Study On Non-Parallel Voice Conversion using Variational Autoencoder with Modulation Spectrum-constrained Training

By Ho Tuan Vu

A thesis submitted to
School of Advanced Science and Technology,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master Science (Information Science)
Graduate Program in Advanced Science and Technology

Written under the direction of
Professor Masato Akagi

March, 2018

Study On Non-Parallel Voice Conversion using Variational Autoencoder with Modulation Spectrum-constrained Training

By Ho Tuan Vu (1610165)

A thesis submitted to
School of Advanced Science and Technology,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Science (Information Science)
Graduate Program in Advanced Science and Technology

Written under the direction of
Professor Masato Akagi

and approved by
Professor Masashi Unoki
Professor Jianwu Dang

February, 2018 (Submitted)

Abstract

This dissertation aims to propose a voice conversion (VC) approach that does not require parallel data or linguistic labeling for the training process. Speech conveys not only the linguistic information but also the non- and para-linguistic information. Linguistic information is the information that can be explicitly described by written language. Meanwhile, para-linguistic information refers to the information added deliberately by the user to alter the linguistic information such as intonation. Lastly, non-linguistic information represents the information related to emotion, speaker individuality, accent, etc... The non-linguistic information is usually unintentionally added by the speaker.

Voice conversion is the process of manipulating the non- and para-linguistic information of speech, such as speaker individuality, emotion, intonation, etc. Voice conversion technique has a wide range of applications. For instance, voice conversion can be used generate emotional speech from neural speech to improve user experience in human-machine interaction. For noisy environment such as train station, voice conversion can be applied to announcement system for enhancing the speech intelligibility.

For another example, the voice conversion technique can be employed in Speech-to-speech translator (S2ST) to generate personalized voice. S2ST is a device that translates a spoken utterance in one language to a spoken output in another language. Conventional S2ST systems focus on processing linguistic information only, ignoring the para- and non-linguistic information of the input speech. In other words, the output voice always sounds the same despite any input voice. It is known that non- and para-linguistic information play an important role in human communication. Therefore, the ultimate goal of this study is a cross-lingual voice conversion system that can be practically integrated into a commercial Speech-to-speech translator.

One of the core-part of realizing a Personalized S2ST system is the cross-lingual voice conversion. The cross-lingual voice conversion differs from conventional voice conversion system that the training speech data are uttered in different languages, which means that the source and target utterances have completely different linguistic information. However, most of the current VC methods require parallel training data. Dictionary-based voice conversion using NMF- factorization is one of the state-of-the-art VC methods where the input spectrum is approximated by a weighted linear combination of a set of dictionary (basis) and weights. However, the requirement for parallel training data in those systems causes several problems: 1) limited practical usability when parallel data are not available, 2) additional error from alignment process degrade output speech quality. In order to alleviate these problems, this paper presents a novel dictionary-based VC approach by incorporating a Variational Autoencoder to decomposed input speech spectrum into speaker dictionary and weights without aligned training data. By replacing the source speaker dictionary with target dictionary as similar to the conventional method, the converted spectrum can be constructed. In our proposed method, we assume that the weights should have normal distribution given a sufficiently large database. Moreover, we believe that the linear combination method put a trivial limitation to dictionary-

based voice conversion as the relation between speaker characteristic and output spectrum should be non-linear in general. Therefore, we generalize the dictionary-based method by utilizing the non-linearity of neural networks in the form of Variational Autoencoder as it has a similar concept to our assumption.

The difference in vocal tract shape mostly corresponds to the speaker individuality in speech. In the source-filter model of speech production, the shape of vocal tract reflects the filter part. Therefore, in the first step of our system, the STRAIGHT vocoder is utilized to decompose the speech waveform into the source-related part (pitch) and filter-related part (spectrum and aperiodicity). After that, 60 Mel-cepstral coefficients derived from the STRAIGHT spectrum is used as the input acoustic feature of our system as it shown the most significant impact on speech naturalness. In addition, according to the previous studies, degradation of modulation spectrum severely affect the quality of output speech. For that reason, we rewrite the training objective of Variational Autoencoder to include the cost of modulation loss.

The proposed system can be divided into two main phases: training phase and conversion phase. In the training phase, the obtained acoustic feature is used to train the Variational Autoencoder model using Stochastic Gradient Descent method. In this phase, the model parameters, source, and target dictionary are learned from the data. In the conversion phase, the activation matrix derived from the source utterance is applied with the target dictionary to generate the target utterance as similar to conventional dictionary based VC method.

From the objective measurement result, the MS of the synthesized speech using proposed training method is improved, indicating that the proposed training strategy is more efficient compared with the conventional method. In addition, the formant frequencies of the synthetic speech are close to those of target speech, indicating that the proposed system can capture and transform the speaker individuality. The proposed system also generate speech with lower Mel-cepstral distortion than the baseline system using NMF.

Acknowledgment

First and foremost, I would like to express my great appreciation to my supervisor at Japan Advanced Institute of Science and Technology, Professor Masato Akagi. Not only the technical knowledge that he taught me but also the way to conduct scientific research. He taught me how to organize ideas in a logical way and how to deeply understand the problem. I really appreciate every comments and encouragement that he gave me in every lab meetings or discussion. He allowed me to freely pursue my own ideas but gave me valuable advice when I get stuck. Without his guidance and persistent help, this thesis would not have been possible.

Also, I am particularly grateful for the insightful comments and suggestions from Professor Masashi Unoki. He taught me how to develop a critical thinking in sciences which is extremely important to be a researcher. He gave me helpful advice for writing a research paper and encourage me to keep trying after each falls. His works and efforts is the inspiration for me to work harder and be more determined to succeed.

I wish to acknowledge the valuable comments and instructions provided by Doctor Rieko Kubo, especially during the experiment process. She gave me a lot of helpful advice and constructive questions to deal with many arisen problems during doing research.

I would like to offer my special thanks to my "tutor" Teruki Toya for helping me settle down in JAIST. He has been helping a lot since I first came to JAIST and introducing many things about Japan.

I would like to thank Professor Hiroyuki Iida, Associate Professor Nguyen Le Minh, and Mr. Bui Nguyen Khanh, who introduced me the wonderful opportunities to study at JAIST. Without their help, I would not be able to study at JAIST and be supervised by Professor Masato Akagi.

I would like to thank Associate Professor Hirokazu Tanaka for helping me complete the minor research project. His efforts and ideas are always my motivation to improve myself.

I would like to thank Mr. Dinh Anh Tuan, Mr. Trinh Kim Dung, Mr. Ngo Van Thuan, Ms. Nguyen Thi Hao, Mr. Li Xingfeng and other lab members of Akagi and Unoki Lab, for their advice and cooperation during my research and their help in my daily life. I really appreciate every moment that we share during the time at JAIST.

My special thanks are extended to the staffs at JAIST for offering me the greatest research environment that I have ever experienced. JAIST has provided me many the cutting-edge technology, which helped me a lot for doing research. Without their help and support, I would be caught in many troubles.

Finally, I would like to express my best thankfulness to my parents, my sister Tram

and her husband Bang, my brother Anh. I cannot find any word to express my gratitude for all the things that they taught me and all their sacrifice for me. Thank you my love Dieu Linh for all the love and faith that you gave me for all the time.

Author
Ho Tuan Vu

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this first chapter, the research context, research objective as well as the contribution of this dissertation is briefly introduced. For the beginning, we explain the definition of Speech-to-speech translator and its significance. In the next parts, we describe the problem statement of current studies. Then the motivation and the scope of our study is presented. Finally, the structure of this dissertation is outlined.

1.1 Motivations

In our daily life, speech is the most common way for us to communicate to each other. However, a common language between speakers must be shared to communicate with others directly. As the world is more open, multinational environments where people speak different languages are becoming ordinary. Therefore, language turns into the major barrier for us to conduct an effective communication. One possible solution to overcome this problem is a speech-to-speech translator (S2ST). The main functions of the S2ST device is to convert a spoken utterance in one language to another language. As shown in figure 1.1, a typical automatic S2ST device consist of 3 main components: 1) Automatic speech recognition (ASR) system translates spoken utterances to texts, 2) Machine translation system convert recognized texts to target language texts, 3) Speech synthesis system synthesizes the output speech from the target language texts.

Speech is an effective tool for the human to communicate because not only linguistic information but also non- and para-linguistic information is conveyed in speech. The linguistic information represents a discrete information which can be explicitly described by written language. The lexical, syntactic, semantic and pragmatic information is contained in linguistic information.

Para-linguistic information refers to the information intentionally added by the speakers to alter the linguistic information such as intonation, intention, and attitude. Depends on the para-linguistic information, the same sentence can be perceived in different ways. For example, a raise at the end of sentence might indicate a question .

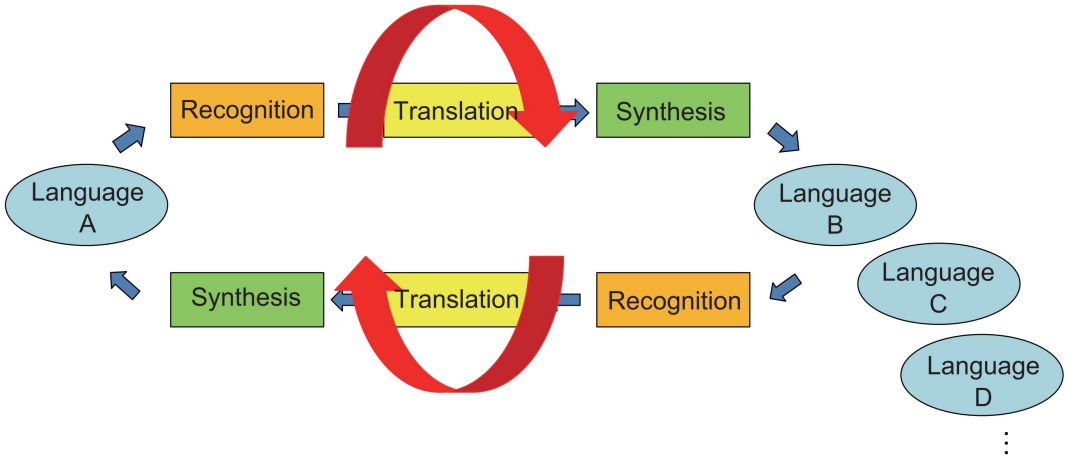

Figure 1.1: An S2ST system described by [1]

On the other hand, non-linguistic information represents the information related to the speaking styles, emotion, and speaker individuality. Non-linguistic information is generally not controlled by the speaker but added without intention.

Although various commercial S2ST systems have been proposed, all of them focus on processing linguistic information only, ignoring the para- and non-linguistic information of the input speech. In other words, the output voice always sounds the same despite any input voice. As stated in [1], para-linguistic information and non-linguistic information play important roles in human communication. Therefore, developing an S2ST system that can also translate para- and non-linguistic information is essential.

This study focuses on developing an S2ST system with personalized output voice. In the Figure 1.2, the expanded S2ST system considers not only linguistic information but also the non-linguistic information related to speaker individuality. One of the central parts for realizing a personalized S2ST system is the cross-lingual voice conversion system.

1.2 Problem Statement

The ultimate goal of this study is a cross-lingual VC system that can be practically integrated into a commercial S2ST system. Therefore, the method for constructing such cross-lingual VC systems must satisfy these criteria: 1) The system can be trained using non-parallel training data, 2) Produce high quality translated speech, 3) Require recorded speech from users as few as possible.

Concatenation method often gives the best naturalness, but it requires an enormous database to achieve this performance. Therefore, it is impractical for this method to be applied in a real S2ST system. Recently, spectral mapping using ANN has reached a comparable performance as concatenation method using fewer data. However, when considering the cross-lingual voice conversion, the spectral mapping method shows its limitation

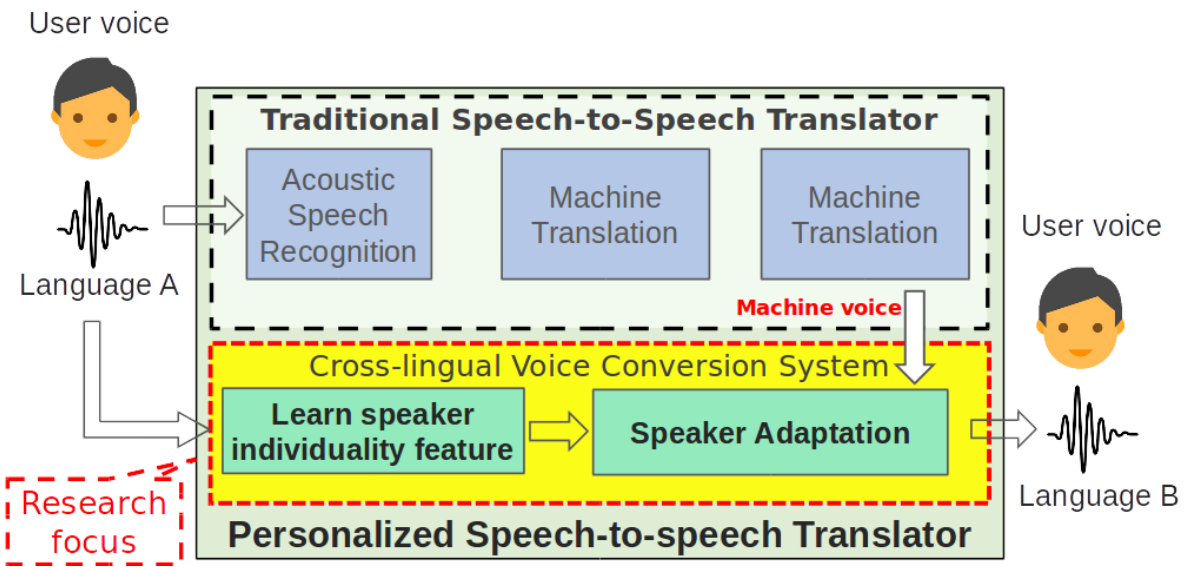

Figure 1.2: A personalized S2ST system

as it cannot be used without parallel training data. This is because the cross-lingual voice conversion must deal with training data containing completely different linguistic content in the source and target utterances. Speech decomposition methods such as Eigenvoice GMM and NMF assume that speech spectrum can be decomposed into two separate factors representing speaker identity and linguistic content. However, those methods still require parallel utterances of the source and target speakers to train the model. The quality of synthesized speech is still poor.

