Doctoral Dissertation

Non-parallel semi-supervised voice conversion with controllable voice characteristics based on Variational Autoencoder

Tuan Vu Ho

Supervisor Professor Masato Akagi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science

September, 2021

Abstract

Voice conversion (VC), in a wide sense, is a method aims to modify the para-/non-linguistic information conveyed in the speech waveform while preserving the linguistic content. Some para-/non-linguistic information of speech can be mentioned as speech expressiveness and speaker individuality features such as age, gender, and accent. In this research, a VC model that focuses on the speaker individuality aspect of speech is studied. In a special case when the source and target voices are in different languages, a cross-lingual VC (CLVC) model that can efficiently work in multi-lingual must be used. This type of VC model is very useful in various applications such as personalizing speech-to-speech translator or language-learning platform. Due to the unavailability of parallel source and target data, conventional mapping methods cannot be applied. To solve this problem, non-parallel VC models have been actively studied in recent years. In contrast with the conventional mapping approaches, these non-parallel VCs aim to disentangle the linguistic information and speaker individuality from the speech waveform. After that, the source speaker individuality is swapped with the target one while the linguistic information in the target is preserved. The most straight-forward approach for CLVC is by cascading automatic speech recognition system and text-to-speech system. As speaker identity and text transcription are both required during the training process, this type of VC model can be referred to as a supervised approach. As another way, semi-supervised CLVC can be trained without text transcription, hence avoiding the use of expensive transcribed speech corpus. Although the semi-supervised CLVC approach can yield better applicability comparing with the supervised CLVC model in practice, however, its performance is often lower compared with the supervised approach. The common approach for semi-supervised CLVC is based on Variational Autoencoder (VAE), which can factorize the linguistic information and speaker information from acoustic features by applying regularization on the latent variables representing the linguistic information. However, most of the previous CLVC methods only focus on mimicking the target speaker individuality without being able to generate new speaker individuality. For some practical applications, such as accent conversion, the ability to actively generate new voice individuality as well as passively mimicking a particular target voice is much more useful than solely mimicking the target voice. Considering the pros and cons of previous studies, the objective goal of this study is to design a semi-supervised

CLVC, which is capable of both mimicking voice and continuously controlling the voice characteristics of generated speech. When modeling continuous controllable degrees of voice characteristics in CLVC, two primary problems must be addressed: (1) how to reliably extract and modify speaker voice individuality from different languages and (2) how to generate high quality speech waveform with desired voice characteristics in cross-lingual setting. To this end, the four following sub-tasks were carried out, in which the first three ones correspond to the first problem and the fourth one corresponds to the last problem:

- Method for non-parallel VC: investigate an effective VC model to mimic a target voice by factorizing linguistic information and speaker individuality information (passive VC).
- Controllable speaker individuality: investigate a method to extract voice characteristics and to generate new speaker individuality (active VC).
- Cross-lingual setting: investigate methods to apply the proposed non-parallel VC for cross-lingual settings with controllable voice characteristics.
- Methods for improving speech naturalness and speaker similarity: investigate methods to improve the performance of the CLVC model.

The main contribution of this study was providing an effective method for controlling the speaker individuality and several enhancements for CLVC. This study can be directly applied in various applications such as customizing audiobook and avatar voices, dubbing, movie industry, teleconferencing, singing voice modification, voice restoration after surgery, and cloning of voices of historical persons. Besides, the results from this study are beneficial for other VC fields such as providing a method for controlling speech intelligibility of speech enhancement models.

Keywords: Voice Conversion, Variational Autoencoder, Unsupervised Learning, Speaker Embedding, Controllable Voice Quality

Acknowledgment

Studying in Japan Advanced Institute of Science and Technology (JAIST) was a life-changing experience that I would never forget. This thesis was completed by the support of many people that I would like to express my sincere appreciation.

First of all, I would like to express my sincere gratitude to my supervisor, Professor Masato Akagi for his tremendous support since my first day at Japan Advanced Institute of Science and Technology (JAIST). Under his supervision, I have been given the opportunity to pursuit my own research ideas. He showed me how to have a precise scientific approach and more importantly, a positive attitude for doing research. His devotion to acoustic science has always been an inspiration for me to keep on exploring new knowledge. Without the guidance from Professor Akagi, I would not be able to fulfil my PhD course.

Then, I would like to extend my gratitude to my second supervisor, Professor Masashi Unoki, for his invaluable suggestion and comments throughout my research. The discussions with Professor Unoki encourage the critical thinking and make my research advance continuously.

I would like to express my indebtedness appreciation to Mr. President of Arcadia Inc. Seiichi Tenpaku, Dr. Ivan Setiawan, Dr. Toshio Hirai and everyone in Arcadia Inc., Osaka, for their numerous help during my internship. They have given me the chance to learn many new things and experience the warm working culture of Arcadia. Special thanks go to Mr. Hirai for his kindness and valuable comments on my research, even after my internship at Arcadia. We have shared many things during our lunch time together in the Toku-Toku restaurant and the Kuruma restaurant. I will never forget the magnificent view of Minoh mountain in the autumn season.

I am also grateful to Toya-san, Ngo-san, Kim-san, Huy-san, Xingfeng-san, Yongwei-san, Xue-san and other members of AIS lab for making my life be a joyful experience at JAIST.

I would like to thank my Japanese teacher, Kuniko Katajima sensei, for her kindness and precious support. Not only the Japanese lesson but also various support for my daily life that I received from her.

I would like to express my gratitude to JAIST for giving me the financial support to study and present my work at many conferences.

Lastly, I would like to express my love to my family for always being there when I needed them. I am especially thankful for my parent for their sacrifice and unconditional love throughout the years. Most of all, I thank my wonderful wife, Dieu Linh, as she has always be my great encouragement and support to complete my PhD course. To me, she is a gift from God and I am proud to cherish every single moment with her.

List of Abbreviations

ASR	Automatic Speech Recognition
CLVC	Cross-lingual voice conversion
CNN	Convolutional neural network
Conv1D	1D convolutional neural network
DNN	Deep neural network
GAN	Generative Adversarial Network
GLU	Gated linear unit
GMM	Gaussian Mixture Model
GV	Gloval variance
HLE	Hierarchical latent embedding structure
LSD	Log-spectral distortion
MCC	Mel-cepstral coefficients
MCD	Mel-cepstral distortion
MOS	Mean-opinion score
MS	Modulation spectrum
NPVC	Non-parallel voice conversion
PCA	Principal Component Analysis
RBM	Restricted Boltzmann machine
RMSE	Root-mean squared error
S2ST	Speec-to-speech translator
SE	Speaker embedding
SE-TJ	English-to-Japanese cross-lingual voice conversion
SJ-TE	Japanese-to-English cross-lingual voice conversion
StarGAN	Star Generative Adversarial Network
TTS	Text-to-speech
VAE	Variational Autoencoder
VC	Voice conversion
VCC2020	Voice Conversion Challenge 2020
VCTK	Voice Cloning Toolkit
VQVAE	Vector-quantized Variational Autoencoder
WN	WaveNet model

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Overview of Voice Conversion

Being the most convenient way for human to communicate, speech contains various information including the linguistic-, para- and non-linguistic information. The linguistic information convey the discrete information which can be explicitly represented by written language. Besides, para-linguistic information refers to the speech prosody that is intentionally added by the speakers to alter the meaning of the sentence, such as intonation, intention, and attitude. Finally, the remaining information in speech is regarded as non-linguistic information, which is often unconsciously added by the speaker such as speaking styles, emotion, and speaker individuality.

With the advent of artificial intelligent, speech is becoming a common human-machine interface for many smart devices, such as virtual assistant and humanoid robot. Speech perception and speech production are the two key technologies to achieve an effective speech communication between human and machine. On one hand, speech perception technique related to the extraction and processing of the information in the speech waveform. On the other hand, speech production technique aims to producing speech waveform that can be easily perceived by the listener. Despite much progress, the speech communication between human and machine is still by far from perfect. In the recent years, the speech production technologies have put more attention on personalized speech synthesis, which can generate speech with different voice characteristics by modifying the para- and non-linguistic information. As stated in [1], the para- and non-linguistic information are crucial for an effective human communication. Thus, the personalized speech can help to improve the user experience when communicating with the machine.

The two main technologies for personalized speech synthesis are the text-to-speech (TTS) and the voice conversion. The main task of text-to-speech technology is to generate speech with the linguistic information provided from the input text. Methods for TTS may vary from the simplest

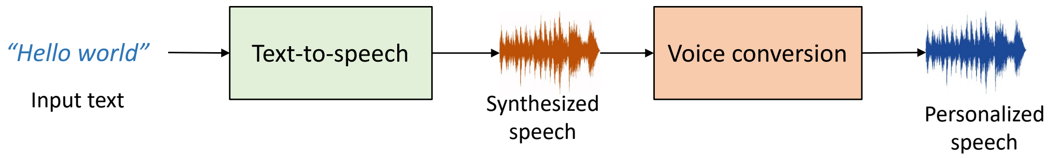


Figure 1.1: Typical process chain of personalized speech synthesis system.

unit-concatenation approach to the more sophisticated neural-network based approach. Initially, the TTS system is only capable of generating speech with a specific speaker individuality. Creating new speaker individuality usually requires re-training the whole TTS system with new data. Recently, studies in TTS area have proposed different multi-speaker models that can vary the speaker individuality in synthesized speech. However, these multi-speaker TTS models often require a large multi-speaker speech database, which is not ideal in practice due to the high cost.

Another technique for personalized speech synthesis is voice conversion, which aims at the control of non-linguistic information in the speech signal while maintaining other information unaffected. Differs from TTS system, the VC system takes the the speech waveform as input and produce another speech waveform with different speaker individuality. Since the VC system itself cannot produce the linguistic content, it is often located after the TTS system system. Figure 1.1 illustrates a typical chain for personalized speech synthesis. The VC system often require fewer speech data for adapting to a new speaker, hence it is more preferable for personalizing speech synthesis.

With the focus on controlling the non-linguistic information, an effective voice conversion system need to understand both the process and mechanism of speech production and speech perception. Due to this requirement, the voice conversion task is often more challenging comparing with the TTS task, which mostly focus on speech production side. At one end, a voice conversion system needs to extract the cues that are relevant to voice individuality from the speech signal. At the other end, the extracted information must be modified in the way that the converted speech sounds natural. Therefore, the speech model of voice conversion system should be capable of representing and modifying these cues efficiently.

The main objective for voice conversion system is to transform the identity of source speaker to that of target speaker while maintaining the naturalness of the converted speech. In practice, the constraint in the training data also poses several challenges for developing a good voice conversion system. In the early years, most voice conversion models based on the popular spectral mapping approach, which requires the parallel data for training process. The parallel data consist of several pairs of utterance from source speaker and target speakers that share a same sentence. The voice conversion model then learn an individuality transform function by minimizing the error between the synthesized waveform and target waveform. Since parallel data is not always available in practice, recent voice conversion studies have put more efforts on non-parallel approaches [2].

The non-parallel voice conversion is a more flexible but challenging framework. In this framework, the source and target utterances are not necessarily have the same linguistic content. There are two typical approach for non-parallel voice conversion: 1) construct the parallel data from non-parallel data [3, 4], and 2) speech features disentanglement approaches [5–9]. The first approach aims to find the source-target frame pairs from the non-parallel utterances using unit-selection [3] or iterative frame alignment method [4]. However, the performance of these approaches still need more improvement due to the imperfect alignments [4].

On the other hand, the speech features disentanglement approaches focus on finding a function to separate the linguistic information and non-linguistic information from speech features. The speech features disentanglement approach can be further divided into two categories: 1) the text-dependent approach (i.e. supervised approach) and 2) text-independent approach (i.e. semi-supervised approach). The first approach is usually a simple cascade of a Automatic Speech Recognition (ASR) system and a Text-to-Speech (TTS) system [5–7]. The ASR system takes the input speech waveform and extract the linguistic information represented as text or phoneme-related features. The TTS system then re-synthesizes the speech waveform from the extracted linguistic information and a label that represents speaker identity. This voice conversion approach requires a speech database with text-annotation for the training process. In the second approach, the linguistic information have to be extracted without any supervision from the text label, hence making it much more difficult than the first approach [8–10]. Therefore, the current performance of text-independent approaches are often lower than those of text-dependent approaches. The advantages of text-independent approach over text-dependent one is that it can be trained on non-annotated data, which can be easier to obtained in practice. Moreover, with the rapid advancement and growth of social media, an enormous amount of un-annotated speech data will be publicly available on the Internet. For these reasons, improving the performance of text-independent approach is becoming an essential topic for voice conversion research.

