Table 5.1: Hermes measure (r), DTW alignment cost. (DTW), DTW-aligned Hermes measure (DTW+r) and perceptual similarity (Perc.) obtained from the test set. The same sentence of target speakers and converted voices are compared.

Chapter VI

DISCUSSION

6.1 Summary of Finding

From the results of comparing our proposed method with the baseline, we can see that using Prosodic transformation has better overall score in Hermes measure (r) and lesser dissimilarity in DTW than the baseline. But for the meaningful measurement DTW-aligned Hermes measurement (**?**) can represent better similarity in the similarity in prosody. The proposed method has significantly outperformed the baseline in this measurement. We can conclude that the prosodic transformation can improve the similarity of the converted voice to the target.

Perceptually, the proposed and baseline have similarity in term of cross gender conversion. From female to male conversion has performed better than male to male. This is due to the higher range in F0 of the female can store more personality than the male one.

6.2 Conclusion

This paper proposes prosody transformation using CycleGAN to improve a non-parallel VC Method. The results demonstrate a superiority of our prosodic CycleGAN method over a conventional non-parallel VC method using standard evaluation procedures. By using a speech dataset in Thai language which depends on the tonal characteristic in speech, it validates the significance of the prosody in the personality of the speaker. Investigation the features in prosody further to improve the voice conversion will be our work in the future.

6.3 Future Work

Prosodic Conversion techniques can be useful in carry over more personality of the speaker to the converted voice. There still are many parts available to be optimized, such as improving the speech synthesis model, create a larger dataset and improving the prosody extraction. The Wavelet decomposition of F0 can capture more variation of prosody in speech synthesis (**?**). Such decomposition might be advantageous in prosodic VC.

Appendix I

LIST OF PUBLICATIONS

A.1 International Conference Proceeding

1. Chuannguleam, C., Kijsirikul, B., & Thubthong, N. (2022). Voice Impersonation for Thai Speech Using CycleGAN over Prosody. In 2022 4th International Conference on Management Science and Industrial Engineering (MSIE) (MSIE 2022), April 28-30, 2022, Chiang Mai, Thailand. ACM, New York, NY, USA, 8 Pages. https://doi.org/10.1145/3535782.3535840

Biography

Chatri Chuannguluеam was born in Nakhon Ratchasima on January 17, 1994. He graduated from Princess Chulabhorn Science High School, Buriram and then went to Chulalongkorn University where he received B.Sc in Physics. His field of interest includes various topics in Speech Processing, Artificial Intelligence, and Machine Learning.