1.3 Research Objectives

The objective of this research is to proposed a voice conversion method that does not require parallel training data for the Speech-to-Speech translator device.

To achieve this goal, there are several problems must be solved. First, it is necessary to build a model that can be trained using non-parallel data. Theoretically, speech decomposition method need not use only parallel data. Therefore, the current work focuses on expanding the speech decomposition method to use non-parallel training data. Based on the same concept of NMF-based VC, we proposed a new method that can decomposed speech into speaker individuality factor and content factor using autoencoder. Although a method for dictionary update for NMF-based VC using autoencoder has been investigated in [28], this method still requires parallel training data. In our method, we aim to replace the conventional autoencoder with the more advance variational autoencoder (VAE)[15]. Attempt to apply VAE model for speech is conducted in [29]. However, no attempt has been made for speech decomposition using VAE model.

In addition, the second issue of the previous NMF-based VC is the low quality of converted speech. Previous studies of Dinh [6] have stated the significance of Modulation Spectrum (MS) of the perceived naturalness of speech. Therefore, this work also incorporates MS to alleviate naturalness of the synthesized speech.

1.4 Structure of Thesis

The remained of this thesis is structured as follows:

- **Literature Review** (Chapter 2): We give an overview of voice conversion system, Non-negative Matrix Factorization-based voice conversion, and the Variational Autoencoder model.
- **Proposed method** (Chapter 3): We describe our approach to building a Variational Autoencoder-based voice conversion system for using with non-parallel training data.
- **Experimental and Results** (Chapter 4): We carried out some objective measurements to asset the correctness of the model. The purpose and detail procedure of subjective test are also described in this chapter. Then we present the results of the objective measurement and subjective test.
- **Conclusion** (Chapter 5): In this chapter, we summarize our work and give out the contribution of our study. Finally, we describe the next steps of our research.

Chapter 2

Literature Review

2.1 Definition of Voice Conversion

Voice conversion (VC) is the process of manipulating the non- and para-linguistic information of speech, such as speaker individuality, emotion, intonation, etc. As shown in Figure 2.1, a typical VC system comprises of two main parts related to two tasks:

- *Training phase*: The training phase contains two stages: the first stage deals with obtaining the corpus of the source and target speaker, then the mapping between phonetic classes and acoustic feature of both speakers is generated in the second stage. The system learns the optimal parameters for spectral transformation using the training data. Training data may contain not only utterance from source and target speaker but other speakers as well.
- *Conversion phase*: Using the learned parameters from the Training phase, the source spectrum into the target spectrum. Several spectral enhancement methods are applied in this phase to improve the synthesized speech quality.

Several methods for voice conversion systems have been proposed. According to [2], these methods can be classified into five categories: statistical techniques (e.g GMM, HMM, PCA, K-means), cognitive technique (ANN), linear algebra technique (SVD), and signal processing techniques (VQ, FW, DFW). A statistical method using GMM is considered to be the most widely used method [2]. Recently, the exemplar-based voice conversion using non-negative matrix factorization has shown its successful to generate good quality synthetic speech in small-data condition [16]. However, most of these methods require pre-recorded parallel data for both source and target speaker, which is inconvenient and expensive in practical application.

The cross-lingual VC is much challenging than the typical VC in the sense that the mapping of acoustical features cannot depend on time-aligned utterances from the source and target speakers. To construct a personalized S2ST device, a cross-lingual voice conversion

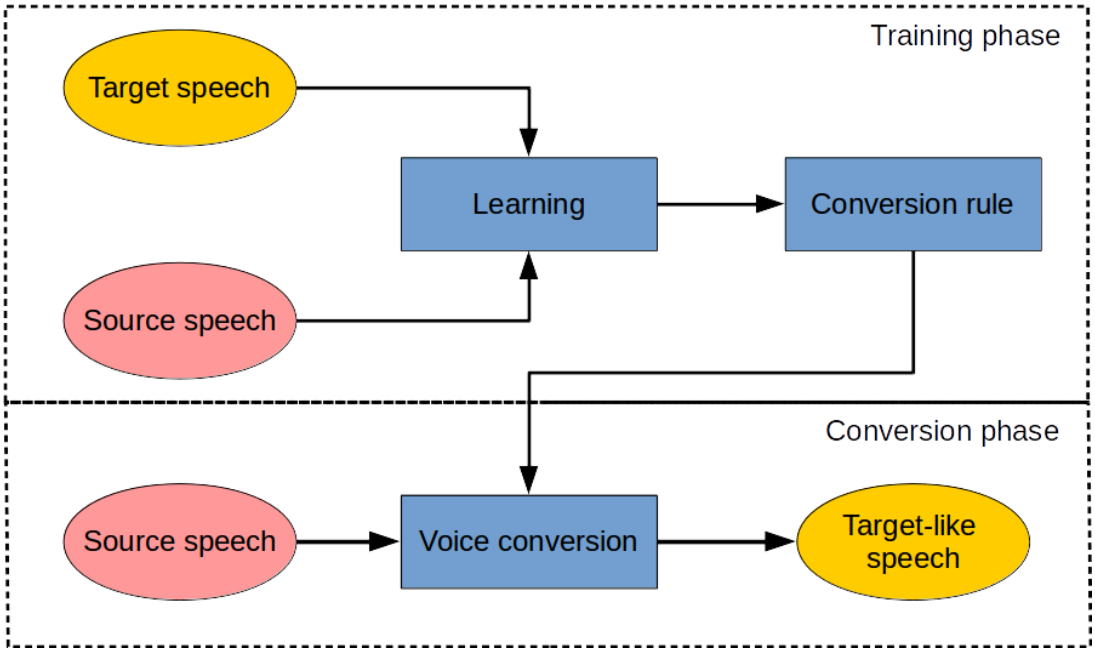

Figure 2.1: A typical voice conversion system

system must be achieved. One of the requirements of cross-lingual voice conversion system is the capability of using non-parallel training data. For decades of research, various methods for cross-lingual voice conversion have been studied so far such as concatenative method, spectral mapping using Gaussian Mixture Model (GMM) or Artificial Neural Network (ANN), speech decomposition using Non-negative Matrix Factorization (NMF) or Eigenvoice GMM (EV-GMM). Table 2.1 summarizes the techniques for cross-lingual voice conversion since the first proposal in 2003.

Table 2.1: Author and techniques in cross-lingual VC

Year	Author	Technique
2003	Kumar [8]	GMM
2006	Duxans [10]	GMM-CART
2006	Sundermann [11]	Unit Selection
2007	Uriz [12]	Frame selection-FW
2009	Zhang [14]	VQ
2010	Erro [13]	GMM
2013	Ariwardhani [3]	ANN-Vocal tract mapping
2015	Percybrooks [4]	HMM
2016	Rammani [5]	GMM

Concatenation methods such as unit selection or frame selection often give the best naturalness. However, they require an enormous database to achieve this high performance. Therefore, it is impractical for these methods to be applied in a real S2ST device.

Recently, a method based on vocal tract mapping using ANN has been proposed [3]. Nevertheless, the precondition of having vocal tract data greatly reduces its feasibility in a real application.

Similar to conventional VC, GMM is the most widely used method for cross-lingual voice conversion. The GMM method can generate acceptable quality of synthetic voice with significant fewer data than the concatenation method. This characteristic enables GMM to be the state-of-art method for cross-lingual voice conversion although there is still a big gap between synthetic voice and natural voice.

Among conventional VC methods, speech decomposition using NMF-based method has proved to be superior to GMM-based methods [16] [17]. However, these methods were proposed for using with parallel data only. The speech decomposition method owns a conceptual simplicity, which assumes that speech can be expressed by two separate factors corresponded to speaker identity and linguistic information. Therefore, this method does not necessarily depend on the parallel training data. The current work focuses on expanding the speech decomposition method to use non-parallel training data.

2.2 Background on NMF-based Voice Conversion

The basic concept of Dictionary-based VC is to decompose speech spectrum into two separate factors representing speaker individuality and speech content. The most common method to accomplished this task is Non-negative Matrix Factorization (NMF). The class of VC methods uses NMF is called NMF-based VC.

For NMF-based VC, a sequence of spectral frames $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}]$ are represented as linear combinations of dictionary matrix $\mathbf{A} = [\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_K}]$ (related to speaker individuality) and activation weight matrix $\mathbf{Z} = [\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_N}]$ (related to speech content) as follows:

$$\mathbf{X} \approx \mathbf{AZ} \tag{2.1}$$

The dictionary matrix $\mathbf{A}$ can be obtained by directly selecting spectral frames from training utterances. This method requires no training phase and the selected spectral frames are called the exemplars. At runtime, given the source spectrogram and the source dictionary, we can derive the activation matrix $\mathbf{Z}$. Then the activation matrix $\mathbf{Z}$ can be applied to the target dictionary to generate corresponding target spectrogram. The merits of this method are only limited data is required. Nevertheless, most of the spectral frames from training utterances are crudely used as exemplars, implying that a large dictionary is implemented. The large dictionary is beneficial for improving the quality of synthesized speech but require a long conversion time, which is unsuitable for applying in real-time application.

In another method, the matrix $\mathbf{A}$ and $\mathbf{Z}$ is learned from the training data by alternatively updating one matrix while keeping the other matrix fixed. The size of constructed dictionary using this method is greatly reduced compared to the exemplar-based NMF

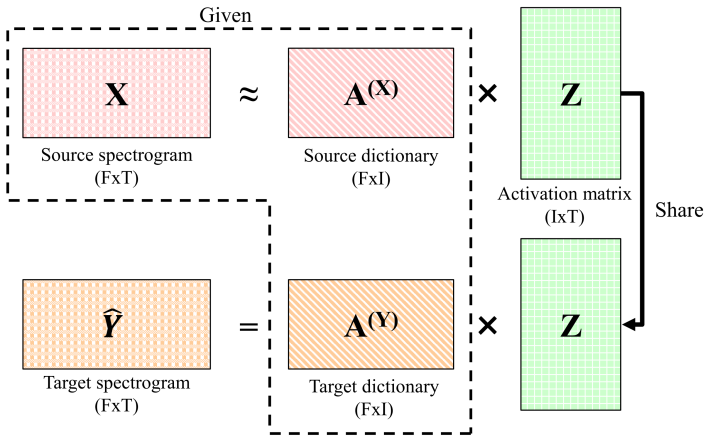

Figure 2.2: Illustration of NMF-based Voice Conversion

method, consequently improving the efficiency of online conversion process [17].

When applying in VC, firstly the source-target dictionaries $\mathbf{A}^{(X)}$, $\mathbf{A}^{(Y)}$ is constructed using parallel dataset. However, because of their different speech rate, the source and target utterances may not align with each other. For that reason, Dynamic Time Warping (DTW) is often used to obtain the frame-wise alignment of source-target utterances.

For generating converted spectrogram, in the next step, we assume the source and target dictionary share the common activation matrix. Given the source spectrogram and source dictionary, the activation matrix is estimated using equation 2.1. The converted spectrogram is obtained by multiply the target dictionary matrix with activation matrix. Figure 2.2 illustrates the detail of NMF-VC.

$$\hat{\mathbf{Y}} = \mathbf{A}^{(Y)}\mathbf{Z} \tag{2.2}$$

2.3 Background on Variational Bayes Autoencoder

2.3.1 Artificial Neural Network

Artifical Neural Network (ANN) is the computation model inspired by the functions and structure of biological neural networks. The most interesting of ANN is the learning process of the network, which is reflected by the reconfiguration of network structure by the information flow through it. The earliest ANN model dates back to 1950s when Frank Rosenblatt proposed the perceptron model [30]. The perceptron was intended to be a machine rather than a program. Figure 2.3 illustrate a simple perceptron model with 3 inputs.