A special application of non-parallel voice conversion is the cross-lingual voice conversion [11–14], in which the target speaker does not speak the same language as the source speaker. This type of voice conversion is very useful in various applications, such as personalized speech-to-speech translator [1] or computer-assisted language learning leveraging accent conversion [15]. In the cross-lingual voice conversion, the phonetic system of source and target speaker are different, hence, parallel training data is impossible. A cross-lingual voice conversion should be able to efficiently model the speaker individuality in a multi-lingual setting. Therefore, the cross-lingual voice conversion is regarded as a more difficult task comparing with the intra-lingual voice conversion [15]. Due to this reason, the performance of cross-lingual voice conversion still remains to be improved.

1.2 Research Motivation

Despite the rapid development of non-parallel voice conversion technique, there are two major problems that still exist. The first issue is that most previous study on non-parallel voice conversion only focus on mimicking a particular target voice without the ability to change the voice characteristics. For some practical applications, such as accent conversion, the ability to actively generate new voice individuality as well as passively mimicking a particular target voice is much more useful than solely mimicking the target voice. The earliest attempt on controlling the voice characteristics in voice conversion system is the Eigenvoice-GMM model [16, 17], which model the speaker identity vector as the mixture of basis identity vectors. This model can flexibly control the speaker individuality by setting the weight parameter of the basis vectors. However, the Eigenvoice-GMM model still requires the availability of parallel training data, which only suitable for intra-lingual task.

The second issue that exists in most non-parallel voice conversion is the degradation in quality of converted speech. In general, over-smoothing problem is the major cause for this degradation. The over-smoothing problem is the result of using Mean-Square Error (MSE) [18,19] or maximum likelihood function [20] as the optimizing criteria. These optimizing objective function based on the assumption that the distribution of speech features follows the simple normal distribution. This over-simplified assumption causes the loss of spectral detail and makes the converted speech sound unnatural. One of the successful method to alleviate this problem is Generative Adversarial Network (GAN), which can avoid explicitly model the likelihood function of the data. However, the current state-of-art GAN-based voice conversion [21–23] does not support for speaker individuality control.

1.3 Research Objective

From the motivations mentioned in Section 1.2, the primary objective of this research is to design a semi-supervised cross-lingual voice conversion (CLVC), which is capable of both speaker individuality mimicry and continuously control of the voice characteristics. The proposed cross-lingual voice conversion can be advantage in various applications. For example, the voice from speech-to-speech translator device can be personalized to sounds like the input voice, hence improving the communication experience. With the ability to control the voice characteristics, such as pitch and gender, the proposed voice conversion model can be used to protect the privacy of the source speaker by modifying the original voice individuality.

To achieve the primary goal of this thesis, these two problems must be addressed: (1) how to continuously modify the speaker voice individuality from different languages and (2) how to generate high quality speech waveform with desired voice characteristics in cross-lingual setting. To solve these two problems, this thesis concentrates on designing a high-quality text-independent speech features disentanglement model. For the first problem, method to control the voice characteristics via a continuous speaker individuality representation for intra- and cross-lingual tasks is investigated. To improve the performance of voice conversion, this thesis focuses on methods for enhancing the spectral features and the prosody features of converted speech.

1.4 Research Methodology

In this thesis, the two problems mentioned in Section 1.3 are addressed by 3 sub-tasks: 1) propose a non-parallel voice conversion with controllable speaker individuality, 2) extend the proposed framework to cross-lingual domain, and 3) improve the performance of cross-lingual voice conversion. As listed below is the detail of 3 sub-tasks in this thesis, in which the first two sub-tasks related to the first problem and the last sub-task related to the last problems.

- **Non-parallel voice conversion with speaker individuality control:** The first study proposes a flexible non-parallel voice conversion (VC) system that is capable of both performing speaker adaptation and controlling speaker individuality. The proposed VC framework aims to tackle the inability to arbitrarily modify voice characteristics in the converted waveform of conventional VC model. To achieve this goal, the speaker embedding learned during the training process is used instead of one-hot encoded vectors to represent and modify the target voice's characteristics. Neither parallel training data, linguistic label nor time alignment procedure is required to train the proposed system. After training on a multi-speaker speech database, the proposed VC system can adapt an arbitrary source speaker to any target speaker using only one sample from a target speaker. The speaker individuality of converted speech can be controlled by modifying the speaker embedding vectors; resulting in a fictitious speaker individuality.

- **Cross-lingual voice conversion with hierarchical discrete latent vector quantized variational autoencoder:** In the second study, the non-parallel voice conversion (NPVC) is improved with hierarchical discrete latent vector-quantized variational autoencoder (VQVAE). The speech signal conveys several levels of information that localized at different temporal scale. However, previous studies on NPVC based on VQVAE use a single codebook to encode the linguistic information at a fixed temporal scale. Therefore, the converted speech may contain unnatural pronunciations which can degrade the linguistic information of speech. To tackle this problem, this study proposes the hierarchical latent embedding structure which comprises several vector quantization blocks operating at different temporal scales. When trained with a multi-speaker database, the proposed model can encode the voice characteristics into the speaker embedding vector, which can be used in one-shot learning settings.
- **Improve voice conversion quality with adversarial training scheme and F_0 injection:** In the final study, the adversarial training scheme of StarGAN [24] is adopted to alleviate the over-smoothing problem exist in VAE-based voice conversion model. In addition, to improve the accuracy of F_0 contour in converted speech, this study proposes the F_0-injection method to condition mel-spectrogram generation with auxiliary $\log F_0$ input. Another problem with cross-lingual voice conversion is the language-dependent speaker embedding behavior, which affect the linguistic information when converting voice. For this issue, a language embedding is introduced in addition to the speaker embedding for disentangling the language factor and speaker individuality.

The first work on non-parallel voice conversion with speaker individuality control discussed in Chapter 3 is published in the conferences: APSIPA 2020 [25], Autumn Meetings of Acoustic Japan Society 2019 [26], and Acoustic Symposium 2019 [27]. The second work on the cross-lingual voice

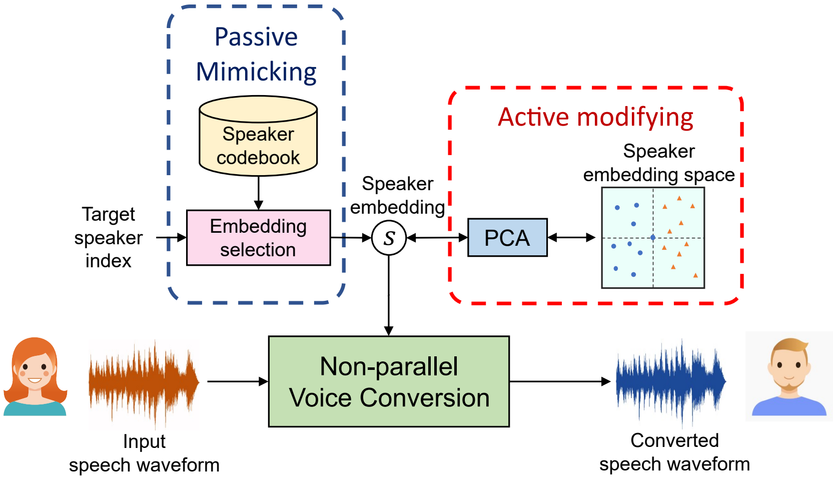


Figure 1.2: Overview framework of the proposed voice conversion system.

conversion discussed in Chapter 4 is published in the conferences: Joint Workshop of Blizzard Challenge and Voice Conversion Challenge 2020 [10], Autumn Meetings of Acoustic Japan Society 2020 [28]. The final work on improving the quality of converted speech is published in the Spring Meetings of Acoustic Japan Society 2021 [29] and the journal IEEE Access [30]. The overview of the proposed framework is shown in Fig. 1.2.

1.5 Dissertation Organization

The rest of this thesis is organized as follows:

- Chapter 2 provides a basic principle of voice conversion, followed by a comparison between voice conversion techniques. Then the review of the state-of-art method on non-parallel voice conversion is provided.
- Chapter 3 proposes non-parallel voice conversion with speaker individuality control, which is the main framework of this study.
- Chapter 4 proposes the extension of non-parallel voice conversion for cross-lingual task, based on the vector-quantized variational autoencoder.
- Chapter 5 proposes the improving method for cross-lingual voice conversion.

Finally, Chapter 6 concludes the dissertations with the summary on the contributions and future research direction. The structure of this thesis is illustrated in Fig. 1.3.

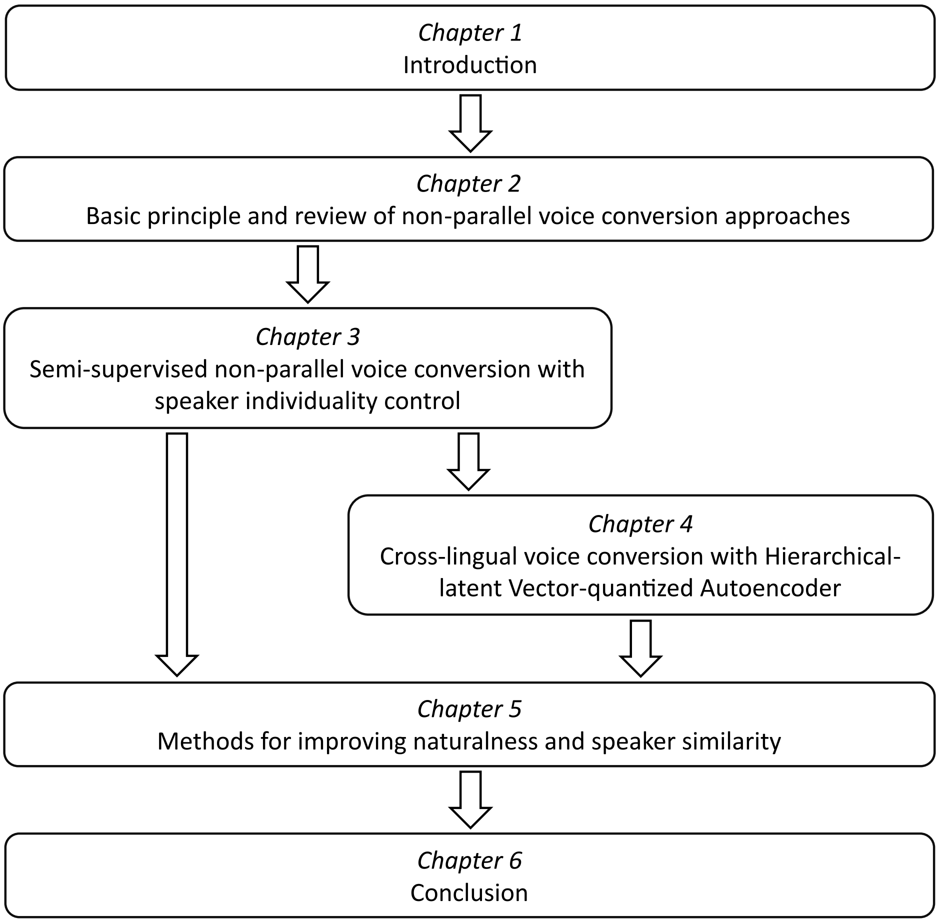

Chapter 1
Introduction
Chapter 2
Basic principle and review of non-parallel voice conversion approaches
Chapter 3
Semi-supervised non-parallel voice conversion with speaker individuality control
Chapter 4
Cross-lingual voice conversion with Hierarchical-latent Vector-quantized Autoencoder
Chapter 5
Methods for improving naturalness and speaker similarity
Chapter 6
Conclusion

Figure 1.3: The overview structure of the thesis.

Chapter 2

Literature Review

This chapter summarizes the background knowledge of voice conversion technique and reviews on the state-of-art methods for non-parallel conversion model. The chapter is organized as follows:

Section 2.1 briefly introduces the speech production mechanism to explain the focus of voice conversion system. Section 2.2 introduces the general flow of voice conversion system and discusses the interesting application of voice conversion. Section 2.3 discusses on the evaluation methods for voice conversion models.

2.1 Speech production mechanism

Whenever someone utters a sentence, they provides not only the message that made up the meaning of the sentence, but also the information about themselves as a person. Recordings from different speakers may sound very different even if it contains the same sentence. This is because the speech production involves the neural, physiological, and physical systems of a specific individual [31]. Differences in these systems contributes to speaker individuality in speech signal, which can be exploited by the listener to identify the characteristics of the speaker, such as age, gender, accent, language, emotion and health state. As the voice conversion system operates on the raw speech waveform, the physical speech production system is focused in the voice conversion study.