In the modern sense, the perceptron is an algorithm for the binary classifier, which means it takes one or more inputs x and outputs a value $f(x)$ in range of $[0, 1]$:

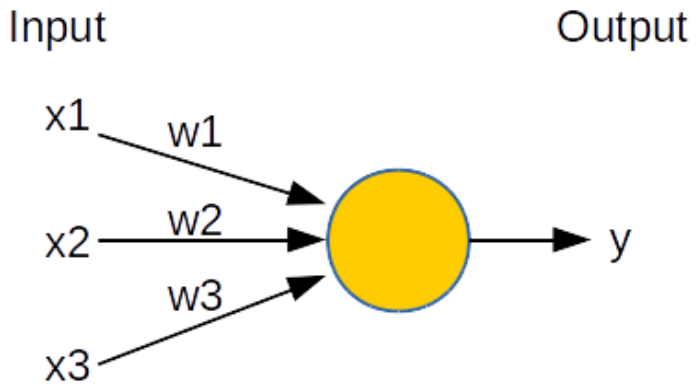

Figure 2.3: A perceptron with 3 inputs

$$f(x) = \begin{cases} 0 & \sum w_i.x_i + b > threshold \\ 1 & \sum w_i.x_i + b \leq threshold \end{cases}$$

where w_i is the weight corresponded to the input x_i and b is the bias.

In case of one perceptron, the only weighted-sum operation can be represented, which is not very helpful. To represent the more complicated functions, multiple perceptrons are combined into a network as illustrated in Figure 2.4. This network has 3 layers. The first layer receives the information from the input so we call it input layer. The second layer takes information from the output of the first layer, therefore it can make more abstract decisions than the input layer. This layer is called hidden layer. The last layer, which is the output layer, receives the information from the hidden layer and has one output perceptron.

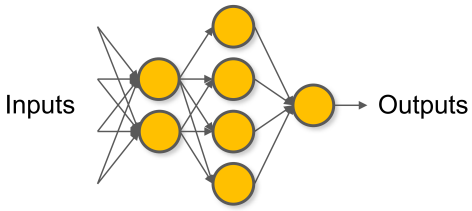

Figure 2.4: A multilayer perceptron

The next important point that make the MLP be a powerful model is the non-linear activation function. In this case, a non-linear function (sigmoid, tanh, etc..) on the weighted-sum input of each perceptron.

$$Output = \sigma(\sum w_i.x_i + b)$$

where $\sigma()$ is a non-linear function. Several widely used non-linear functions are:

- Sigmoid: $\sigma(z) = \frac{1}{1+e^{-z}}$

- Tangent hyperbolic (tanh): $\sigma(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$
- Rectifier linear unit (ReLU): $f(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}$

2.3.2 Autoencoder

Autoencoder is a special type of ANN where the output is the same as the input. It means that the autoencoder is an unsupervised learning model, i.e. we only need to know the input [33]. The aim of such network is to learn a representation of the set of data, usually for reducing the dimensionality of the data space. Autoencoders were first introduced in the 1980s by Rumelhart [32] to address the problem of "backpropagation without a teacher", by using the input data as the teacher. A simple autoencoder is illustrated in Figure 2.5, which consists a feed-forward and non-recurrent neural network. An autoencoder always has two distinct parts: the encoder and decoder network.

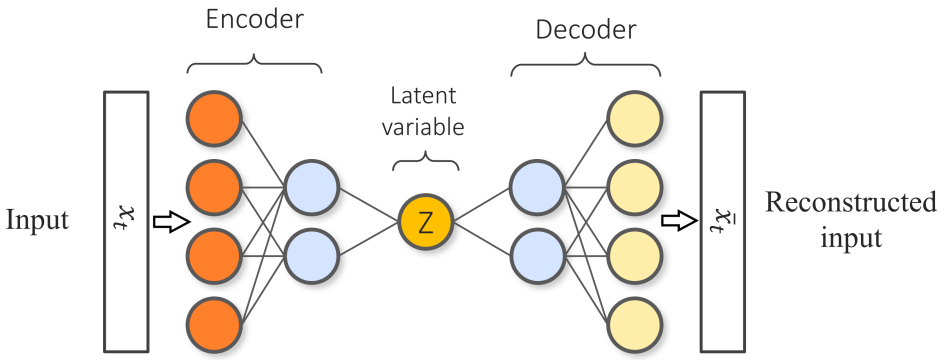

Figure 2.5: A deep autoencoder

In the simplest case, when both encoder and decoder networks is made of a single layer, the encoder stage takes the input $\mathbf{x} \in \mathbb{R}^d$ and maps it to $\mathbf{z} \in \mathbb{R}^p$.

$$z = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$$

The image $\mathbf{z}$ is usually referred as latent variables, latent representation, or code. σ is an element-wise activation function such as sigmoid or tanh. $\mathbf{W}$ is the weight matrix and $\mathbf{b}$ is the bias vector of the encoder network. After that, the decoder network maps the latent variables $\mathbf{z}$ to the reconstruction $\mathbf{x}'$ of the same shape of $\mathbf{x}$:

$$\mathbf{x}' = \sigma'(\mathbf{W}'\mathbf{z} + \mathbf{b}')$$

where $\sigma', \mathbf{W}', \mathbf{b}'$ may differs in general to the $\sigma, \mathbf{W}, \mathbf{b}$ of the encoder network. Since the error criterion of the autoencoder is the reconstruction criterion, this allows the autoencoder to learn an efficient coding of the data[31]. The most simplest training objective for autoencoder is to minimize the mean-square error between the input and output:

$$\mathcal{L}(x, x') = \|x - x'\|^2 = \|x - \sigma'(\mathbf{W}'(\mathbf{W}\mathbf{x} + \mathbf{b}) + \mathbf{b}')\|$$

If the dimension $\mathbf{p}$ of the latent variables is usually smaller than the dimension of $\mathbf{x}$, the latent variables $\mathbf{z}$ can be regarded as a compressed representation of the input $\mathbf{x}$.

2.3.3 Variational Autoencoder

Variational Autoencoders (VAEs), which is a special variant of autoencoder, is a both discriminative and generative model proposed by Kingma et al. and Rezende et al. in 2013 [15]. The VAE is different from the conventional autoencoder in several points. Firstly, the variational autoencoder defines a probabilistic generative model:

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}) \tag{2.3}$$

From Equation 2.3.3, the observed data $\mathbf{x}$ is assumed to be generated by a random process involving some underlying random variables $\mathbf{z}$. The latent variable $\mathbf{z}$ is assumed to have a posterior distribution belongs to a much simpler distribution family (Gaussian distribution):

$$p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Secondly, the data $\mathbf{x}$ is assumed to be generated by directed graphical model $p(x|z)$ and the encoder is learning an approximation $q_\phi(x|z)$ to the posterior distribution $p_\theta(z|x)$ where θ and ϕ denote the parameters of the encoder and decoder respectively. The objective function of the VAE has the following form:

$$\mathcal{L}(\phi, \theta, \mathbf{x}) = D_{KL}(q_\phi(z|x)||p_\theta(z)) - \mathbb{E}_{q_\phi(z|x)}(logp_\theta(x|z)) \tag{2.4}$$

Where the first term D_{KL} is the Kullback-Leibler divergence to regularize the distribution of the latent variables. The second term is the reconstruction cost. Training process is equivalent to iteratively estimate the autoencoder parameters θ and ϕ to maximize the equation (2.3.3). However, back propagation is not possible through random sampling ($z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$). Therefore, a sampling trick is applied as follow:

$$\begin{gathered}\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathbf{z} = \mu + \sigma \odot \epsilon \\ \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})\end{gathered}$$

where ϵ can be regarded as a random input drawn from a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. By using this trick, the objective function of the VAE can be interpreted in a close form as follows:

$$Regularisation\ cost = -D_{KL}(q_\phi(\mathbf{z}—\mathbf{x})||p_\theta(\mathbf{z})) = \frac{1}{2}\sum(1 + log(\sigma_\mathbf{z}^2) - \mu_\mathbf{z}^2 - \sigma_\mathbf{z}^2)$$
and
$$Reconstruction\ cost\ = -log(p(\mathbf{x}—\mathbf{z}) = \sum\left(\frac{1}{2}log(\sigma_\mathbf{x}^2) + \frac{(\mathbf{x}-\mu)^2}{2\sigma_\mathbf{x}^2}\right)$$

For a constant variance σ_x, the reconstruct cost becomes the least square error, similar to the convention autoencoder cost function. Figure 2.6 illustrate the VAE network.

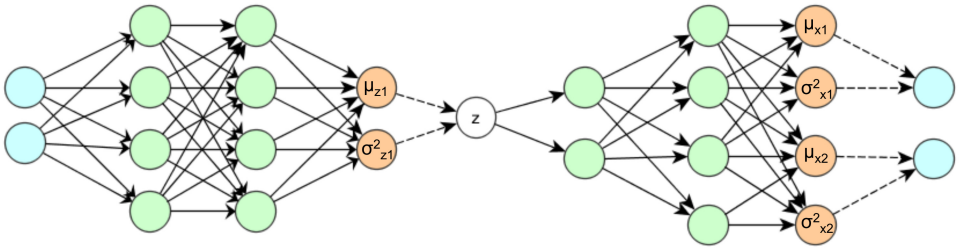

Figure 2.6: Illustration of VAE model

Chapter 3

Non-parallel Voice Conversion using Variational Autoencoder

3.1 Overview of Proposed System

In this section, the VC system proposed by the author is presented. The ultimate goal of VC system is to transfer the speaker individuality contained in the target voice to the source voice. As illustrated in figure 3.1, the proposed voice conversion system can be separated into two main stages: the first stage corresponds to the training phase and the second stage corresponds to conversion phase. Both stages share a common Acoustic Feature Processing unit, which is responsible for extracting the acoustic features from the input utterances. As illustrated in Figure 3.2, this unit consists of 3 smaller parts including STRAIGHT analysis, Mel-cepstral analysis, and mean-variance normalization. STRAIGHT, a high-quality vocoder system is used with the purpose to extract the spectrum from speech waveform. Then, the spectrum from STRAIGHT vocoder is transformed into Mel-cepstral coefficients (MCC) using SPTK toolkit. The output acoustic features are obtained by normalizing the MCC using mean-variance scaling.

The purpose of training phase of the proposed voice conversion system is to derive parameters for VAE model (consists of encoder and decoder parts) and dictionaries for target and source speaker. The detail of training process is mentioned in **Section 3.5**.

The conversion phase aims to generate the converted waveform (adapted to target voice) from the source waveform. Using the target dictionary obtained from training phase combine with the Activation matrix derived from source speech, the acoustic features of target-adapted speech is constructed. Finally, the waveform generation unit, as shown in Figure 3.3, does the reversal jobs to acoustic processing unit by taking the input acoustic feature then generating the output waveform.

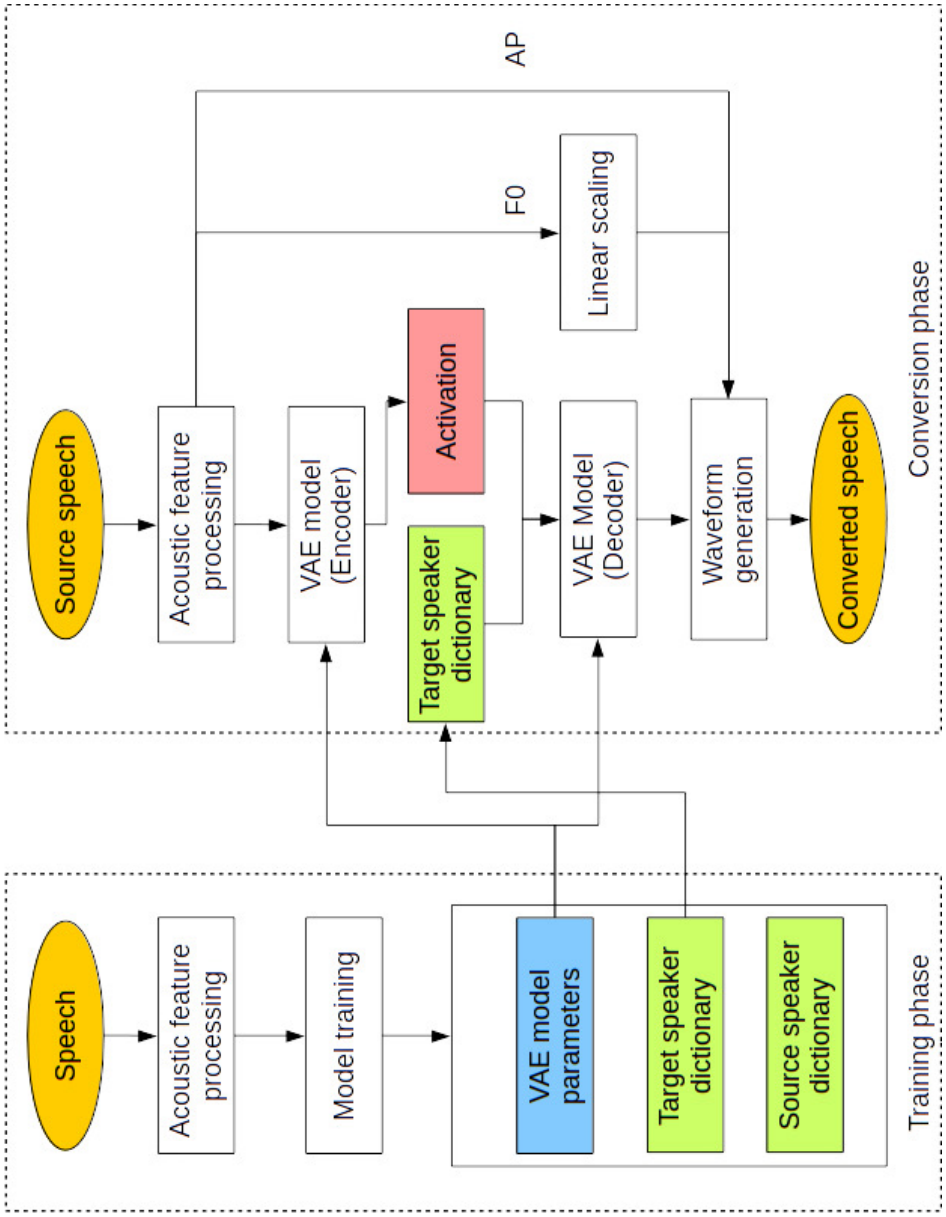

Figure 3.1: Block diagram of proposed voice conversion system

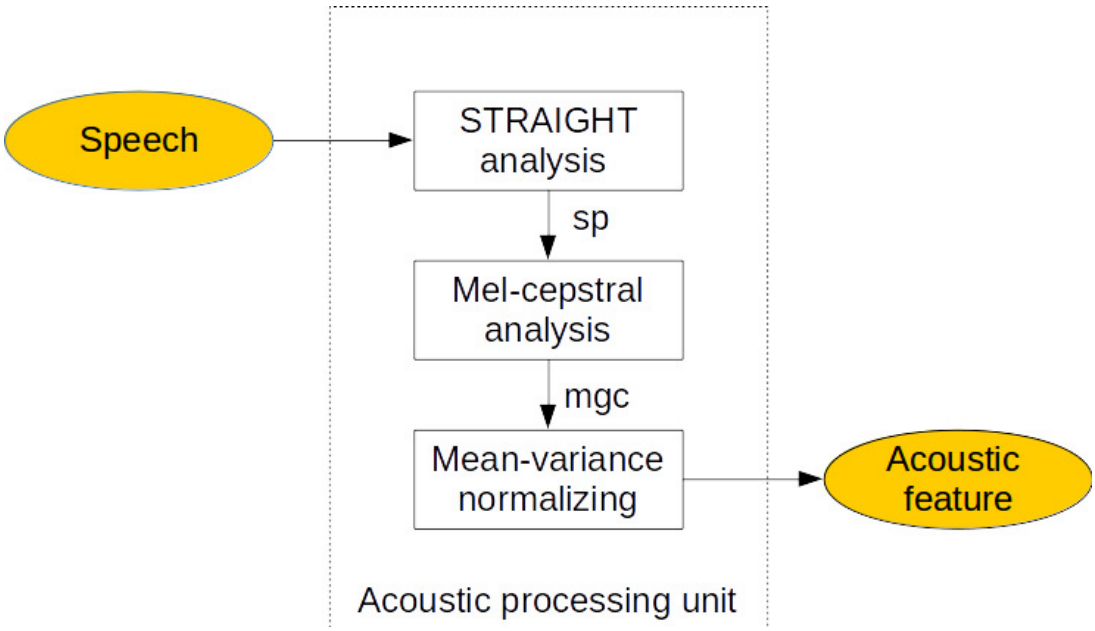

Figure 3.2: Block diagram of acoustic processing unit

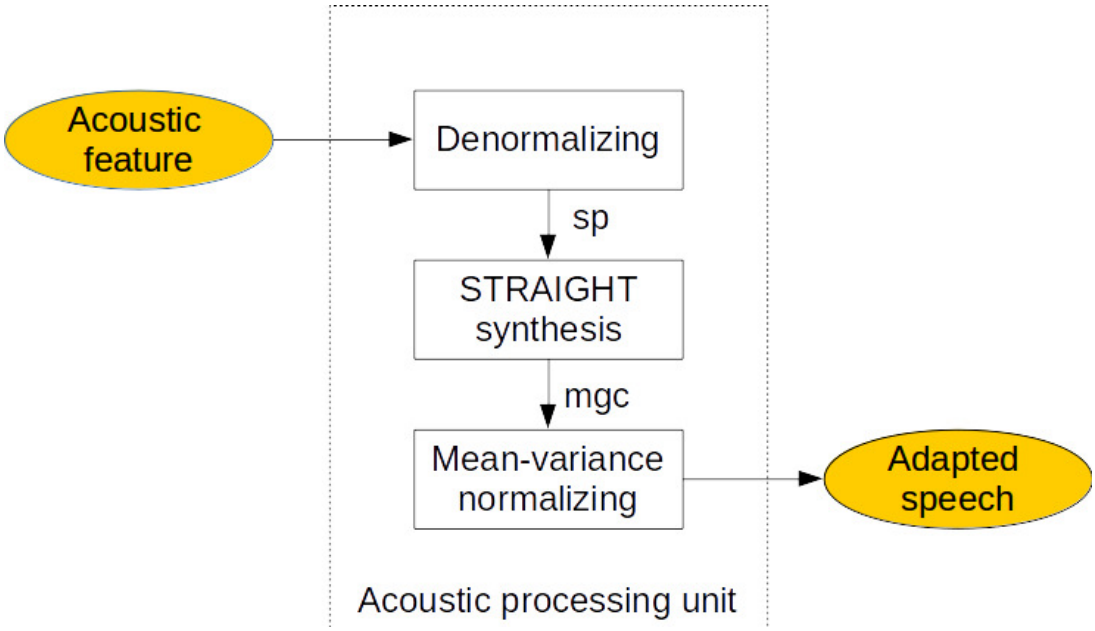

Figure 3.3: Block diagram of waveform generation unit

3.2 STRAIGHT Vocoder

The STRAIGHT system [18] is a high-quality vocoder based on the source-filter model. A vocoder is a specific-coder for human voice signal that relies on speech models and is focused on producing perceptually intelligible speech without necessarily matching the waveform [19]. The central idea behind the STRAIGHT vocoder is to extract spectral information that does not consist of the periodic structure in both the time and frequency domains [20]. In other words, STRAIGHT decomposes the speech into source information (pitch and aperiodicity) and spectral information (spectral envelope). With conceptual simplicity plus the flexibility of controlling of speech parameters, STRAIGHT system is a powerful tool for speech processing research as well as other speech-related application.

STRAIGHT, which stand for "Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum), consist of analysis and synthesis parts as illustrated in Figure 3.4. There are 3 fundamental concepts of STRAIGHT foundations. The first one is the reconstruction of time-frequency surface by a two-step procedure: 1) extract power spectra that minimize temporal variation by using a complementary set of time windows, 2) removal of frequency domain periodicity caused by source interference by inverse filtering in a spline space [18]. The second one is the accurate F0 extractor based on instantaneous frequency. The third is an excitation source design base on phase manipulation to reduce the buzzy timbre resulting from a conventional pulse excitation. For a more detail mathematical implementation of STRAIGHT, please refer to [18].

There are two main reasons for applying STRAIGHT to our system. Firstly, STRAIGHT provides a high-performance speech analysis/synthesis framework that is robust to speech parameters manipulation without introducing further degradation. This vocoder has been being applied in many speech application areas including voice conversion, speech recognition, speech synthesis, etc.. Secondly, the obtain speech spectrum from STRAIGHT is very smooth, which means that the MCC derived from the spectrogram are highly correlated among frames, thus enhancing the accuracy in parameters generation process.

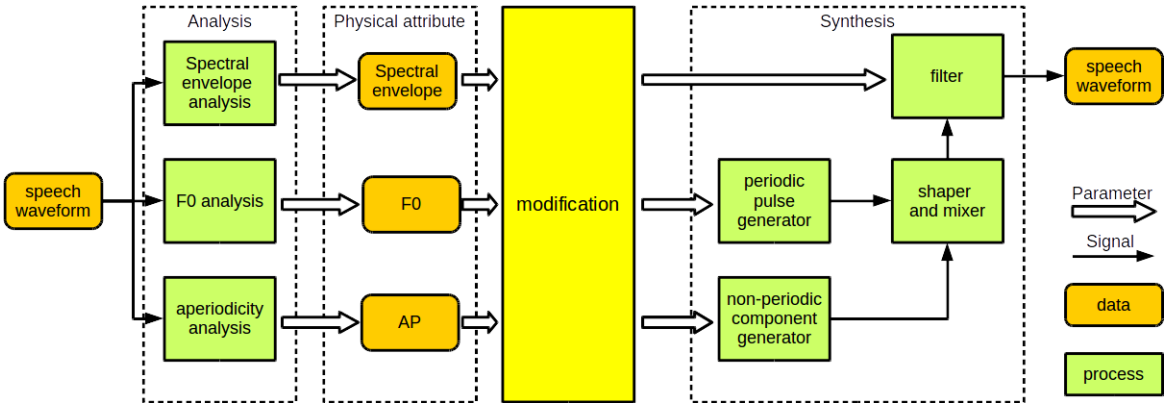

Figure 3.4: Flow chart of STRAIGHT vocoder system

3.3 Acoustic Feature Processing

3.3.1 Mel-generalized Cepstral Coefficients

Cepstral analysis is a popular feature extraction techniques based on the frequency domain. The cepstrum coefficients are defined as the coefficients of the Inverse Fourier transform of the log-magnitude spectrum:

$$S(e^{j\omega}) = F[s(n)]$$
$$C(m) = F^{-1}[log|S(e^{j\omega}|]$$

where $S(e^{j\omega})$ is the Fourier transform of signal $s(n)$, $C(m)$ is the cepstrum.

In the purpose of unifying the cepstral method with linear prediction method, a generalize-cepstral analysis has been proposed [21]. In addition, since human ear has higher resolution at low frequency, therefore a non-linear frequency scale is applied to compensate for this characteristic. The cepstrum $c(m)$ of a real sequence x(n) is defined as the inverse Fourier transform of the generalized logarithmic spectrum calculated on a non-linear frequency scale $\beta_\alpha(\omega)$ as follows:

$$s_\gamma(X(e^{j\omega})) = \sum_{m=-\infty}^{\infty} c_{\alpha,\gamma}(m) e^{-j\beta_\alpha(\omega)m}$$

The generalized logarithmic function is defined as:

$$s_\gamma(\omega) = \begin{cases} (\omega^\gamma - 1)/\gamma, & 0 < |\gamma| < 1 \\ log(\omega), & \gamma = 0 \end{cases}$$

where $X(e^{j\omega}$ is the Fourier transform of $x(n)$. The warped-scaled $\beta_\alpha(\omega)$ approximate the non-linearity of auditory frequency scale and is defined as the phase response of an all-pass system:

$$\Psi_\alpha(z) = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}}\Big|_{z=e^{j\omega}} = e^{-j\beta_\alpha(\omega)} \text{ where } \beta_\alpha = tan^{-1}\frac{(1-\alpha^2)sin\omega}{(1+\alpha^2)cos\omega - 2\alpha}$$

The speech spectrum $H(e^{j\omega})$ can be represent by the $M+1$ Mel-generalized cepstral coefficients as follows:

$$H(z) = \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\Psi_\alpha^m(z)\right)^{1/\gamma}, & 0 < |\gamma| \leq 1 \\ exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\Psi_\alpha^m(z) \quad , & \gamma = 0 \end{cases} \tag{3.1}$$

By selecting the appropriate (α, γ), the models spectrum will have the form of all-pole representation (LPC) or cepstrum representation. Table 3.1 describe the form of spectrum representation given the (α, γ) [21]. It is noted that the value of α is empirically selected based on the sampling rate to give good approximation to human auditory scale.

Table 3.1: Spectral representation based on Mel-generalized cepstrum

	$\alpha = 0$	$\|\alpha\| < 1$
$\gamma = -1$	all-pole	warped all-pole
$\gamma = 0$	cepstral	Mel-cepstral
$\gamma = 1$	all-zero	warped all-zero
$\|\gamma\| \leq -1$	generalized cepstral	Mel-generalized cepstral

In this thesis, the mel-cepstral representation ($\alpha = 0.42, \gamma = 0$) of spectrum is used since it shown the most influence on speech naturalness as described in [6]. Firstly, the speech signal is processed using STRAIGHT vocoder to calculate the spectrum of each frames. Smooth transition from frame to frame is ensured by overlapping of consecutive frames (the next and the previos frame). Then the obtained spectrum from STRAIGHT is represented as 60 mel-cepstrum coefficients using the equation 3.3.1. The Speech Signal Processing Toolkit (SPTK) [22] is used to calculate the Mel-cepstrum from the STRAIGHT spectrum.

3.3.2 Modulation Spectrum

Spectral envelope is known as the prime carrier of the phonetic information as well as the individuality of speaker. The modulation spectrum of speech is defined as the spectral analysis of temporal trajectories of the spectral envelope. The modulation spectrum is closely related to the dominant rate of change of the vocal tract shape. The modulation spectrum of continuous speech between 2 Hz to 8 Hz is mostly related to the phonetic information in speech [26]. Therefore, it is not surprising that the modulation spectrum around 4 Hz is most sensitive to human auditory system.

Being one of the most important features, the modulation spectrum has a wide range of application in speech research. In the automatic speech recognition field, the modulation spectrum can be applied to yields higher accuracy rate in the noisy environment [27]. In the field of speech synthesis, various studies have shown that the over-smoothing effect of synthesized speech originated from the degradation of modulation spectrum [24] [23]. Especially, the modulation spectrum also contributes to the speaker individuality as described in [25]. Therefore, by compensating the modulation spectrum of the synthetic speech for being close to the target natural speech, the quality of synthetic speech can be improved.