The physical of speech production system can be described by the source-filter model [32] as shown in Fig. 2.1, which consists of two primary components: the sound source, such as the vocal folds, and a time-varying acoustic filter, the vocal tract. For a voiced sound, the sound source is the periodic waveform produced by the regular vibration of the vocal folds and the filter is the whole vocal tract shape. For un-voiced sound, such as fricative and plosive sound, the sound source is white noise created by the air turbulence from the constriction in vocal tract, and the filter is the rest of vocal tract after the constriction. The resonance of the time-varying filter

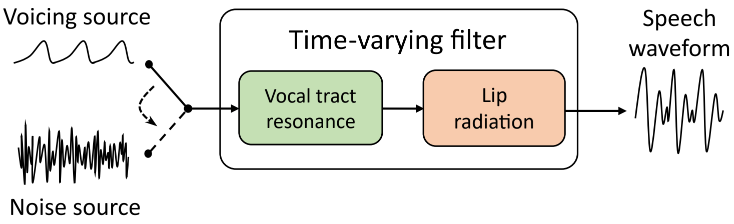


Figure 2.1: Source-filter model

emphasizes or attenuates the source signal at some frequencies and creates the formants in speech. The locations of formant correspond to a specific vowel. From the source-filter model, it is apparent that the generation of speech waveform can be controlled via the source signal and the filter shape. The source signal can be characterized by the fundamental frequency, pitch contour and intensity, while the filter shape can be represented by the spectral envelope. Previous study [33] has shown that cues for speaker individuality can be found in both the fundamental frequency contours and the spectral envelopes. For this reason, an effective voice conversion is expected to modify both the source-related features and filter-related features to transform the speaker individuality in speech.

2.2 Overview of Voice Conversion System

Voice conversion is a process that transform the speaker individuality in the source speech to sound like it was uttered by the target speaker. Voice conversion can be regarded as a learning problem that consists of two phases: the offline training phase and the run-time conversion phase. Figure 2.2 depicts the typical framework of a voice conversion system. Typically, there are two main modules in a voice conversion system, namely the analysis/synthesis system or the vocoder, and the conversion model. Since the voice conversion model is a parametric and data-driven model, its parameters need to be optimized in the training stage. A set of speech data, which contains the utterances from source and target speakers, is used to train the voice conversion model. The speech analysis of vocoder extracts the relevant acoustic features set from the input speech waveform. The speech features consist of the spectral features, e.g. mel-cepstrum coefficients, and the source features, e.g. fundamental frequency. The parameters of the voice conversion model are iteratively updated to optimize a predefined objective function, such as mean-square-error between converted features and target features. In the run-time conversion phase, the trained voice conversion model input the

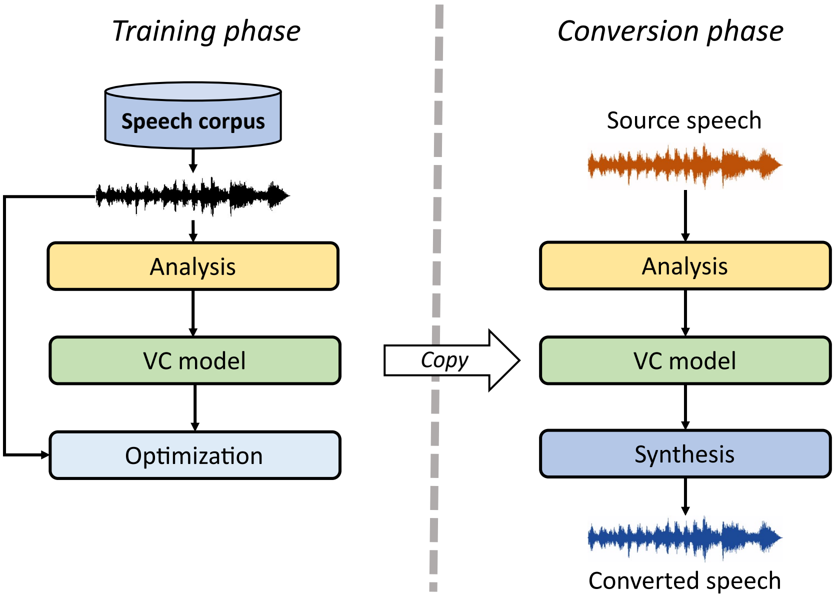


Figure 2.2: General framework of a voice conversion system.

speech features from the analysis and generates the converted speech features. The converted speech features is then re-synthesized to the converted speech waveform.

As mentioned above, a typical voice conversion systems consists of two modules: the a speech synthesis/analysis and a voice conversion model. In the following sections, the review of the typical approaches for each module is provided.

2.2.1 Speech Analysis and Synthesis

The speech analysis and synthesis module, or speech vocoder, is a signal processing system designed to synthesize the speech waveform from the feature representation. The purposes of vocoder in voice conversion system are: 1) extract the speech features, e.g acoustic features and prosodic features, and 2) re-synthesize speech waveform from speech features. Most of the voice conversion models are designed to operate on acoustic features. Therefore, an effective vocoder should be able to reconstruct the speech waveform from extracted speech features with high naturalness. Moreover, the extracted speech features should be allowed to be flexibly modified without affecting the quality of generated waveform. The most popular vocoders for voice conversion can be categorized into two types: the deterministic vocoder and data-driven vocoder.

Deterministic vocoder

The most common deterministic vocoder is the STRAIGHT vocoder [34], which is based on the source-filter model with mixed excitation signal. This vocoder is a voice-specific coder which focuses on producing perceptually intelligible speech without necessarily matching the waveform. It decomposes the speech signal into three components: the fundamental frequency, the speech spectrum and the aperiodicity. It features the pitch-adaptive analysis window to reduce the inference from the excitation signal, therefore, the extracted speech spectra is very smoothed and can be easily controlled. Due to its high-quality and flexibility, STRAIGHT vocoder is widely used in text-to-speech systems [35, 36] as well as voice conversion systems [37–39].

Data-driven vocoder

With the advent of deep learning techniques, speech generation can be entirely done using deep neural network. The first and most famous neural vocoder is the WaveNet architecture [40], which can predicts the speech waveform from the past samples and can be controlled by the common speech features, such as mel-cepstral coefficients. The architecture of WaveNet consists of stacks of causal dilated convolutional layers to achieve the a wide receptive field. The auto-regressive WaveNet model estimates the joint distribution of the quantized speech waveform sample $\mathbf{x} = [x_1, ..., x_T]$ as:

$$p(x) = \prod_{t=1}^{T} p(x_t \mid x_1, ..., x_{t-1}). \tag{2.1}$$

The distribution of each audio samples x_t is conditioned on the samples at all previous timesteps. Since the autoregressive model can only output one sample at each forward step, the generation speed of WaveNet is very slow. The recently proposed vocoder models, such as Parallel WaveNet, WaveGlow, and Parallel WaveGAN, have greatly improve the speech quality and the generation speech.

2.2.2 Voice conversion model

Being the central part in voice conversion system, the voice conversion model aims to modify the speech features in order to change the speaker individuality while maintaining the linguistic information. Since the low-dimensional prosody features such as fundamental frequency can be transform using simple linear mapping, most voice conversion models focus on modifying the

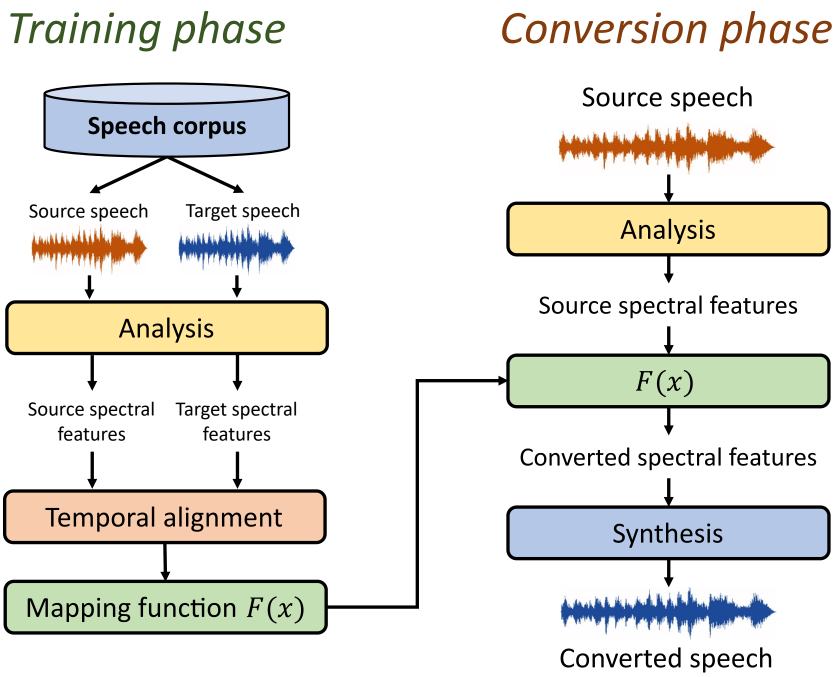


Figure 2.3: Typical flow of spectral mapping approach.

high-dimensional spectral features. Depends on the training data, the voice conversion model can be categorised into parallel and non-parallel methods:

- **Parallel voice conversion model** is trained to minimize the frame-by-frame error between the converted acoustic features and target ones. Since the mapping function for spectral features is focused, these models can be regarded as a spectral mapping function between source and target acoustic features. These models require the parallel training data, which consists of pairs of linguistic-identical utterance from source and target speakers. For frame-based methods, Dynamic Time Warping algorithm is used to aligned the source and target spectral features to account for the duration mismatch.
- **Non-parallel voice conversion model** aims to learn the disentanglement between the linguistic information and speaker individuality in the acoustic features. As opposed to the parallel methods, these models has more flexibility as they can be trained with both parallel and non-parallel data.

The overview for spectral mapping and feature disentanglement approaches are described in the following parts.

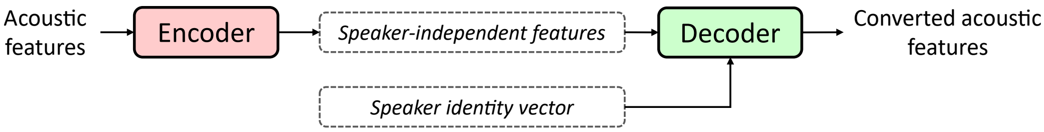


Figure 2.4: Typical framework for speech feature disentanglement approach.

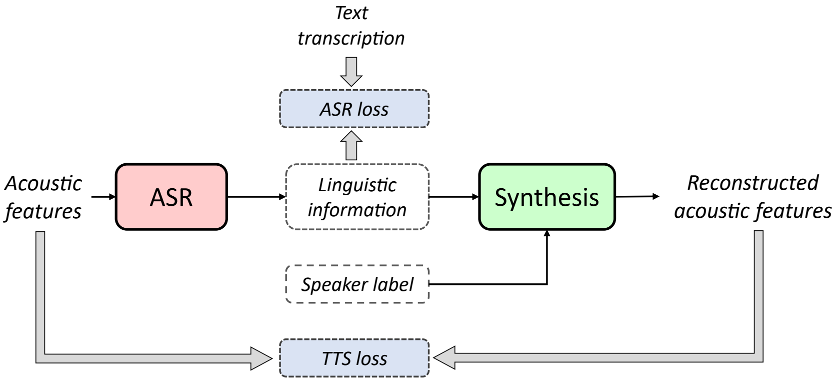
2.2.2.1 Spectral mapping approach

The early studies of voice conversion focus on the spectral mapping between the aligned source speaker's and target speaker's spectrum. Two parallel utterances are aligned in a frame-by-frame basis using Dynamic Time Warping to account for the difference in speech duration. Then a mapping function takes the source speaker spectrum as input and outputs the converted spectrum. By minimizing the error between the converted spectrum and target speaker spectrum, the mapping function can learn the transformation between the source speaker individuality and the target one. Figure 2.3 describes a typical processing flow of the spectral mapping approach. Initially, the vector quantization [41] and fuzzy vector quantization [42] approaches have been successfully applied to learn the mapping function. When more data is available, statistical parametric approaches such as Gaussian mixture model [37, 43–45], Hidden Markov model [46] and partial least square regression [47] can improve the performance of voice conversion model with better speech naturalness and speaker similarity.