In this study, the modulation spectrum of parameter sequence x is defined as follow:

$$\mathbf{s}(\mathbf{X}) = \left[\mathbf{s}(1)^\top, \cdots, \mathbf{s}(d)^\top, \cdots, \mathbf{s}(D)^\top\right] \tag{3.2}$$

$$s(d) = [s_d(0), \cdots, s_d(f), \cdots, s(D_s)] \tag{3.3}$$

$$s_d(f) = abs(FFT(\mathbf{x}(d)) \tag{3.4}$$

3.4 Proposed Variational Autoencoder-based Voice Conversion

3.4.1 Dictionary-based voice conversion using Variational Autoencoder

The major drawbacks of NMF-based voice conversion is the requirement of parallel training data. This implies the NMF-based voice conversion may not be suitable for personalized S2ST device, where no parallel training is available. Furthermore, the use of DTW for aligning source and target utterance may introduce additional error which degrade the converted speech quality. Therefore, to overcome those issues, we aim to apply different method to decompose speech for using non-parallel dataset.

There are three important points in our proposed method. Firstly, we expand spectrum decomposition into non-linearity domain by using neural network with non-linear activation function (tangent hyperbolic):

$$\mathbf{X} = f_{dec}(\mathbf{A}^{(X)}\mathbf{Z}) \tag{3.5}$$

where $f_{dec}()$ is realized by a neural network.

In the next step, the activation matrix Z is extracted from the input spectrum also using a neural network:

$$\mathbf{Z} = f_{enc}(\mathbf{X}) \tag{3.6}$$

The parameters of encoder network f_{enc} and decoder network f_{enc} can be learned by jointly train the two networks as an Autoencoder. However, without any constraint on the activation matrix, the source and target dictionary cannot share the same activation matrix. In other words, the converted spectrogram cannot be constructed by target dictionary and activation matrix extracted from source spectrogram. Therefore, we introduce one additinal constraint by assuming the activation matrix has the standard norm distribution $\mathbf{N}(\mathbf{0}, \mathbf{I})$ over the whole utterance. This leads the network to have the form of Variational Autoencoder (VAE). The training objective function of our proposed network have the similar form of VAE model [15] as follows:

$$\mathcal{L}(\theta, \phi; \mathbf{x}_n) = -D_{KL}(q_\phi(\overline{\mathbf{z}}_n|\mathbf{x}_n)||p(\mathbf{z}_n)) + log\ p_\theta(\mathbf{x}_n|\overline{\mathbf{z}}_n, \mathbf{A}^{(X)}) \tag{3.7}$$

where the first term K_{LD} is the Kullback-Leibler divergence constraining the activation to have standard normal distribution, the second term is the log-probability of acoustic features x_n given the activation z_n and speaker dictionary $\mathbf{A}$. The speaker dictionary $\mathbf{A}$ is obtained by multiply global dictionary A_{global} with speaker specific *adjustment term* A_X as equation 3.4.1. Each target and source have separate speaker adjust ment term A_Y and A_X respectively.

$$A = A_{global} \cdot A_X \tag{3.8}$$

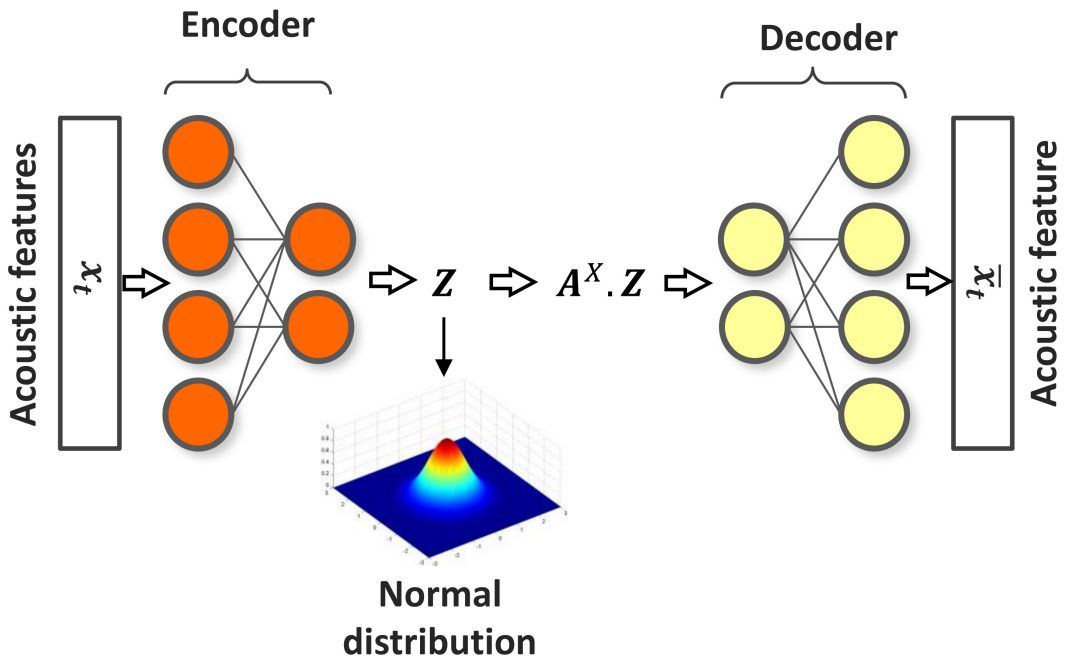

Figure 3.5: Proposed speech decomposition method using VAE

Training process is equivalent to iteratively estimate the autoencoder parameters θ and ϕ to maximize the equation (3.7):

$$\{\overline{\theta}, \overline{\phi}\} = \underset{\theta,\phi}{argmax}\ \mathcal{L}(\theta, \phi; \mathbf{x}_n) \tag{3.9}$$

Similar to the conversion process of NMF-based voice conversion, in our proposed method, the converted spectrogram is generated by multiply the target dictionary with activation extracted from the source utterance.

3.4.2 Modulation Spectrum-constrained training

As stated in the Section 3.3.2, modulation spectrum is an important acoustic cues for speech application. In the previous study by Dinh et al [6], a method for improving the naturalness of HMM-based synthesized speech based on Asymetric Bilinear model (ABM) is proposed. With the simmilar concept as NMF, ABM also factorizes the input data into two separate factors. In Dinh's study, various acoustic features are exchanged between natural speech and synthesized speech for comparing their effect on speech naturalness and intelligibility. The results from the subjective tests proved that the temporal modulation spectrum of MCC sequences has the most significant impact on speech naturalness.

To improve naturalness of the synthesized speech, we also incorporate the MS in the proposed model because of significance on speech naturalness [6][23]. In order to constrained the MS of output synthesized speech, an additional cost for MS is added to the

training function of VAE model. The modified log-likelihood function for VAEs model considering the MS is defined as follow:

$$\begin{aligned}\mathcal{L}_{ms}(\theta,\phi;\mathbf{x}_n) = -D_{KL}(q_\phi(\overline{\mathbf{z}}_n|\mathbf{x}_n)||p(\mathbf{z}_n)) + log\ p_\theta(\mathbf{x}_n|\overline{\mathbf{z}}_n,\mathbf{y}_n) \\ +w.log\ p(s(\mathbf{x})|\overline{\mathbf{z}}_n,\mathbf{y}_n)\end{aligned} \tag{3.10}$$

The final term in (3.10) explicitly constrains the model to increase the log-likelihood of the modulation spectrum conditioned on given latent variable $\overline{\mathbf{z}}_n$ and speaker identity y_n. Furthermore, we also assume that the modulation spectrum have a Gaussian distribution with diagonal covariance matrix: $s(x) \sim N(s(x)|s(\overline{x}), diag(\sigma_s))$. Therefore the final log-probability term in (3.10) can be expressed in closed-form:

$$log\ p(s(\mathbf{x})|\overline{\mathbf{z}}_n,\mathbf{y}_n) = -\frac{1}{2}\sum\left(log(2\pi\sigma_s^2) + \frac{(s(\mathbf{x}) - s(\overline{\mathbf{x}}))^2}{\sigma_s^2}\right) \tag{3.11}$$

3.4.3 Pretraining Procedure

To optimize the network parameter for efficiently encoding the input acoustic feature to the latent space z, the model is firstly pre-trained using global speaker dictionary only. Therefore, the all the voices will share the same speaker dictionaries in the pre-training phase. After the pre-training phase, the global speaker dictionary is multiplied with corresponded speaker specific term in the training process.

Chapter 4

Evaluation and Discussion

In this chapter, several objective measurements and subjective tests are conducted to evaluate the performance of the system.

4.1 Objective measurement

4.1.1 Formant frequency measurement

To measure the adaptation ability of VAE model, the formant frequencies of the adapted voice are compared with the original voice and target voice. The formant is the harmonic resulted from the resonance in the human vocal tract. As different speaker have distinct vocal tract shape, the formant frequencies can represent the speaker individuality. We measure the formant frequencies only for vowel sound where the formant frequencies are stable. The formant frequencies of vowel /o/ obtained from Pratt tool are reported in table 4.1. It can be clearly seen that the formant frequencies of **bdl** voice is shifted to those of **slt** voice. This proves that the proposed system can extract and transform the speaker individuality.

Table 4.1: Formant of vowel /o/ of natural and adapted speech

Formant	F1	F2	F3	F4
Natural bdl	862	1085	2272	3467
Natural slt	849	1433	3224	3994
bdl adapt to slt	**875**	**1439**	**3164**	**4014**

4.1.2 Pretraining evaluation

The purpose of this evaluation is to assess the effectiveness of pretraining method. In order to obtained the adapted voice, the system must first be able to precisely reconstruct

the source voice. We perform mel-cepstral distortion (MCD) measurement and the standardize Perceptual Evaluation of Speech Quality (PESQ) measurement from ITU [35] on adapted voice from the system with and without pretraining. The MCD is calculated by the following equation:

$$MCD[dB] = \frac{10}{log(10)}\sqrt{2\sum_{d=1}^{N}(c_d - \bar{c}_d)^2} \tag{4.1}$$

where N is the number of cepstral coefficients, c_d and $\bar{c}_d$ are reference and generated MCC respectively.

From the results in Figure 4.1 and 4.2, the mel-cepstral distortion and PESQ MOS score of the system with pretraining are clearly better. Therefore, the pretraining phase can improve the performance of the system.

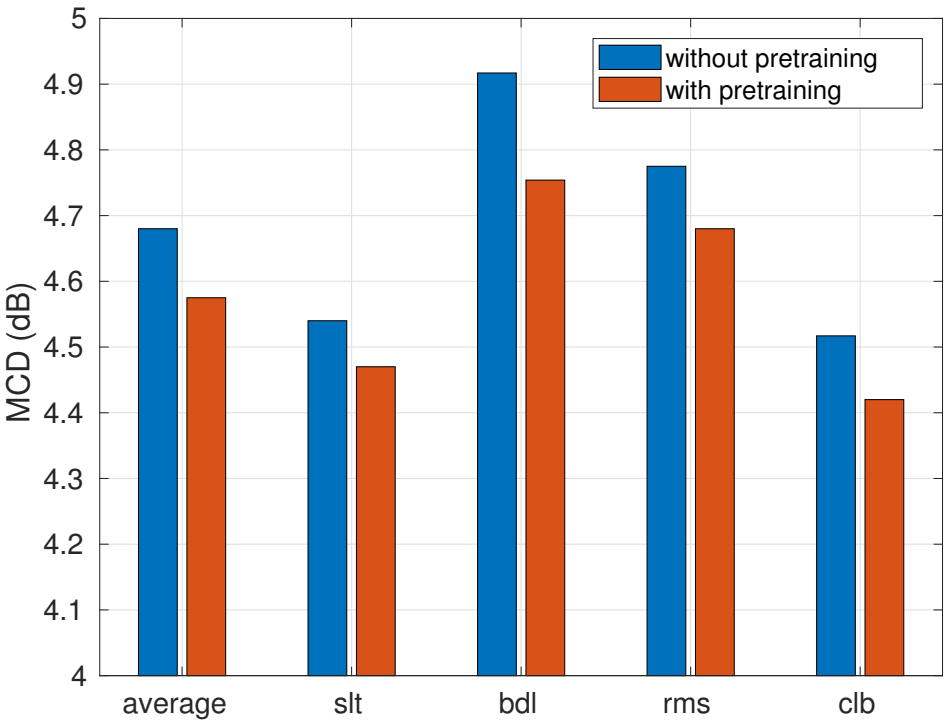

Figure 4.1: Mel-cepstral distortion of reconstructed speech from proposed system with and without pretraining (lower is better)

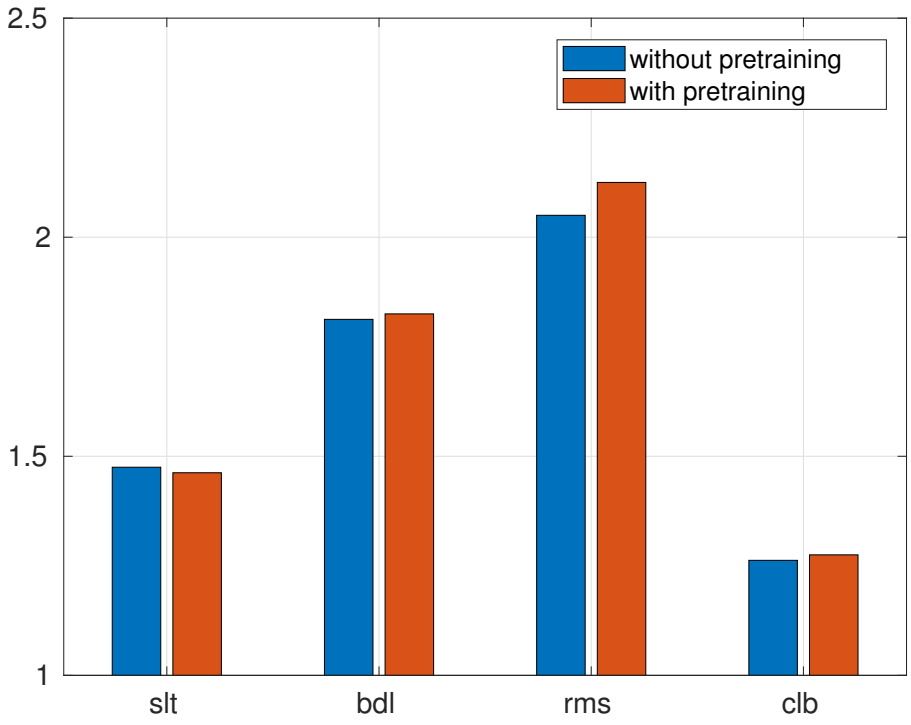

Figure 4.2: PESQ MOS score of reconstructed speech from proposed system with and without pretraining (higher is better)

4.1.3 Modulation Spectrum measurement

In order to assess the effectiveness of MS-constrained training, the MS of converted speech from VAE model with and without- MS-constrained training is measured. In both cases, the system is trained using the same set of data and training epochs. According to Figure 4.3,4.4 and 4.5, the MS of the 32^{th}, 64^{th}, 128^{th} spectral sequence at 4 Hz , which carries most of the linguistic information [36], from VAE model with MS-constrained training are better which indicate the effective of our proposed method. From Figure 4.6 and 4.7, it can be seen that the formant structure of synthetic speech using MS-constrained training is much clearer than the other, which implies a better synthetic speech quality.

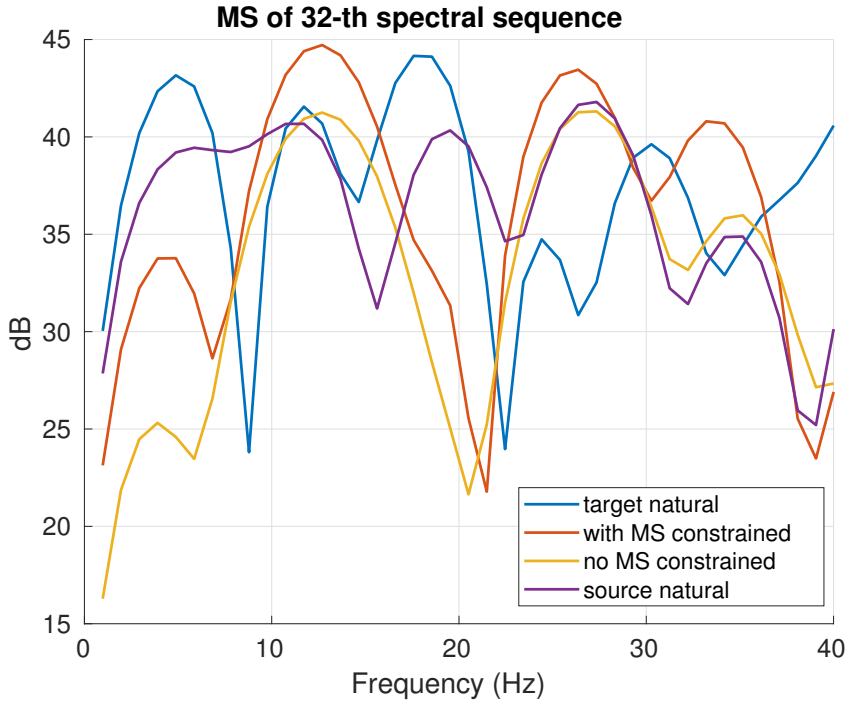

Figure 4.3: Modulation spectrum measurement of 32^{th} spectral sequence

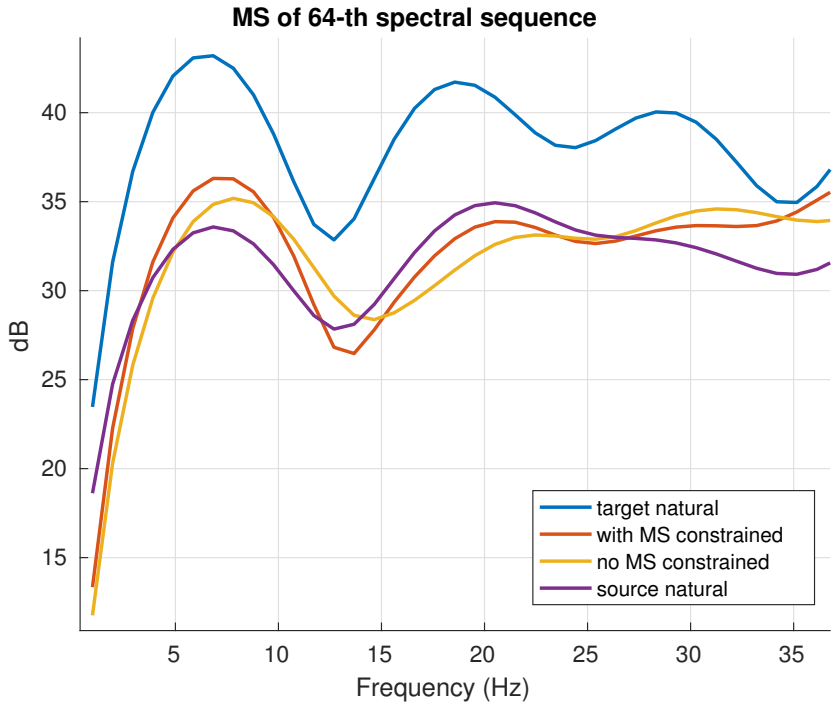

Figure 4.4: Modulation spectrum measurement of 64^{th} spectral sequence

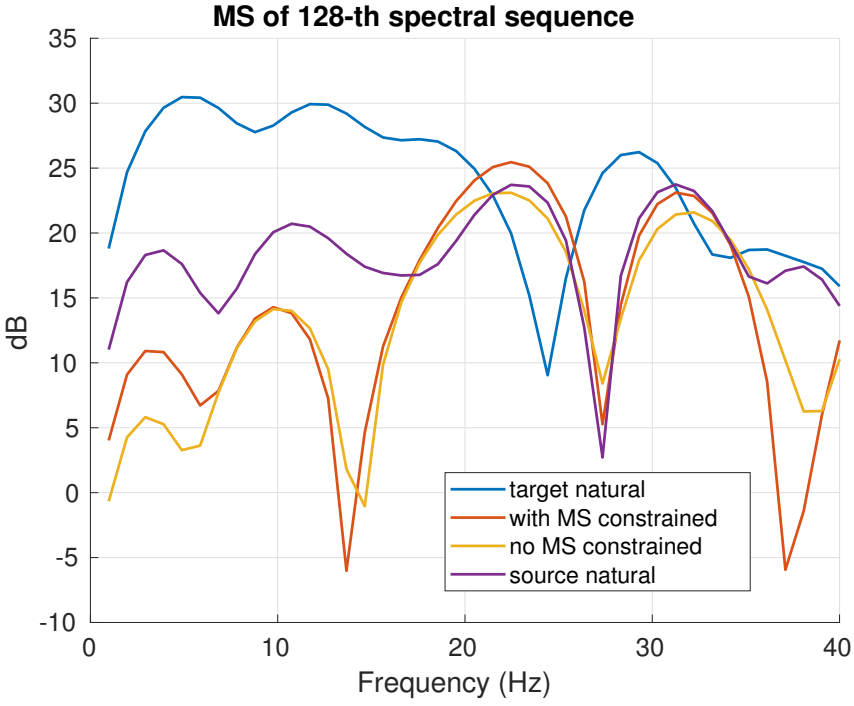

Figure 4.5: Modulation spectrum measurement of 128^{th} spectral sequence

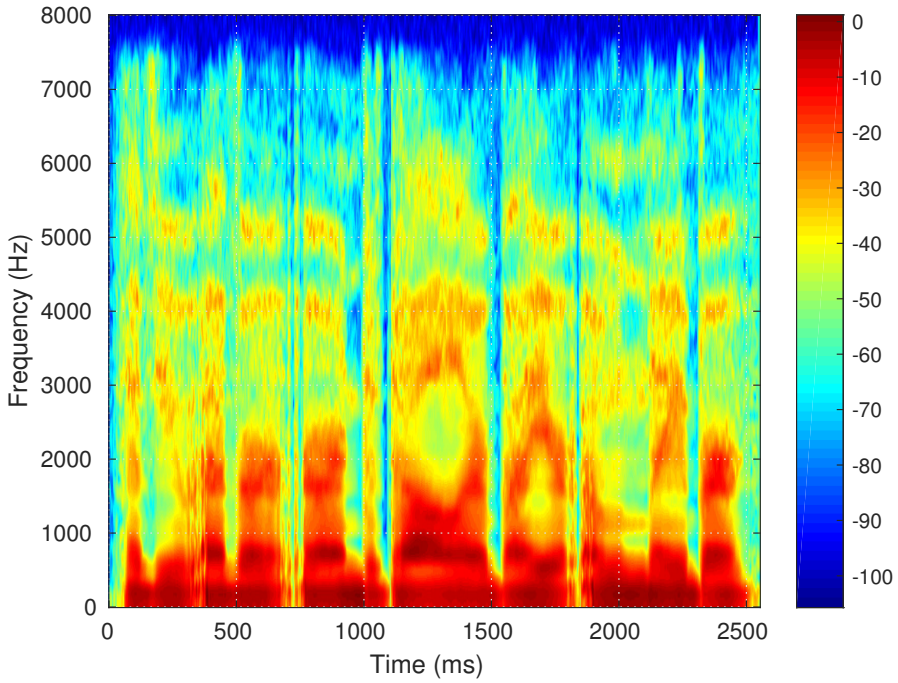

Figure 4.6: Modulation spectrum measurement of 128^{th} spectral sequence

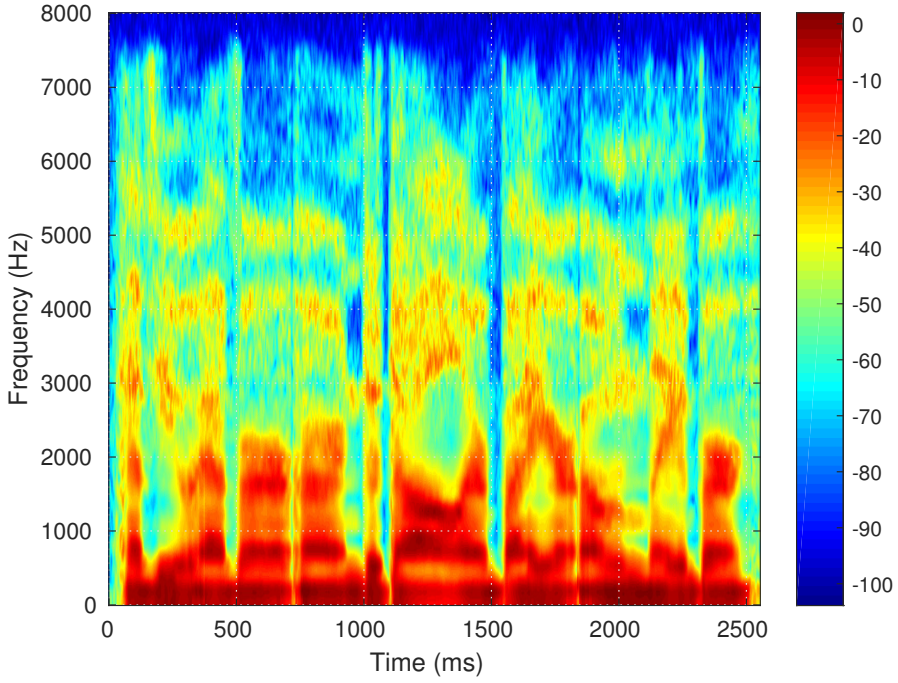

Figure 4.7: Modulation spectrum measurement of 128^{th} spectral sequence

4.1.4 Mel-cepstral distortion measurement

In this evaluation, we measure the MCD between the adapted speech and target speech by the baseline and our proposed method trained on different amount of data. In order to perform this test, the converted speech from the proposed method is aligned to target speech to DTW. The converted speech from the baseline method is already aligned, therefore no further alignment process is conducted. The measure MCD from 20 utterances is averaged to produce the final result. According to figure 4.8, the MCD of the proposed method is significantly lower than the proposed method although un-aligned training data is used.