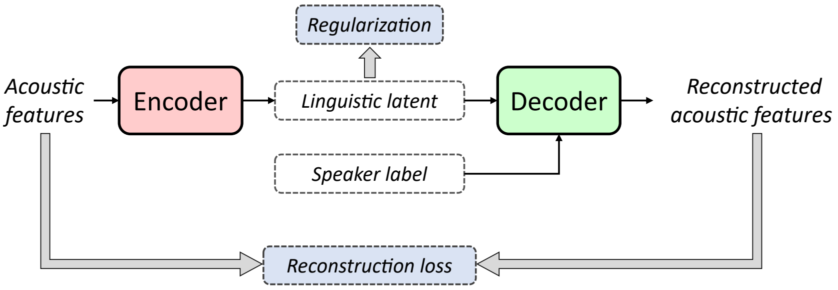
2.2.2.2 Feature disentanglement approach

In the feature disentanglement approach, the acoustic features is first encoded into speaker-independent (SI) features representation, such as text, Phonetic Posteriorgram (PPG) [48–50], or latent code [10, 25, 51]. The decoder then estimate the acoustic features with the SI features, conditioned on the speaker identity vector, e.g one-hot vector, i-vector, or neural speaker embedding. During training, the decoder learns the mapping function between the SI features and the corresponding acoustic features of the same utterance, therefore, parallel data is not needed. The encoder and decoder are usually trained with a multi-speaker speech corpus. Figure 2.4 illustrates the typical framework of speech feature entanglement approach.

The voice conversion model based on feature disentanglement method can also be further categorized into fully-supervised [48–50] and semi-supervised approaches [10, 25, 51] depending on the the availability of text-transcription. In fully-supervised models as shown in Fig. 2.5(a), the automatic speech recogniser (ASR) is trained to map the acoustic features



(a) Typical training pipeline of supervised non-parallel voice conversion model. The inputs consist of acoustic features, text-transcription and speaker label represent speaker identity. The ASR and synthesis modules can be trained independently.



(b) Typical training pipeline of semi-supervised non-parallel voice conversion model. Only acoustic features and speaker label are used as inputs.

Figure 2.5: The training pipeline of (a) supervised non-parallel voice conversion model and (b) semi-supervised non-parallel voice conversion model.

to speaker-independent features extracted from text-transcription. A decoder then reconstructs the acoustic features from the linguistic features conditioned on the speaker label. In the case of semi-supervised model , an encoder learns the disentanglement without any text-transcription by applying regulation on the output linguistic features. The typical pipeline of semi-supervised model is shown in Fig. 2.5(b).

2.3 Evaluation metrics

The performance of the voice conversion system can be evaluated via both objective and subjective tests. The objective evaluation usually measures the spectral distortion, the spectral over-smoothness, and the pitch error between the converted speech and the target speech. While the subjective measures the perceived naturalness and speaker similarity of the converted speech via listening tests.

2.3.0.1 Objective evaluation metrics

The typical metrics to measure the spectral distortion are Log-Spectral distortion (LSD) and mel-cepstral distortion (MCD). Besides, one of the most common problem of converted speech is the over-smoothing degradation, a situation in which the converted speech sounds muffle and unnatural. This problem happens due to the reduction in the global variance or modulation spectrum of the acoustic features. Therefore, global variance and modulation spectrum metrics are often used to objectively measure the naturalness of converted speech.

Log-spectral distortion
The log-spectral distortion measures the difference between two sequence of log-spectra. The LSD between two spectral squence is calculated in the below equation:

$$d(\mathbf{s}, \hat{\mathbf{s}}) = 10\frac{1}{T}\sum_{T}^{t=1}\sqrt{\sum_{i=0}^{D}(\log_{10}(x_{i,t}) - \log_{10}(\hat{x}_{i,t}))^2} \tag{2.2}$$

where D is the total number of frequency bins, T is the total number of frames.

Mel-cespstral distortion
The mel-cepstral distortion (MCD) is the common metric for measuring the difference between two sequences of mel-cepstra. The mel-cepstral distortion between two sequence of mel-cepstrum is defined as:

$$d(\mathbf{c}, \hat{\mathbf{c}}) = \frac{10\sqrt{2}}{\ln 10}\frac{1}{T}\sum_{t}^{T}\sqrt{\sum_{i=0}^{D}(c_{i,t} - \hat{c}_{i,t})}, \tag{2.3}$$

where D is the total number of frequency bins, T is the total number of frames.

Global Variance The global variance describes the fluctuation of a parameter trajectory in utterance-level [52, 53]. For a given sequence of speech parameter vectors $Y = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T]$, the global variance on channel k of the parameter vector y_t is defined as:

$$v(k) = \frac{1}{T}\sum_{t=1}^{T}(y_t(k) - \bar{y}(k))^2, \bar{y}(k) = \frac{1}{T}\sum_{t=1}^{T} y_t(k) \tag{2.4}$$

, where $\bar{y}(k)$ is the intra-utterance mean value of the parameter at channel k.

Modulation Spectrum

The modulation spectrum is defined as the power spectrum of the parameter trajectory [54]. The modulation spectrum is closely related to the over-smoothing degradation of synthesized speech. In particular, the modulation spectrum of synthesized speech usually lower than those of natural speech, especially in the low modulation frequency region. The MS of parameter sequence x is defined as follows:

$$\begin{aligned} \mathbf{s}(\mathbf{X}) &= \left[\mathbf{s}(1)^\top, \cdots, \mathbf{s}(d)^\top, \cdots, \mathbf{s}(D)^\top\right], \\ s(d) &= \left[s_d(0), \cdots, s_d(f), \cdots, s_d(D_s)\right], \\ s_d(f) &= |DFT(\mathbf{x}(d)|, \end{aligned} \tag{2.5}$$

where D is the number of channel of $\mathbf{x}$, D_s is the number of frame of $\mathbf{X}$, $s_d(f)$ is the discrete Fourier transform of channel d at frequency bin f.

2.3.0.2 Subjective evaluation metrics

The subjective evaluations are conducted to measure the speech naturalness and speaker similarity of the converted speech via listening tests.

Naturalness test

The naturalness test measures how natural is the converted speech. The naturalness of the speech is defined as the speech output that sounds natural or normal to the listener as if it was uttered by human. The naturalness of an utterance is usually quantified by a standard 5-point-scale MOS score, which range from 1 (very unnatural) to 5 (very natural). The MOS of a system is represented as the average MOS score evaluated by each listener. However, due to the biased opinions across listeners, the MOS difference between voice conversion systems might be small and insignificant.

Alternately, the AB test is used to give better discrimination between voice conversion systems. At each time, the listener is presented with a pair of utterance from different voice conversion system. The generated utterance from both systems are presented in random order (A-B or B-A) to avoid any bias. Then the listener decides which utterance (A or B) has better speech naturalness. The higher selection rate for a given voice conversion system indicates a better naturalness performance.

Similarity test

The similarity test measure how close is the speaker individuality of generated speech to the target speaker individuality. The most common tests for speaker similarity are ABX test and the same/difference MOS test.

The ABX test is used to compare the similarity performance of two voice conversion system. The subjects are presented with 2 generated utterance (A and B) from each system, and 1 reference utterance (X) from either source speaker or target speaker. Then subjects decide which utterance, A or B, is closest to the reference utterance X in terms of speaker similarity. The voice conversion system which is given the higher selection rate when comparing with target speaker has the better similarity performance.

In the MOS test, the subjects are asked to listen to a pair of utterances, one is generated from the voice conversion system, and the other is the natural target speech. The subjects then judge the speaker similarity of the utterance pair in the 4-point scale: "Same, absolutely sure", "Same, not sure", "Different, not sure", and "Different, absolute sure". The similarity score of the voice conversion system is presented as the rate of each choice averaged across all subjects.

2.4 Chapter conclusion

This chapter has given an introduction to the basic principles of voice conversion technique and the evaluation metrics for voice conversion system. This was provided by first describes the speech production mechanism and the components of a typical voice conversion system. Then, the reviews of some common voice conversion technique were given with some comparisons. Next, several state-of-art methods for semi-supervised non-parallel voice conversion model were reviewed. Finally, the methods for evaluating the voice conversion systems were described.

Chapter 3

Semi-supervised non-parallel Voice Conversion with Variational Autoencoder

To achieve the primary goal of this thesis as described in Chapter 1, the first and most important step is to control speaker individuality in the non-parallel voice conversion. To this end, this chapter proposes a non-parallel intra-lingual voice conversion framework with speaker individuality control via a continuous speaker representation vector (i.e. speaker embedding). The continuous speaker embedding is learned by training the Variational Autoencoder model on a multi-speaker speech database. By analyzing the speaker embedding using principle analysis component (PCA), it is found that the voice characteristics of the speaker, such as gender, can be explored. Modifying the speaker embedding shows correspondence to the change of speaker individuality in the converted speech.

The rest of this chapter is organized as follow. In the first section, the introduction of this chapter is provided. Then in Section 3.2, the non-parallel voice conversion based on variational autoencoder is described. The proposed voice conversion model is introduced in Section 3.3. The experiment procedure and the results are provided in Section 3.4 before the conclusion in Section 3.5.

3.1 Introduction

Voice conversion (VC) is a special type of voice transformation (VT) whose aim is to manipulating speaker characteristics in the speech signal while preserving linguistic information [55]. This technique is beneficial in many practical applications such as intelligibility enhancement for speech disorder patients, or enhancing Human-Machine Interface experience. VC approaches can be categorized into 2 groups: rule-based approaches and statistical approaches.

Rule-based approaches [56–58] aim to modify acoustic features that correspond to the speaker individuality such as fundamental frequency (F_0) and formants by some manually derived rules. However, since different rules must be applied for different speakers, these approaches are impractical and less preferred than statistical approach.

On the other hand, statistical approaches use machine learning technique to modify the acoustic features. These approaches are more flexible to adapt to new speaker than rule-based method. The most straight-forward statistical approach for VC is to perform mapping from source acoustic features to target acoustic features. This approach requires a parallel training data, in which the source and target utterances contain identical linguistic information so that the differences in speaker voice characteristics could be learned. The conventional method for this approach is using Gaussian Mixture Model (GMM) to model the joint probability of source and target acoustic features [20]. However, synthesized speech using GMM-based method often suffered from over-smoothing degradation. Therefore, lately, Deep Neural Network (DNN) has been employed to perform the mapping task. With sufficient training data, DNN-based model outperforms GMM-based model in both speech naturalness and target similarity.

Despite the simplicity of mapping approach, parallel training data is often expensive to obtain. Therefore, a new set of method that can perform speaker adaptation using non-parallel data has been investigated. The first non-parallel VC method utilize an Eigen GMM-based model to describe speaker characteristic as combinations of base speakers [20]. However, although the speaker adaptation phase can work with non-parallel data, it requires parallel-data in the training phase. Later, various methods were proposed that can use non-parallel data in both training phase and adaptation phase. Some of the most popular methods are Restricted Boltzmann machine (RBM), Variational Autoencoder (VAE), and Generative Adversarial Network (GAN). All these three methods share the same principle of disentangling speaker-related information and linguistic information from speech waveform.

However, most prior non-parallel VC methods only focus on categorized speaker adaptation since a target voice is required as a reference to perform voice conversion. In other words, controllability of the degree of speaker individuality has not been much interested. These limitations restricted the use of VC system in some situations, such as in a storyteller system, when collecting utterances from a large number of target voices is unrealistic. In this situation, the VC system with the controllable voice characteristics is desirable as it can freely manipulate the source voice to generate any new fictitious voice without the recordings from the target speakers. Moreover, most VC model requires retraining when adapting to an unseen-target speaker. The controllability can also avoid this problem as the VC model can synthesize waveform with the desired voice characteristics extracted from the reference utterance. This controllability is also beneficial in many other voice transformation fields such as emotional voice conversion, voice dubbing in movie post-production, creating new voices for text-to-speech system, speech enhancement, and voice editing software.

To achieve this goal, this study proposes a VC framework based on VAE that can simultaneously disentangle speaker-related information with linguistic information and discover the latent structure of speaker characteristic. After training on a multi-speaker dataset, a continuous speaker embedding (SE) that represents voice characteristics is obtained. By manipulating the speaker embedding vector,the synthesized waveform with desired voice characteristics can be obtained. The experimental results show that the proposed VC system with continuous SE input (SE-VAE) has comparable performance as using VAE with one-hot speaker identity input (OH-VAE).

The significant of the proposed VC system are:

- Controlling the characteristics of converted voice using non-parallel training data.
- Performing speaker adaptation using a minimum of one utterance from target speaker.
- Converting waveform from both seen- and unseen-source speaker to unseen-target speaker and fictitious speaker.

3.2 Variational Autoencoder-based Voice Conversion

The VAE is a probabilistic model that can discover the latent structure of data [59]. In VC, a previous study by Hsu et al. [8] showed that linguistic information can be interpolated via latent representation of the VAE. The latent variable $\mathbf{z}$ is assumed to follow the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ that is independent from the speaker information. Hence, the latent variable $\mathbf{z}$ can be regarded as linguistic information conveyed in speech. From the input acoustic feature $\mathbf{x}$, the encoder of the VAE f_{enc} outputs the estimated parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ of the posterior $p_\theta(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$. Then $\mathbf{z}$ is sampled from the posterior as $\mathbf{z} \sim p(\mathbf{z}|\mathbf{x})$. However, back-propagation is impossible if $\mathbf{z}$ is directly sampled from the posterior $p_\theta(\mathbf{z}|\mathbf{x})$. Therefore, a re-parameterization trick is applied by sampling an independent variable $\boldsymbol{\epsilon}$ from normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ then executing a scale and shift operation.

The procedure of estimating $\mathbf{z}$ is as follows:

$$\begin{aligned} \boldsymbol{\mu},\ \boldsymbol{\sigma} &= f_{enc}(\mathbf{x}) \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathbf{z} &= \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}, \end{aligned} \tag{3.1}$$

where $\circ$ is the Hadamard product.

To reconstruct $\mathbf{x}$, in addition to the linguistic information in $\mathbf{z}$, a variable $\mathbf{s}$ that contains speaker information is introduced. The $\mathbf{s}$ can be expressed as a one-hot encoded vector or continuous vector that represents the speaker's identity. From $\mathbf{z}$ and $\mathbf{s}$, the decoder of the VAE reconstructs $\mathbf{x}$s as follows:

$$\widehat{\mathbf{x}} = f_{dec}(\mathbf{z}, \mathbf{s}). \tag{3.2}$$

The encoder and decoder are jointly trained by minimizing the variational objective function:

$$\mathcal{L}_v = -D_{KL}(p_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - \mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{x})}(p(\mathbf{x}|\mathbf{z}, \mathbf{s})), \tag{3.3}$$

where D_{KL} is the Kullback-Leibler divergence between the estimated posterior $p_\theta(\mathbf{z}|\mathbf{x})$ and the true prior distribution $p(\mathbf{z})$. Since $p(\mathbf{z})$ is assumed to follow a normal distribution, D_{KL} can be expressed in closed form as

$$D_{KL}(p_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = -\frac{1}{2}\sum(1 + \log\boldsymbol{\sigma}^2 - \boldsymbol{\mu}^2 + \boldsymbol{\sigma}^2). \tag{3.4}$$

The second term on the right side of (3.3) is the reconstruction loss. Assuming that $\mathbf{x}$ also follows a Gaussian distribution, the term $\mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{x})}(p(\mathbf{x}|\mathbf{z}, \mathbf{s}))$ can be described by a simple mean-squared difference between reconstructed acoustic features and original acoustic features as

$$\mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{x})}(p(\mathbf{x}|\mathbf{z}, \mathbf{s})) = -\frac{1}{2}\sum(\widehat{\mathbf{x}} - \mathbf{x})^2. \tag{3.5}$$

According to Rolinek et al. [60], the optimization of (3.3) will lead to a *polarized regime* situation, in which only a subset of the latent variables (active subset) encodes meaningful information, while the other subset (passive subset) purely encodes noise. Clearly, the passive subset has $D_{KL} \approx 0$. Therefore, the second term in (3.3) encourages a bottleneck in the latent variable, where useful information is restricted only in the active subset. Figure 3.1 illustrates the inferred latent statistical parameters from an input utterance. Since most of the dimensions are invariant with $\mathbf{x}$, the decoder is unable to fully reconstruct the $\mathbf{x}$s without any additional information. In this situation, the decoder network has to rely on the speaker information contained in the input speaker embedding to minimize the reconstruction loss (second term in (3.3)). This is the cause of the disentanglement of linguistic information and speaker information in the VAE.

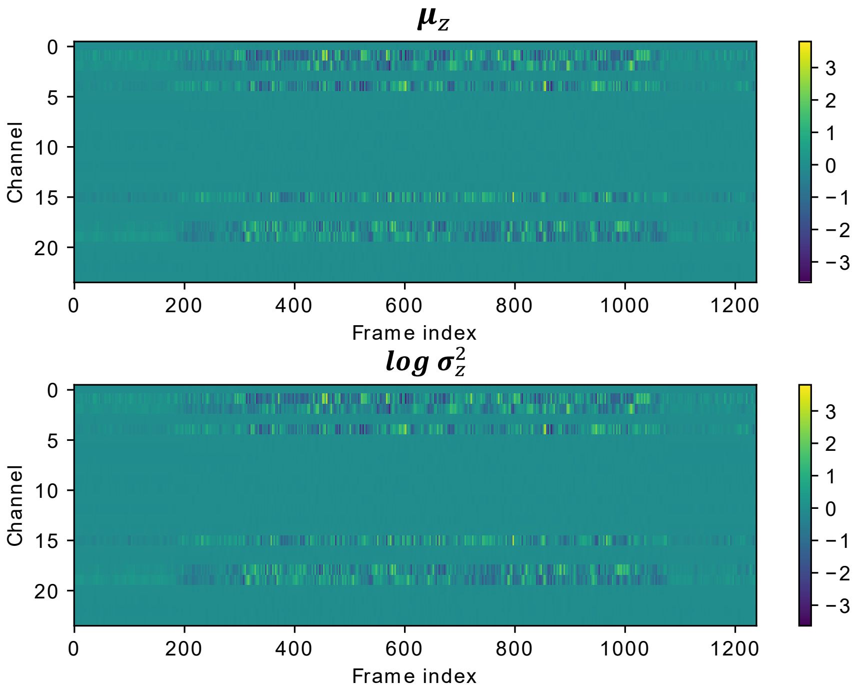


Figure 3.1: Generated parameters for posterior $q(\mathbf{z}|\mathbf{x})$. Most dimensions of the latent mean μ_z and log variance $\log \sigma^2$ are invariant with respect to input.

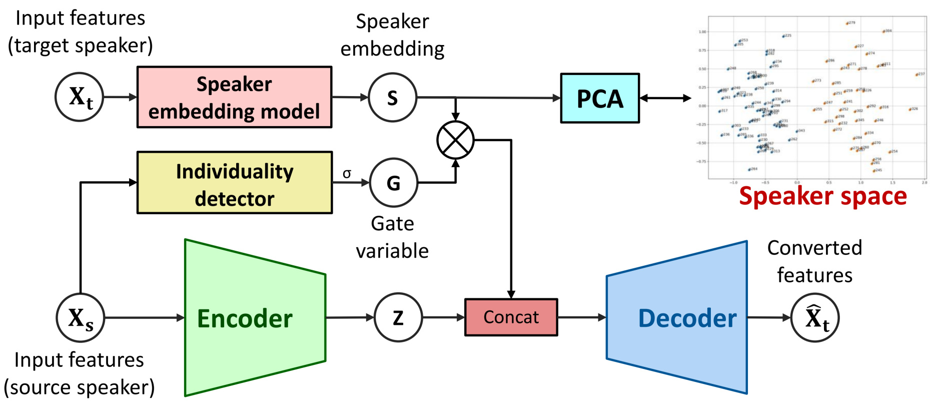


Figure 3.2: Overview of proposed VC system.

3.3 Proposed model

This section describes the proposed method for controlling speaker individuality and the modified training loss to account for the over-smoothing problem. The overview of the proposed voice conversion system is shown in Fig. 3.2.

3.3.1 Infer Speaker Embedding using Back-propagation

In conventional VAE-based VC, speaker identity is represented as a one-hot vector. However, this type of encoding does not include any other information on the speaker's voice characteristics such as gender or age. To overcome this problem, a different interpretation of speaker identity is used by letting the model self-derived the most suitable speaker embedding during the training process. Let $\mathbf{y}$ is the one-hot vector represent speaker identity, the continuous speaker embedding vector $\bar{y}$ is:

$$\bar{y} = \mathbf{W} \cdot y^{\intercal} + \mathbf{B}, \tag{3.6}$$

where $\mathbf{W}$ and $\mathbf{B}$ is a learnable kernel and bias in a fully-connected NN layer. In this interpretation, the one-hot encoded vector y acts as a switch to select correspond row vector in matrix $\mathbf{W}$. With this interpretation, 2 speakers with similar voice characteristics may have almost identical speaker embedding.

This interpretation can be expanded into by adding more layer and applying non-linear activation such as tanh or sigmoid. In this case, the speaker embedding $\bar{y}$ is

$$\bar{y} = \mathbf{W_n} \cdot ...f(\mathbf{W_1} \cdot f(\mathbf{W_0} \cdot y^{\intercal} + \mathbf{B_0}) + \mathbf{B_1})... + \mathbf{B_n}, \tag{3.7}$$

where f is a non-linear function. Although this interpretation is convenient to explain voice characteristics, however, the speaker embedding is only available for speakers in the training set. Therefore, to perform voice conversion on a new target speaker that is not in the training set, a speaker embedding model is trained to predict the corresponding speaker embedding from acoustic features. The learned speaker embeddings obtained after training VAE model are used as the ground truth. After the speaker embedding model is trained, a speaker embedding vector from new target speaker can be estimated using only a few seconds of their recording (10 seconds in the actual experiments).

3.3.2 Modulation Loss

To improve the naturalness of the synthesized speech, the Modulation Spectrum (MS) loss is adopted in the proposed model because of its beneficial effect on speech naturalness. Similar to [61], the MS of parameter sequence x is defined as follows:

$$\begin{aligned}\mathbf{s}(\mathbf{X}) &= \left[\mathbf{s}(1)^\top, \cdots, \mathbf{s}(d)^\top, \cdots, \mathbf{s}(D)^\top\right] \\ s(d) &= \left[s_d(0), \cdots, s_d(f), \cdots, s_d(D_s)\right] \\ s_d(f) &= |FFT(\mathbf{x}(d)|\end{aligned} \tag{3.8}$$

where D is the number of channel of $\mathbf{x}$, D_s is the number of frame of $\mathbf{X}$, $s_d(f)$ is the FFT of channel d at frequency bin f.

The modified log-likelihood function for the VAE model considering the modulation spectrum is defined as follow:

$$\begin{aligned}\overline{L}_{ms}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}_n) =& - D_{KL}(q_\phi(\overline{\mathbf{z}}_n|\mathbf{x}_n)||p(\mathbf{z}_n)) \\ &+ log\ p_\theta(\mathbf{x}_n|\overline{\mathbf{z}}_n, \overline{\mathbf{y}}_n) + w.log\ p(s(\mathbf{x})|\overline{\mathbf{z}}_n, \overline{\mathbf{y}}_n)\end{aligned} \tag{3.9}$$

The final term in Eq. 3.9 explicitly constrains the model to increase the log-likelihood of the modulation spectrum conditioned on the given latent variable $\overline{\mathbf{z}}_n$ and speaker identity y_n. Furthermore, the modulation spectrum is assumed to follow a Gaussian distribution with a diagonal covariance matrix: $s(x) \sim N(s(x)|s(\overline{x}), diag(\sigma_s))$. Therefore, the final log-probability term in Eq. 3.9 can be expressed in the following closed form:

$$\begin{gathered}log\ p(s(\mathbf{x})|\overline{\mathbf{z}}_n, \mathbf{A}^{(X)}) = \\ -\frac{1}{2}\sum\left(log(2\pi\sigma_s^2) + \frac{(s(\mathbf{x}) - s(\overline{\mathbf{x}}))^2}{\sigma_s^2}\right)\end{gathered} \tag{3.10}$$

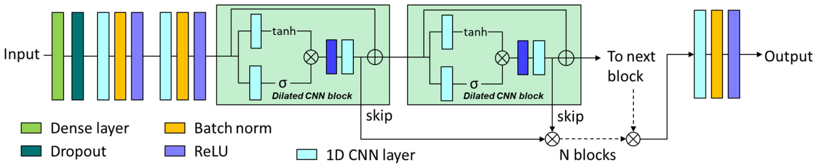


Figure 3.3: Multi-scale architecture with dilated residual CNN block

3.4 Experiment and discussion

3.4.1 Model configuration

Figure 3.2 illustrates an overview of the proposed SE-VAE VC model. The encoder and decoder network utilize the multi-scale convolutional Neural Network (CNN) architecture as shown in Fig. 3.3. In addition to the basic VAE framework, the auxiliary gate variable g is introduced to control the amount of the speaker individuality in the output features. The reason for this controlling is that some speech segments, such as silence, may not contain any speaker individuality. By introducing the gating variable, the model can ignore these segments by outputting the gate variable $g = 0$. The gating variable is inferred directly on the input features by the *individuality detector* block.

3.4.2 Dataset

The VCTK corpus [62] is used for training the models, which contains 44 hours of recordings from 109 English speakers. The data is divided into 2 subsets: training set (containing 100 speakers) and testing set. The testing set consisted of 2 groups of utterances. One group contains utterances from 9 held out speakers from the training set (unseen speakers). The second group contains 2 held out utterances of each speaker from the training set (seen speakers).