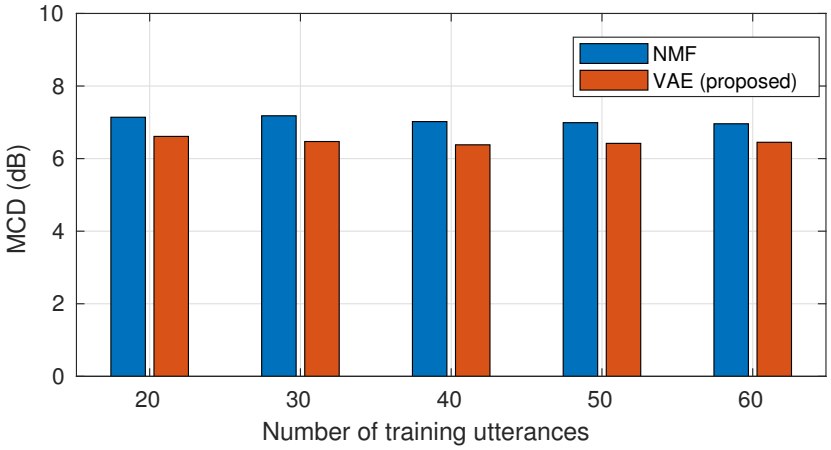

Figure 4.8: Mel-cepstral distortion of synthesized speech from proposed and baseline system using different amount of training utterances

4.2 Subjective measurement

4.2.1 Experimental setup

The baseline system

The baseline system is a NMF-based Voice Conversion using parallel data described in [17]. The dictionaries have $r = 100$ bases. 50 utterances of 2 speaker bdl (male) and slt (female) from CMU-ARCTIC database is used for training process. Alignment between source and target utterance is done by DTW. For input acoustic feature, the baseline method uses 513-dimension STRAIGHT spectrum. Aperiodicity (ap) remains unchanged while $logF_0$ is linearly scaled.

The proposed system

The configuration of the proposed system is shown in table 4.2. The decoder part have the same configuration of the encoder and in reverse order. The training database is the same as the baseline system. For the input acoustic features, 60 MCCs extracted from STRAIGHT spectrum using SPTK toolkit is used. Stochastic Gradient Descent (SGD) algorithm is used to optimize the parameters. The network is trained through 400 epochs, which takes approximately 20 minutes on GPU NVIDIA GTX1060.

Table 4.2: Network configuration

	units	activation
Input layer	128	linear
Encoder	1024-512-512-256-256	tanh
Output layer	180	linear

4.2.2 Similarity Test

In this experiment, the speaker similarity between natural voice and synthesized voice by different methods is evaluated. There are 20 stimuli for each voice conversion method. Each pair of stimuli contains one sample from natural voice and one sample from conversion system. Those two samples are selected randomly and contain different sentences. Not only the synthesized voice is compared to the target speaker but also the source speaker and target speaker. In summary, there are total 140 pairs of stimuli as described in table 4.3.

Table 4.3: Stimuli for each source-target pair in similarity test

Trials	Number of pair
Source voice - Target voice	20
Source voice - Source voice	20
Target voice - Target voice	20
Baseline - Source voice	20
Baseline - Target voice	20
Proposed - Source voice	20
Proposed - Target voice	20
Total	**140**

For each time, the listener will listen to a pair of stimuli and judge whether if those samples were produced by the same speaker or not. The listener is instructed to ignore the distortion and concentrate on identifying the voice [34]. After that, the listener indicates his/her confidence by a five-point scale:

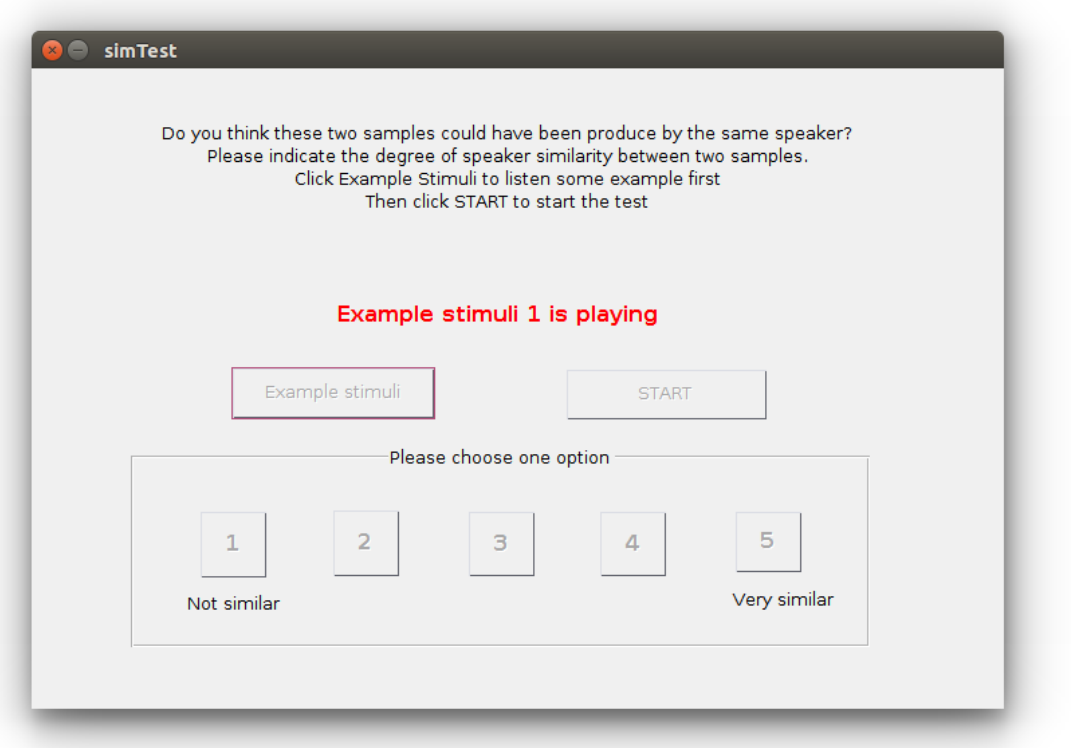

Figure 4.9: Graphic User Interface for Similarity Test

1. Completely different
2. Slightly different
3. Neither different nor same
4. Slightly same
5. Completely same

The GUI usage process for speaker similarity test is as follows:

1. Click **Start** to begin.
2. Listen to A and then B, wait until it finished.
3. Select one in five options base on the speaker similarity in two samples.
4. Then another pair of stimuli will be played automatically, wait until it finished.
5. Back to step 3 until the experiment is finished.

The experiment is carried out using a graphic user interface designed in Matlab (GUIDE) (figure 4.9). The listeners are seated in a soundproof room and listen to the samples through a headphone (HDA200, SENNHEISER) connected to an audio interface (FIREFACE UCX, Syntax Japan). The volume is set to a comfortable level. Each stimulus lasts for approximately 3 seconds. After every 50 pairs, the listener is asked to take a 2-minute break. The total time for the experiment including break time is around 30 minutes.

The listeners are asked to fill out a short questionnaire after completing the experiment with questions regarding name, student ID, gender, native language and any hearing problems if they have.

4.2.3 Naturalness Test

In this experiment, the naturalness between natural voice and synthesized voice by different methods is evaluated. There are 20 stimuli for each voice conversion method. Each pair of stimuli contains two samples from different voice conversion methods or natural voice. Those two samples are selected randomly and contain same sentences. In summary, there are total 120 pairs of stimuli as described in table 4.4.

Table 4.4: Stimuli for each source-target pair in similarity test

Trials	Number of pair
Baseline - Natural voice	20
Natural voice - Baseline	20
Proposed - Natural voice	20
Natural voice - Proposed	20
Baseline - Proposed	20
Proposed - Baseline	20

For each time, the listener will listen to a pair of stimuli A âĂŞ B (with A is the first utterance, B is the second one) and decide which one is more natural. After that, the listener select one of two options below:

1. A is more natural than B
2. B is more natural than A

The method of more natural sample receives the score of '1' while the other receives none. The experiment is carried out using a graphic user interface designed in Matlab (GUIDE) (figure 4.10). The listeners are seated in a soundproof room and listen to the stimuli through a headphone (HDA200, SENNHEISER) connected to an audio interface (FIREFACE UCX, Syntax Japan). The volume is set to a comfortable level. Each stimulus lasts for approximately 3 seconds. After every 50 pairs, the listener is asked to

Figure 4.10: Graphic User Interface for Similarity Test

take a 2-minute break. The total time for the experiment including break time is around 30 minutes.

The listeners are asked to fill out a short questionnaire after completing the experiment with questions regarding name, student ID, gender, native language and any hearing problems if they have.

The GUI usage process for naturalness test as follows:

1. Click **Start** to begin.
2. Listen to A and then B, wait until it finished.
3. Select one in five options base on the speaker similarity in two samples.
4. Then another pair of stimuli will be played automatically, wait until it finished.
5. Back to step 3 until the experiment is finished.

4.2.4 Results

The results of subjective test with 95-percent confidence interval are shown in Figure 4.12,4.11, and 4.13. The two-tail student t-test is used for analyzing the statistical significance of the results.

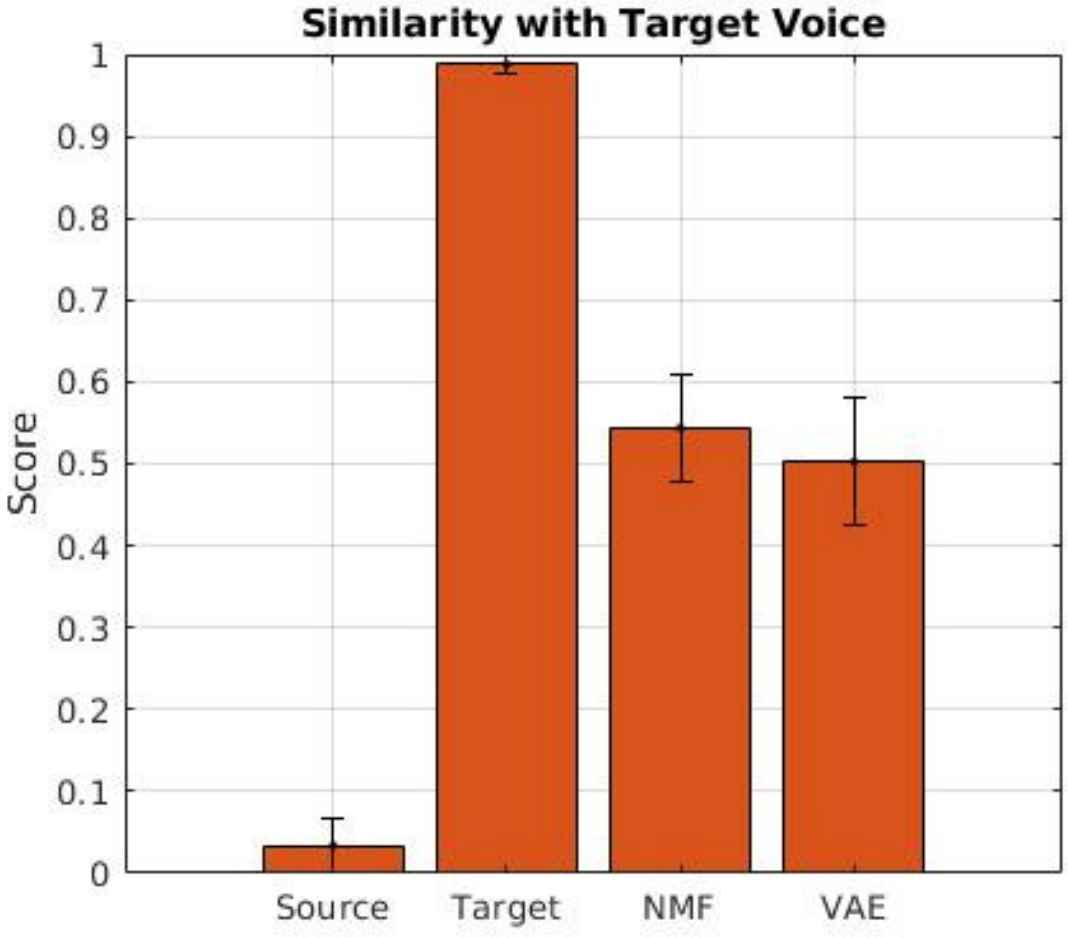

Figure 4.11: Similarity to target speaker with 95-percent confidence interval ($p = 0.44 > 0.05$).

4.3 Discussion

From the results in Figure 4.12,4.11, and 4.13, the subjective evaluations demonstrated significantly higher naturalness of the proposed VAE-based system over that of the NMF-based system. Meanwhile, the speaker similarity between two methods is comparable. These subjective results also conform with the objective results shown in Section 4.1.