As speech features, the WORLD vocoder [63] is used to extract the fundamental frequency value F_0, the spectral sequence sp, and the aperiodicity from speech waveform. Then the spectral sequence sp is transformed to 60^{th}-order Mel-cepstral coefficients (MCC). Since the spectral envelope from WORLD vocoder is very smooth, high-order cepstral coefficients, which capture the fine fluctuation in the spectral envelope, can be neglected during the conversion process. Therefore, only the 1^{st} to 31^{th} MCC coefficients is used along with interpolated F_0 and voice/unvoiced flag as the input features.

For the proposed system, the VC model and speaker embedding model are trained separately. The VC model are firstly trained to obtained the speaker embedding table. Then a feed-forward neural network is trained to estimate the speaker embedding from the from the speech features. Both VC model and speaker embedding estimation model are trained on the same training set.

For the baseline models, the GMM-based VC (denoted as GMM) used in Voice Conversion Challenge 2018 and the VAE-based VC model that uses the fixed one-hot encoded speaker vector (OH-VAE) [9] is used. For the baseline onehot-VAE model, most of the model architecture is kept identical to the proposed model for a fair comparison. Since the baseline model cannot convert voice to unseen target speaker, the baseline model is only evaluated in seen source to seen target (*seen-seen*) and unseen source to seen target (*unseen-seen*) conversion scenarios. To perform voice conversion, the baseline VAE model uses the one-hot encoded vector, while the proposed model uses the speaker embedding extracted from a 10-second utterance of the target speaker. For each source-target speaker pair, a separate GMM-based VC model is trained using 100 pair of parallel utterances.

3.4.3 Speaker Embedding Space

After the VC model is trained,the speaker embedding space is visualized by analyzing the speaker embedding using Principle Component Analysis (PCA). As shown in Fig. 3.4, the speakers are well separated by genders, with all female speakers lie on the left and male speakers lie on the right. This indicates that the model can learn meaningful voice characteristics of the speakers.

3.4.4 Fictitious Speaker

To generate the fictitious voices that are not exist in the training data, the speaker embedding vectors are sampled on the speaker embedding space as shown in Fig. 3.5. To evaluate the naturalness of the fictitious voices, 9 utterances from a female speaker in VCTK dataset (seen speaker) are generated with the above sampled speaker embedding vectors.

3.4.5 Subjective Evaluations

To evaluate the quality of the converted waveform, two listening tests are conducted: the speech naturalness test and speaker similarity test. Eight listeners (6 males, 2 females) with normal hearing ability enrolled in these tests. All the listeners rate the same sets of test stimuli.

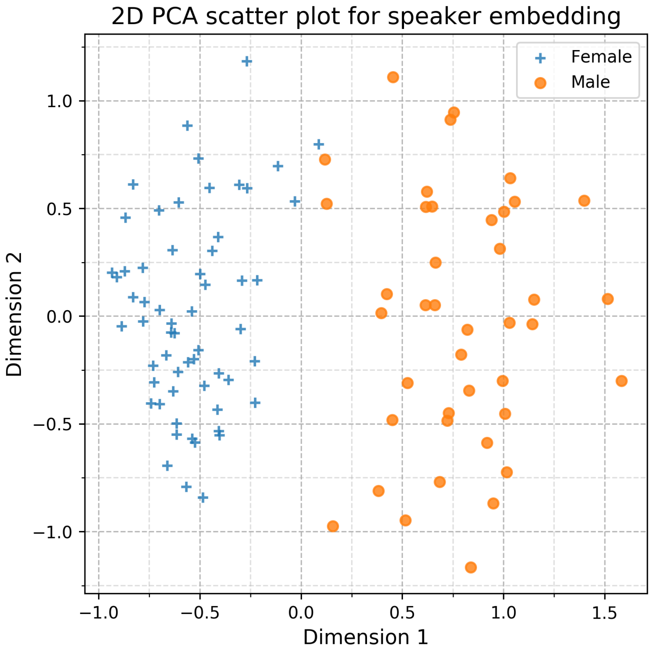


Figure 3.4: Learned speaker embedding map of VCTK dataset

3.4.5.1 Speech naturalness test

The naturalness of converted speech from the baseline models and the proposed model are measured using Mean-Opinion Score evaluation in 5 test scenarios: 1) seen-seen, 2) unseen-seen, 3) seen-unseen, 4) unseen-unseen and 5) fictitious target speaker. Two target speakers (1 male (M), 1 female (F)) were selected for each test scenarios except scenario 5. For each target speakers, all the models will perform both intra-gender (M-M, and F-F) and cross-gender (M-F, and F-M) conversion with the same source speakers. The listeners are instructed to concentrate on the quality of the speech and rate the sample using 5 point-scale that consisted of "bad" (1), "poor" (2), "fair" (3), "good" (4) and "excellent" (5). The order of test stimuli is randomized for each speaker. The result shown in TABLE 3.1 and Fig. 3.6 indicates that the speech waveform generated from the proposed model have higher naturalness than those generated from the GMM-based model in all conversion scenarios. When compared with the onehot-VAE model, the proposed model can synthesize waveform with equivalent naturalness, although only one utterance from the target speaker is required as the reference. The results in TABLE 3.1 marked with an asterisk are significantly

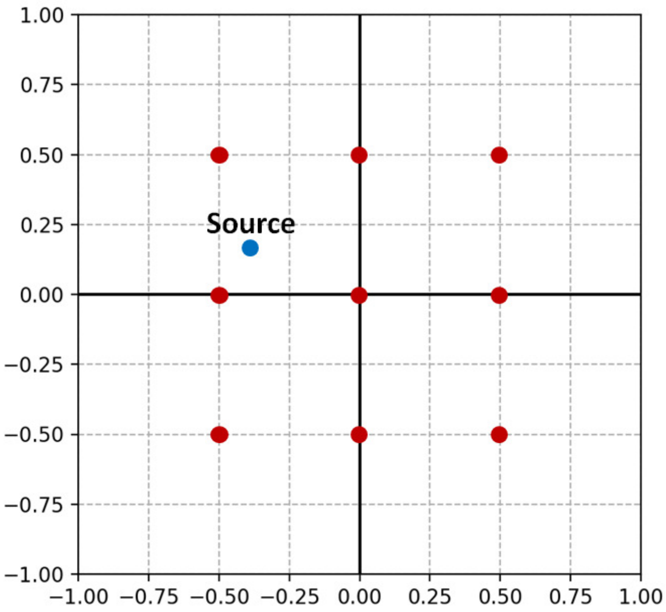


Figure 3.5: **Blue**: Position of source speaker embedding vector, **Red**: Position of selected target speaker embedding vector for synthesizing fictitious voices

different ($p < 0.05$) as compared to the proposed SE-VAE model. Moreover, the generated speech of fictitious speakers also has fair naturalness of 3.1 MOS.

3.4.5.2 Speaker similarity test

In this experiment, the speaker similarity between the converted waveform and the target waveform is evaluated in 4 test scenarios: 1) seen-seen, 2) unseen-seen, 3) seen-unseen and 4) unseen-unseen. The listeners are given a reference utterance from target speaker and several converted utterances from different source speakers. All the test stimuli are identical to the test stimuli in naturalness test. The listeners were instructed to concentrate on the voice characteristics and ignore any distortion or degradation in the test stimuli. Then the listener judges the voice similarity between the converted utterances with the reference utterance using the 5-point scale “not at all similar” (1), “slightly similar” (2), “moderately similar” (3), “very similar” (4) and “extremely similar” (5). The result of the similarity test is reported

Table 3.1: Result of MOS test for speech naturalness in intra-gender and cross-gender conversion

	GMM	**OH-VAE**	**SE-VAE** (*proposed*)
F-F	1.55±0.32*	3.06±0.67	**3.14±0.40**
M-M	1.66±0.43*	3.31±0.71	**3.30±0.74**
F-M	1.34±0.24*	2.25±0.35	**2.23±0.42**
M-F	1.48±0.28*	2.88±0.45	**2.72±0.57**
All	1.51±0.16*	2.88±0.32	**2.85±0.32**

Table 3.2: Result of MOS test for speaker similarity in intra-gender and cross-gender conversion

	GMM	**OH-VAE**	**SE-VAE** (*proposed*)
F-F	2.07±0.53	2.18±0.62	**2.39±0.65**
M-M	2.11±0.62	2.89±0.98	**2.54±0.72**
F-M	1.77±0.44	2.18±0.69	**1.82±0.72**
M-F	1.52±0.27*	2.93±0.82	**2.57±0.62**
All	1.87±0.25*	2.54±0.41	**2.33±0.35**

in TABLE 3.2 and Fig. 3.7, with the asterisk indicates that the different is statistically significant ($p < 0.05$) as compared to the proposed SE-VAE model. From the result, it is clear that there is no difference between the proposed SE-VAE model and OH-VAE, despite the lacks of a large number of training examples from target speaker in the proposed scheme. The overall performance of the proposed VC system is significantly better than the GMM-based VC.

3.5 Chapter Conclusion

In this chapter, a non-parallel voice conversion with speaker individuality control has been proposed. The proposed method provides a flexible way to control speaker individuality of converted speech by modifying speaker embedding vectors. Even only a single utterance is required as the reference, the subjective test results show that the proposed model can convert speech with better perceived naturalness and speaker similarity to the baseline

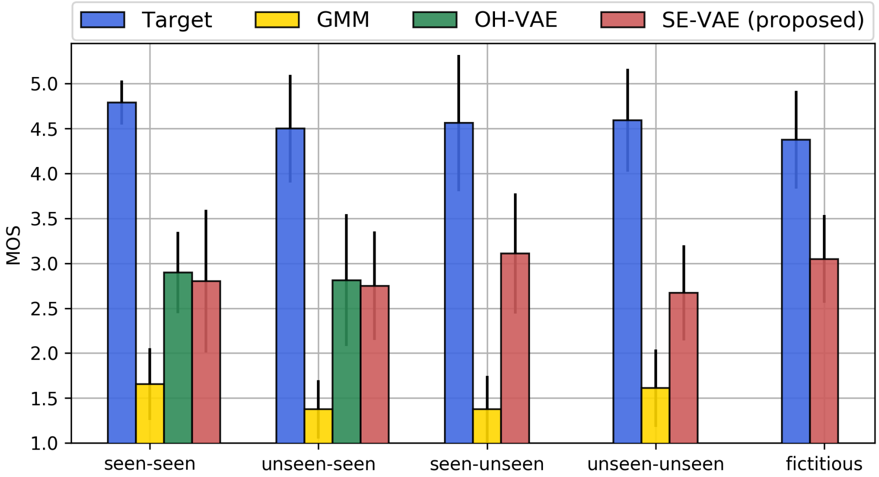


Figure 3.6: Result of MOS test for speech naturalness when adapting from seen/unseen-source to seen/unseen-target speaker with 95% confidence interval

GMM-based model and comparable to the VAE model using one-hot speaker identity vector. Moreover, since the proposed model can synthesize arbitrary voices with good naturalness, it is beneficial in various practical application to generate new voice individuality. Based on these results, it has been confirmed that the controllability of speaker individuality for non-paralel voice conversion, which is the first task of this thesis, has been achieved. The next chapter will focus on applied this framework to cross-lingual voice conversion.

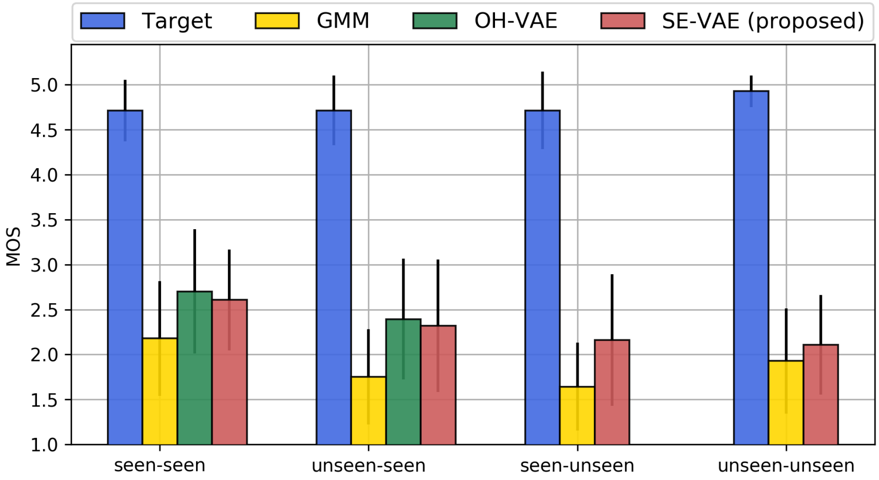


Figure 3.7: Result of MOS test for speaker similarity when adapting from seen/unseen-source to seen/unseen-target speaker with 95% confidence interval

Chapter 4

Cross-lingual Voice Conversion with speaker individuality control

In the previous chapter, a non-parallel voice conversion with speaker individuality control has been proposed. In the second task of this thesis, it is necessary to applied the speaker individuality control for cross-lingual voice conversion tasks. To this end, this chapter focuses on a non-parallel voice conversion that can operate in both intra-lingual and cross-lingual task with the limited training data for the target speaker. The proposed model share the same framework as the model in Chapter 3, with the difference in the regularization of the latent representation. To avoid the mode collapse problem in variational autoencoder, the vector-quantized variational autoencoder is applied with the hierarchical latent structure. To adapt to the cross-lingual settings, the proposed voice conversion model is fine-tuned on the cross-lingual target data after training on the intra-lingual data. The cross-lingual speaker embedding of the target speaker obtained after the fine-tuning process is used to condition the mel-spectrogram generation. The experimental results show that the proposed model delivers good performance in the speech naturalness in both intra-lingual and cross-lingual task. However, the speaker similarity in cross-lingual task remains to be improved.