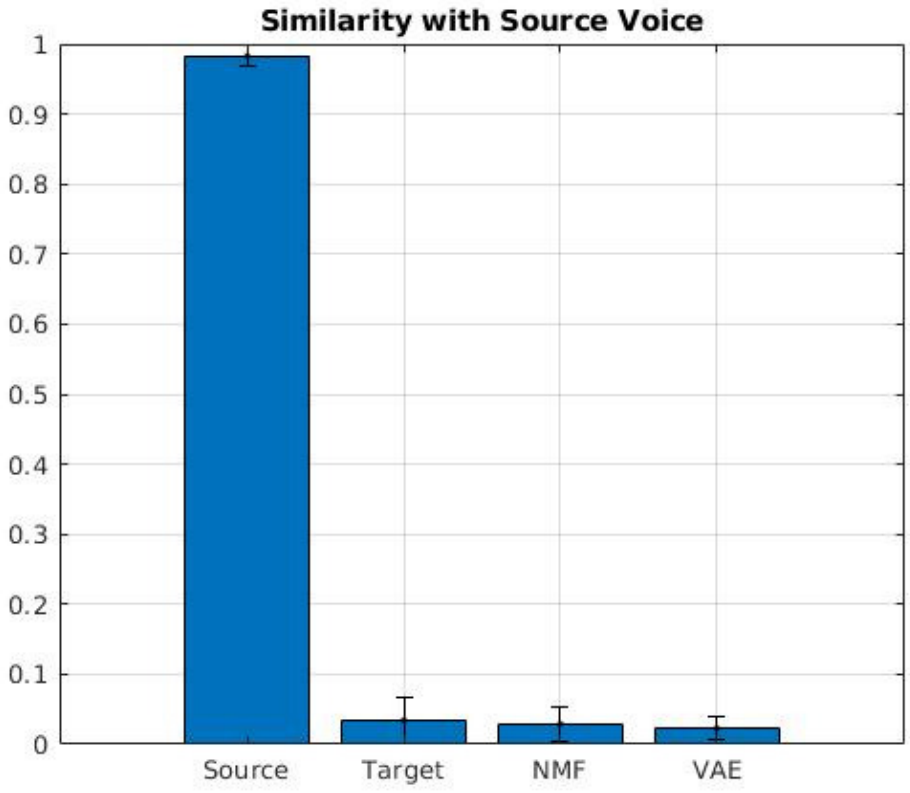

Figure 4.12: Similarity to source speaker with 95-percent confidence interval $p = 0.69 > 0.55$.

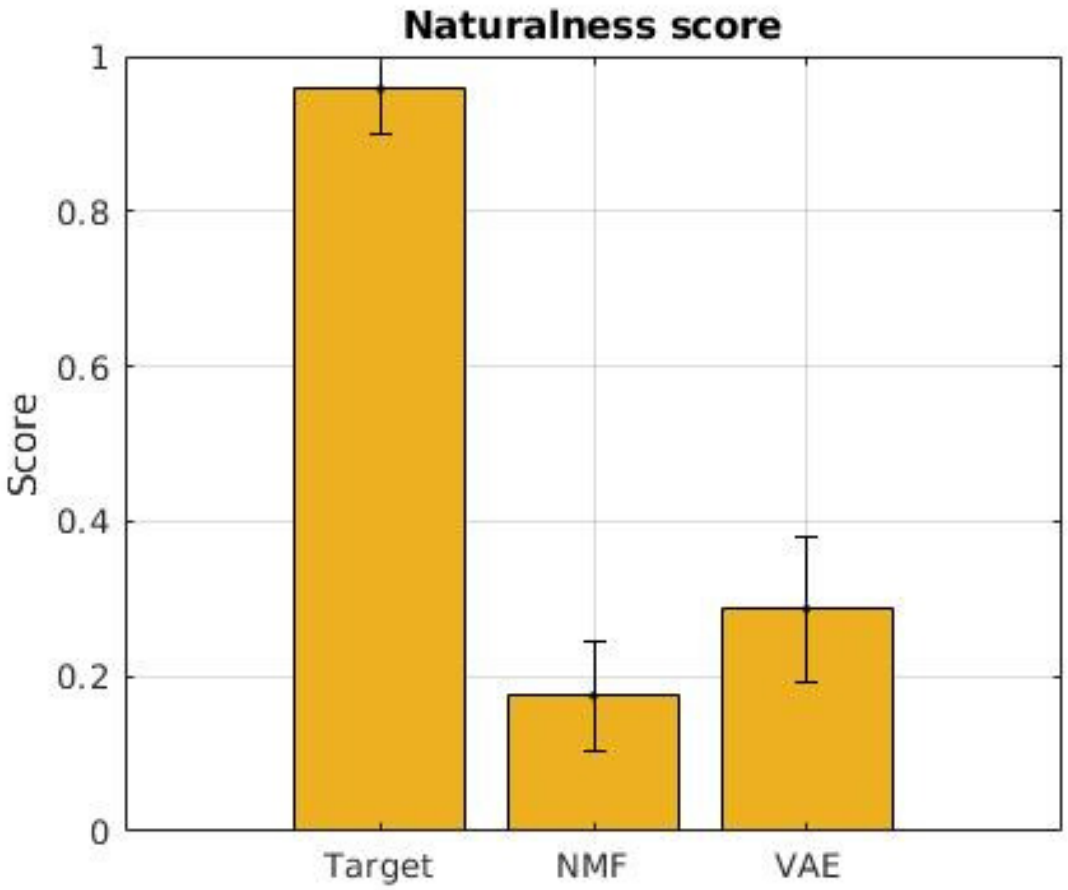

Figure 4.13: Naturalness MOS score with 95-percent confidence interval ($p = 0.04 < 0.05$).

Chapter 5

Conclusion

5.1 Summaries

In this study, we presented a dictionary-based voice conversion system for using with non-parallel training data based on the deep learning model Variation Autoencoder. From the experimental result, the MS of the synthesized speech using proposed training method is improved, indicating that the proposed training strategy is more efficient compared with the conventional method. The results from the subjective evaluation indicate that the proposed method can give the intended speaker individuality perception similar to the previous NMF-based VC system. And the naturalness of the converted speech using our proposed system achieve the average score of 2.87/10, much better than the NMF-based voice conversion whose average naturalness score is 1.75/10. However, when comparing with the average score of natural speech (9.58/10), there is still much room for improvement.

The advantage of this method is two-fold. Firstly, parallel training data are no longer required for dictionary-based voice conversion. Second, this method outperforms the conventional NMF-based voice conversion in term of naturalness while retaining comparable speaker similarity.

In conclusion, a voice conversion system utilizing the Variational Autoencoder model has been proved to achieve better-converted speech quality compared to the previous NMF-based method.

5.2 Contributions

The voice conversion system utilizing Variational Autoencoder model is proposed in this study. The proposed method can give the intended speaker individuality perception similar to the previous method using NMF but with a better naturalness which is a great enhancement. Although there is still much room for improvement, this study put the

first step toward the realization of the personalized S2ST device. Moreover, this work can contribute to much other application such as Story Teller System, Foreign Language Learning apps, etc.., all of which can give great improvement to human daily life.

5.3 Remaining problems

The final goal of this research is to construct high-quality cross-lingual voice conversion based on Deep Learning model. As the proposed method does not depend on linguistic information, in the next step, we will generalize our method to use with the cross-lingual dataset, making it suitable for personalized S2ST devices. In addition, since there is still a big gap between synthetic speech and natural speech, the cause that degrades speech quality must be further investigated. After that, a solution to improve speech quality will be proposed.

Bibliography

[1] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, J. Li, "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," APSIPA, 2014.

[2] A. F. Machado, M. Queiroz, "Techniques for Crosslingual Voice Conversion," IEEE International Symposium on Multimedia, 2010.

[3] N.W. Ariwardhani, Y. Iribe, K, Katsurada, T. Nitta, "Voice conversion for arbitrary speakers using articulatory-movement to vocal-tract parameter mapping," IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2013.

[4] W.S. Percybrooks, E. Moore, "A New HMM-Based Voice Conversion Methodology Evaluated on Monolingual and Cross-Lingual Conversion Tasks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Volume: 23, Issue: 12, Dec. 2015.

[5] B.Ramani, M.P. Actlin Jeeva, P. Vijayalakshmi, T. Nagarajan, "A Multi-level GMM-Based Cross-Lingual Voice Conversion Using Language-Specific Mixture Weights for Polyglot Synthesis," Circuits, Systems, and Signal Process, vol. 35, issue 4, pp. 1283-1311, 2016.

[6] A.T Dinh, M. Akagi, "Quality Improvement of HMM-based Synthesized Speech Based on Decomposition of Naturalness and Intelligibility using Non-negative Matrix Factorization," O-COCOSDA, 2016.

[7] P.C. Nguyen, T. Ochi, M. Akagi, "Modified restricted temporal decomposition and its application to low rate speech coding," IEICE Transactions on Information and Systems, vol. E86-D, no. 3, 2003.

[8] A. Kumar, A. Verma, "Using phone and diphone-based acoustic models for voice conversion: a step towards creating voice fonts," ICASSP, pages 720âĂŞ723, 2003.

[9] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," ASRU, 2003.

[10] H. Duxans, D. Erro, J. Perez, F. Diego, A. Bonafonte, and A. Moreno, "Voice conversion of non-aligned data using unit selection," TC-STAR WSST, 2006.

[11] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text independent voice conversion based on unit selection," ICASSP, 2006.

[12] A. J. Uriz, P. D. Aguero, A. Bonafonte, and J. C. Tully, "Voice Conversion using K-Histograms and Frame Selection,", Interpeech, 2009.

[13] D. Erro and A. Moreno, A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, issue 5, pp. 944-953, 2007.

[14] M. Zhang, J. Tao, J. Nurminen, J. Tian, and X. Wang, "Phoneme cluster based state mapping for text-independent voice conversion," ICASSP, 2009.

[15] D.P. Kingma et al: Auto-Encoding Variational Bayes, The International Conference on Learning Representations (ICLR), 2014.

[16] Z. Wu, E. S. Chng, H. Li, "Exemplar-based voice conversion using joint nonnegative matrix factorization", Multimedia Tools and Applications, vol. 74, pp. 9943-9958, 2015.

[17] S.Z. Fu, P.C. Li, Y.H. Lai, C.C. Yang, L.C. Hsieh. Y. Tsao, "Joint Dictionary Learning-Based Non-Negative Matrix Factorization for Voice Conversion to Improve Speech Intelligibility After Oral Surgery," IEEE Transactions on Biomedical Engineering, vol: 64, issue: 11, 2017.

[18] H. Kawahara, I. Masuda-Katsuse, and A. de CheveignÃĺ, "Reconstructing the speech representations using pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol. 27, no. 3-4, pp. 187-207, 1999.

[19] A.S. Spanias, "Speech coding: A tutorial review," Proceedings of IEEE, vol. 82, no. 10, pp. 1541-1540, 1994.

[20] H. Kawahara, "TANDEM-STRAIGHT, a research tool for L2 study enabling flexible manipulations of prosodic information,", Proceedings of Speech Prosody, pp. 619-628, 2008.

[21] K. Tokuda, T. Kobayashi, T. Masuko, S. Imai, "Mel-generalized Cepstral Analysis - A Unified Approach to Speech Spectral Estimation," ICSPL, 1994.

[22] K. Tokuda, Speech Processing Toolkit (SPTK), http://sp-tk.sourceforge.net

[23] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, S. Nakamura, "Postfilters to Modify the Modulation Spectrum for Statistical Parametric Speech Synthesis," IEEE/ACM Transactions on Audio, Speech, and Language Processing vol. 24, issue: 4, pp. 755-767, 2016.

[24] S. Takamichi, T. Toda, A. W. Black, S. Nakamura, "Parameter generation algorithm considering Modulation Spectrum for HMM-based speech synthesis," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[25] Zhi Zhu, Ryota Miyauchi, and Masashi Unoki, âĂIJAnalysis of Speaker Individual Differences on Modulation Spectrum,âĂİ Proc. 2015 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, Kuala Lumpur, Malaysia, February 2015.

[26] Hynek Hermansky, "The Modulation Spectrum in the Automatic Recognition of Speech," IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, 1997.

[27] L.C Sun, L.S Lee, "Modulation Spectrum Equalization for Improved Robust Speech Recognition," IEEE Transaction on Audio, Speech, and Language Processing, vol. 20, no. 3, 2012.

[28] C.C. Hsu, H.T. Hwang, Y.C. Wu, Y. Tsao, H.M. Wang, "Dictionary update for NMF-based voice conversion using an encoder-decoder network," International Symposium on Chinese Spoken Language Processing (ISCSLP), 2016.

[29] M. Blaauw, J. Bonada, "Modeling and Transforming Speech using Variational Autoencoders," Interspeech, 2016.

[30] F. Rosenblatt, "The Perceptron, A Probabilistic Model for Information Storage and Organization In The Brain," Psychological Review, vol. 65, No. 6, 1958.

[31] S. H. Mohammadi, A. Kain, "Voice Conversion using Deep Neural Networks with Speaker Independent Pretraining," IEEE Spoken Language Technology Workshop (SLT), 2014.

[32] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning internal representations by error propagation," Parallel Distributed Processing, vol 1: Foundations. MIT Press, Cambridge, MA, 1986.

[33] Y. Bengio, "Learning Deep Architectures for AI," Foundation and Trends in Machine Learning, 2009.

[34] M. Wester, Z. Wu, J. Yamagishi, "Analysis of the Voice Conversion Challenge 2016 Evaluation Results," Interspeech, 2016.

[35] http://www.itu.int/rec/T-REC-P.862/en, "P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,"

[36] H. Hermansky, "Modulation Spectrum in Speech Processing," Signal Analysis and Prediction, pp 395-406, 1998.

Publications

1. Ho, T.V and Akagi, M, (2018). "Non-parallel Training Dictionary-based Voice Conversion with Variational Autoencoder," Proceedings of International Conference (NCSP'18), Guam, US.