The rest of the chapter is organized as follow. In the first section, the introduction of this chapter is provided. In Section 4.2, the non-parallel voice conversion based on vector-quantized variational autoencoder is described. Then, the proposed cross-lingual voice conversion model is introduced in Section 4.3. The experiment procedure and the results are provided in Section 4.4 before the chapter conclusion in Section 4.5.

4.1 Introduction

Voice conversion (VC) is a subset of voice transformation method for altering speaker characteristics while preserving the linguistic information [55]. Conventionally, VC can be seen as a mapping problem between source waveform and target waveform [20]. This perspective requires learning a mapping function using parallel training data, in which the source and target waveform shares the same linguistic information. However, parallel training data cannot be collected in some situations such as in cross-lingual VC. Therefore, VC methods for non-parallel data are increasingly gaining more attention in recent years.

One of the straightforward methods for non-parallel VC (NPVC) is to concatenate speech recognition (ASR) with text-to-speech (TTS) model [64–66]. These methods often achieve the highest performance with highly natural converted speech [67]. However, both the ASR and TTS models must be trained on an enormous amount of transcribed speech data, which is often very expensive to construct. This constraint limits the applicability of the ASR-TTS approach in a practical situation.

In contrast, NPVC based on deep generative model such as Generative Adversarial Network (GAN) and Variational Autoencoder (VAE) can be trained without transcribed data. Therefore, this type of NPVC model can be easily constructed from scratch using vastly available of untranscribed speech, thus reducing the development cost. With the recent advances of deep generative model, state-of-art GAN based VC [8, 22, 68] and VAE based VC [69, 70] have narrowed down the performance gap with ASR-TTS approaches. Although GANs come with a nice theoretical justification that the generated data should match the distribution of true data, it is widely known that the adversarial training is fragile and unstable. Moreover, while there are many studies on GAN-based VC, neither of them give strong evidence that the data distribution learned by Discriminator corresponds to human speech perception. In contrast, VAE can be easily trained. However, the VAE often suffers from the posterior collapse problem caused by Kullback-Leibler divergence (KLD) [71], which reduces the useful information received by the decoder for speech reconstruction.

A recently proposed Vector Quantized VAE (VQVAE) [72] model with discrete latent space avoids the posterior collapse problem by not optimizing the KLD but learning the categorical prior instead. Since linguistic information can be regarded as categorical data, discrete latent space is suitable to represent linguistic information. The VQVAE has been successfully applied in various speech processing tasks [73–75]. However, the linguistic information conveys different levels of semantic structure that spans at different temporal scales (e.g phonemes, syllables). Therefore, a single vector quantizer operating at a fixed temporal scale is inefficient to capture various levels of semantic structure, hence reducing the naturalness of converted speech. To tackle this problem,the hierarchical latent embedding VQVAE (HLE-VQVAE) is proposed to capture the linguistic information at various temporal scales. As shown in the next sections, the proposed scheme can improve the performance of VC system and provide highly natural converted speech for both intra-lingual and cross-lingual tasks.

4.2 Vector-quantized Variational Autoencoder

The VQVAE can be regarded as a communication systems, in which the input feature vector $\mathbf{x}$ is compacted into latent vector $\mathbf{z}$ by a non-linear transformation (encoder). The latent vector $\mathbf{z}$ is then quantized to discrete variable $\mathbf{q}$ based on its distance to pseudo-vectors in the codebook $\mathbf{e}_k, k \in 1...K$.

$$\mathbf{q} = \mathbf{e}_k \text{ where } k = \underset{k}{argmin} \left\| \mathbf{z} - \mathbf{e}_j \right\| \tag{4.1}$$

Finally, the decoder reconstructs the input vector from the discrete latent vector $\mathbf{q}$ and one-hot speaker embedding $\mathbf{s}_m$ of the source speaker. The latent codebook is updated simultaneously with other parameters of the model during training process. Due to the use of *argmin* function in quantization process, the computation graph is disconnected and the model cannot be trained with back-propagation. Therefore, straight-through reparameterization trick [72] is used to avoid this problem:

$$\begin{aligned} \mathbf{z} &= Enc(\mathbf{x}) \\ \mathbf{q} &= Quantize(\mathbf{z}) \\ \mathbf{q}_{st} &= \mathbf{z} + sg(\mathbf{q} - \mathbf{z}) \\ \mathbf{x}_{dec} &= Dec(\mathbf{q}_{st}, \mathbf{s}_m) \end{aligned} \tag{4.2}$$

where $\mathbf{x}_{dec}$ is the reconstructed feature vector, $\mathbf{q}_{st}$ is straight-through variable from which gradient is copied to $\mathbf{z}$, $Enc(\cdot)$ is the encoder function, $Dec(\cdot)$ is the decoder function, $Quantize(\cdot)$ is quantization function, and $sg(\cdot)$ is the stop-gradient operator. The model parameters are obtained by minimizing the following objective function:

$$\begin{aligned} \mathcal{L}_{\text{VQVAE}} = \left\| \mathbf{x} - \mathbf{x}_{dec}) \right\|_2^2 + \left\| \mathbf{z} - sg(\mathbf{q}) \right\|_2^2 + \\ \beta \left\| sg(\mathbf{z}) - \mathbf{q} \right\|_2^2 \end{aligned} \tag{4.3}$$

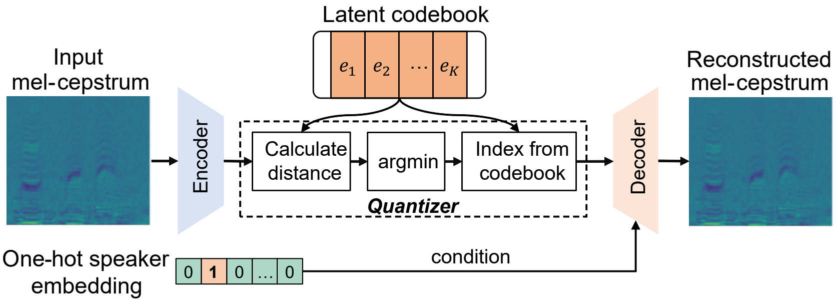


Figure 4.1: Conventional VQVAE-based VC.

where $\|\mathbf{x} - \mathbf{x}_{dec})\|_2^2$ is the reconstruction loss, $\|\mathbf{z} - sg(\mathbf{q})\|_2^2$ is the quantization loss, $\|sg(\mathbf{z}) - \mathbf{q}\|_2^2$ is the codebook loss, and β is a hyper-parameter to control the reluctance to change of the codebook loss.

At the inference step, providing the source mel-cepstrum and the speaker embedding of target speaker, the model outputs the converted mel-cepstrum containing the target voice characteristics. The overview of conventional VQVAE based VC is shown in Fig. 4.1.

4.3 Proposed method

In this section, the VQVAE model with hierarchical latent embedding structure (HLE-VQVAE) is proposed. Following this, the method to adapt the intra-lingual VC model for cross-lingual VC task is described.

4.3.1 Hierarchical Latent Embedding VQVAE

In conventional VQVAE, input data are encoded to latent embedding variable at a fixed temporal scale. However, the semantic structure of speech contains different levels that span across different temporal scale. Inspired by the work of [76] on image generation, a hierarchical structure is used to better capture different information at different temporal scales.

The overview of the proposed model with 3 stages of hierarchical structure is shown in Fig. 4.2. Each stage consists of an encoder network, a quantizer and a decoder network. At stage n, the encoder downsamples its input and the decoder upsamples its input by the same factor. Except for the top encoder, each encoder output is split along channel dimension into 2 parts: the latent variable $\mathbf{z}_n$ and hidden variable $\mathbf{u}_n$. The latent variable $\mathbf{z}_n$ is then discretized to $\mathbf{q}_n$, while hidden variable $\mathbf{u}_n$ is passed to the next encoder. On the decoder side, the discrete latent variable of the current stage is concatenated with the decoded hidden variable $\mathbf{v}_n$ from previous stage before passing through the decoder network. Similar to vanilla VQVAE

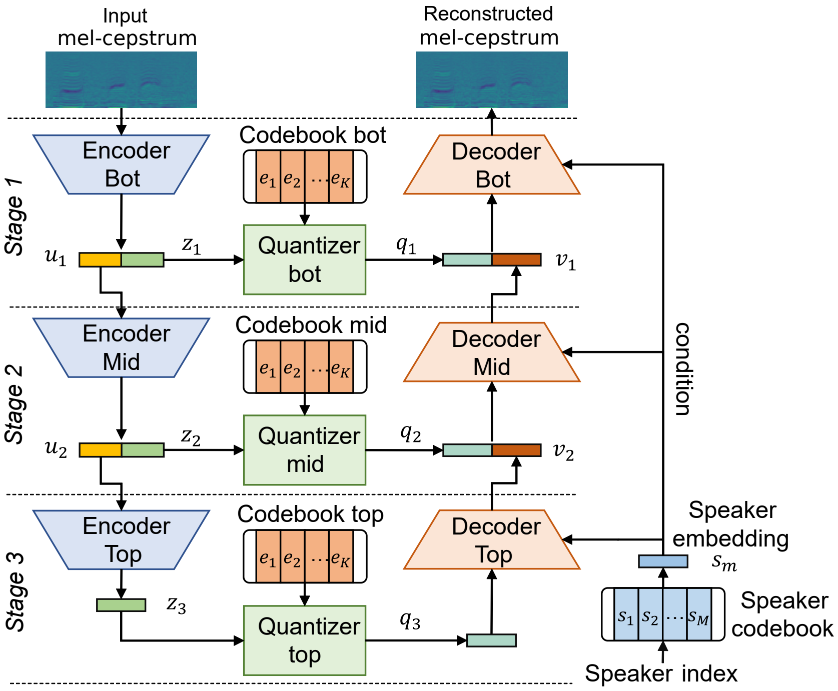


Figure 4.2: Diagram of the proposed 3-stage HLE-VQVAE.

based VC, each decoder in the proposed model is conditioned by the same speaker embedding $\mathbf{s}_m$.

At the training phase, providing the mel-cepstral sequence with speaker embedding of source speaker, the model is trained to minimize the following objective functions:

$$\mathcal{L}_{\text{HLE-VQVAE}} = \|\mathbf{x} - \mathbf{x}_{dec})\|_2^2 + \sum_{n=1}^{N} \left(\|\mathbf{z}_n - sg(\mathbf{q}_n)\|_2^2 + \beta \, \|sg(\mathbf{z}_n) - \mathbf{q}_n\|_2^2 \right) \quad (4.4)$$

where N is the number of hierarchical stage, β is set to 0.25 in this study.

4.3.2 Learnable speaker embedding

One-hot speaker embedding has the drawback that the number of speaker embedding is fixed by the dimension of the one-hot vector. Follow the study [25], the proposed model uses learnable speaker embeddings which are jointly optimized with other models parameters during the training phase by using back-propagation. The speaker index is used to select the corresponding speaker embedding in speaker codebook.

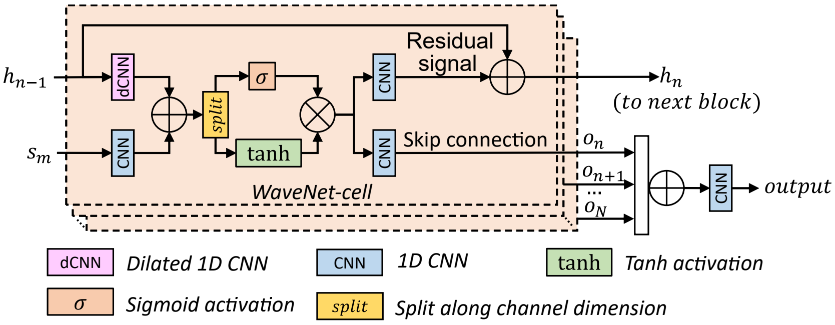


Figure 4.3: Stacks of non-causal dilated Wavenet-like structure in encoder and decoder.

4.3.3 Cross-lingual adaptation

The advantage of the proposed VC scheme is that only the target speaker embedding is needed to mimic the voice characteristics of the target. To obtain the target speaker embedding of foreign language, the latent codebook from the pretrained intra-lingual model and the random-initialized speaker embedding are fine-tuned on the target data. After the target speaker embedding is obtained, the model generates converted mel-cepstrum using the similar inference step described in Section 4.2.

4.4 Experiments

In this section, the results of the objective and subjective measurements are described to explain the model selection for Voice Conversion Challenge 2020 (VCC2020). Then the official results of the VCC2020 are shown to demonstrate the performance of the submitted system. To conveniently compared the models that are tested, the models are named as follows:

- **VQVAE**: conventional VQVAE model with 1 stage of quantization.
- **HLE-VQVAE-2**: the proposed HLE-VQVAE model with 2 stage of quantization.
- **HLE-VQVAE-3**: the proposed HLE-VQVAE model with 3 stage of quantization.

4.4.1 Dataset

The VCC2020 training set consists of 4 source English speakers, 4 target English speakers, and 2 target speakers of each foreign language (Finnish, German, and Mandarin). Each speaker in the VCC2020 training set utters a sentence set consisting of 70 sentences. Besides, a subset of the CSTR

VCTK dataset [62] containing all utterances from the first 100 speakers was used in combination with the VCC2020 training set to train the models. The VCC2020 evaluation data is used for testing.

In the pre-processing step, the audio file is down-sampled to 24 kHz and normalized to $[-1.0, 1.0]$ range. Then, an 80-dimension mel-spectrogram is extracted using the Short-time Fourier Transform (STFT) and mel-filterbank. The window length of STFT is set to 2048 and the hop-length is 300. The mel-spectrum is transformed into mel-cepstrum by applying Inverse Discrete Fourier Transform on the log-magnitude mel-spectrum. The Parallel WaveGAN neural vocoder [77], which has been trained on the VCTK dataset for 1000k iterations, is used to reconstruct the waveform,

4.4.2 Implementation details

For the proposed model, the downsampling and upsampling factors for each encoder and decoder are set to 2. The codebook at each stage contains 128 atoms of 32 dimensions. The encoder and decoder are implemented by stacking multiple non-causal dilated WaveNet-like structures [40] as shown in Fig. 4.3.

For the baseline model, a conventional VQVAE model is implemented with a similar encoder and decoder structure as the proposed model. As the baseline model has 1 stage, the feature vector is downsampled by the factor of 2 before quantized using a codebook containing 256 atoms of 64 dimensions.

The dimension of speaker embedding in all models is 16. The model parameters were optimized using Adam [78] with learning rate of 0.0005 and gradually reduced to 0.0002 after 10 epochs. For intra-lingual task, all models were trained with 200 epochs with batch size 32. For cross-lingual adaptation, all models are fine-tuned with 1000 epochs for each target speaker.

4.4.3 Objective test

The modulation spectrum (MS) of the parameter temporal trajectory is one of the well-known metrics to measure the quality of synthetic speech [79]. The MS of converted mel-cepstrum is measured by taking Discrete Fourier transformation on each cepstral sequence. Then, root-mean-squared errors (RMSEs) between the logarithmic MS of target natural speech and converted speech from different models are calculated. It should be expected that the lower the RMSEs, the better quality of converted speech. The RMSEs are measured on all the converted utterances and averaged across all mel channels and modulation frequencies. The results shown in Table 4.1 indicate that the mel-spectral sequences obtained from the proposed models are closest

Table 4.1: Comparison of RMSE between target and converted logarithmic MS averaged over all mel channels and modulation frequencies. Smallest RMSE value is highlighted in bold.

Method		VQVAE	HLE-VQVAE-2	HLE-VQVAE-3
Intra-lingual	Same-gender	0.267	0.258	**0.238**
	Cross-gender	0.313	0.302	**0.280**
Cross-lingual	Same-gender	0.431	0.427	**0.422**
	Cross-gender	0.434	**0.414**	0.430
Average		0.375	0.364	**0.359**

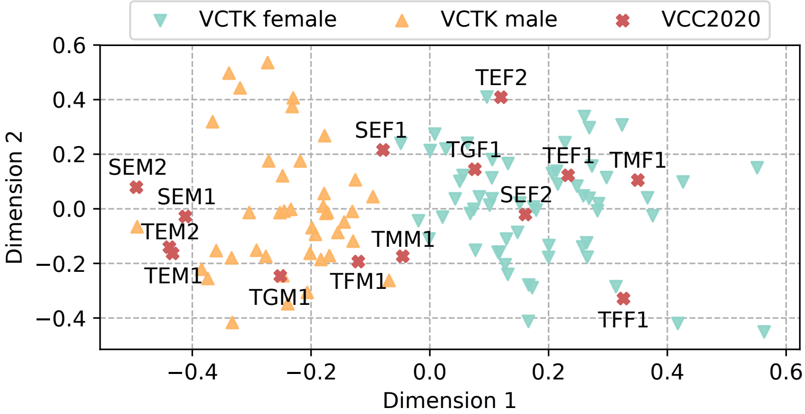


Figure 4.4: 2D PCA visualization of learned speaker embedding by HLE-VQVAE-3 model from VCC2020 dataset (**VCC2020**) and VCTK dataset (**VCTK male** and **VCTK female**). The horizontal and vertical axes are the first and second principal components, respectively.

to the target speaker in terms of MS. In particular, the HLE-VQVAE-3 outperformed the HLE-VQVAE-2 in most cases except for cross-lingual and cross-gender VC.

4.4.4 Visualization of Speaker Embedding

Principle component analysis (PCA) is used to visualize the learned speaker embedding. As shown in Fig. 4.4, it can be seen that the speakers are well clusterized by genders. This indicates that the speaker embedding can encode meaningful voice characteristics of the speakers without any additional speaker information.

4.4.5 Subjective test

The AB naturalness test and ABX similarity test are conducted to compare the performance of 3 models. Due to time constraint, the subjective tests only consist the converted speech between 2 source speakers (**SEF1** and **SEM1**) and 4 target speakers (English speakers: **TEF1** and **TEM1**, German speakers: **TGF1** and **TGM1**). Two sentences (E30001 and E30002) were selected from each source-target pairs to form the listening test set. Therefore, the listening test set consisted of 48 converted utterance pairs (2 sentences × 8 source-target speaker pairs × 3 model pairs). As for reference stimuli in the ABX similarity test, the original utterances of the target speakers are selected from the VCC2020 training set. There were 12 participants with good English proficiency joined both listening tests. Each participant rated 24 random pairs of converted utterances for each test.

The results of the AB naturalness test are shown in Fig. 4.5. It can be seen that the HLE-VQVAE-3 model outperformed the VQVAE and HLE-VQVAE-2 in terms of naturalness performance for both intra-lingual and cross-lingual conversion. The result of the ABX similarity test shown in Fig. 4.6 indicates that the HLE-VQVAE-3 model was slightly better than the HLE-VQVAE-2 model in cross-lingual VC. In other cases, the HLE-VQVAE-3 significantly outperformed the HLE-VQVAE-2 and VQVAE model. These results were also aligned with the objective measurement shown in Section 4.4.3.

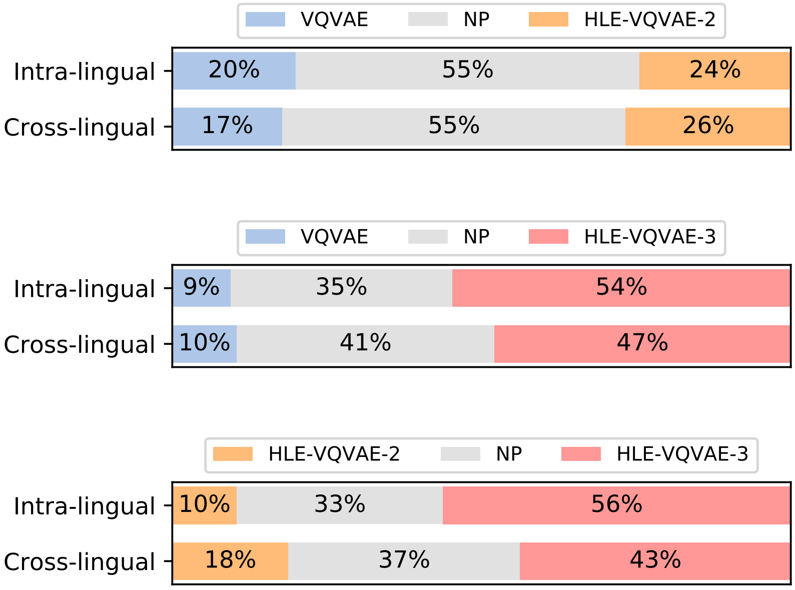


Figure 4.5: Preference score of AB naturalness test. **NP** means no preference.

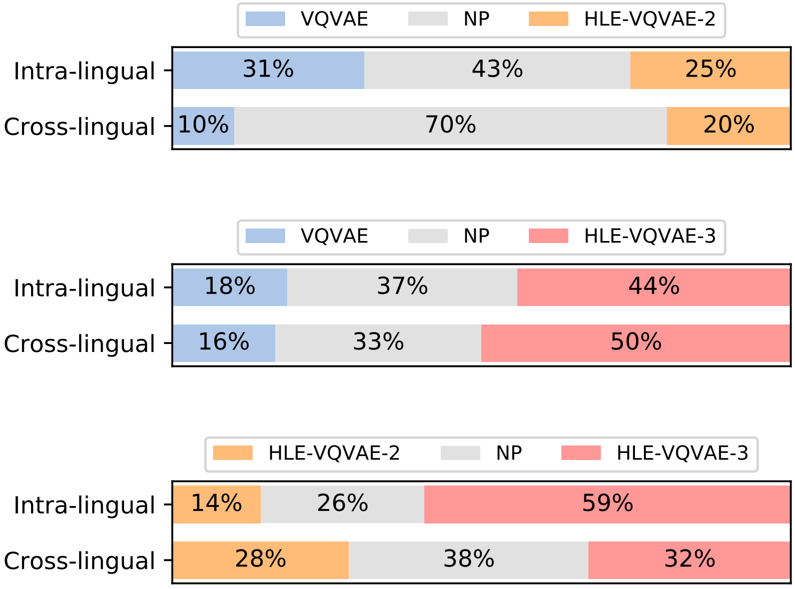


Figure 4.6: Preference score of ABX speaker similarity test. **NP** means no preference.

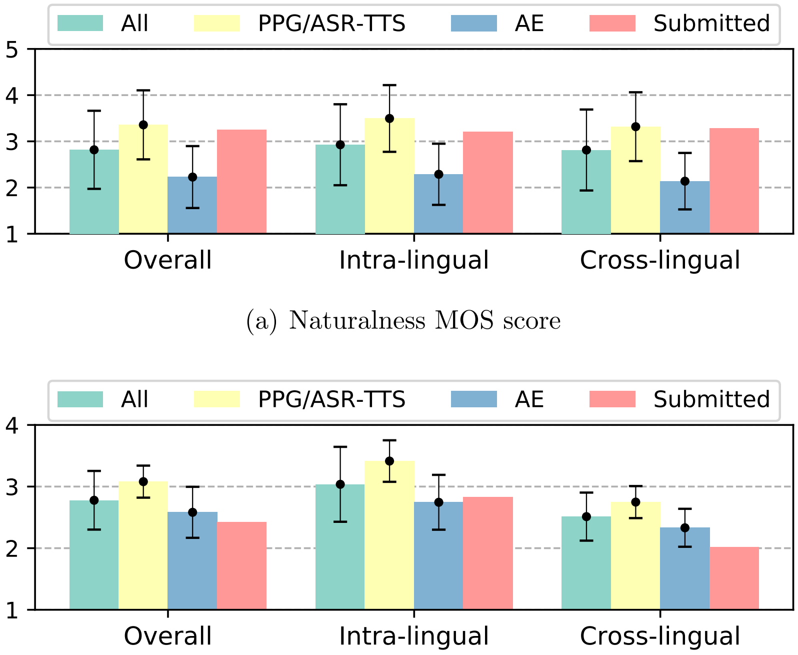
4.4.6 Voice Conversion Challenge 2020 results

The VCC2020 organizers conducted 2 large-scale listening tests to evaluate the speech naturalness and speaker similarity of converted speech [15]. In the naturalness test, listeners were asked to evaluate voice quality on a scale from 1 (Bad) to 5 (Excellent). In the speaker similarity test, listeners were asked to judge whether or not the converted and target utterances were spoken by the same person, and then evaluate using a 4-point scale that varies from "Different (sure)" to "Same (sure)".

To conveniently evaluate the performance of the submitted systems, the score of the proposed system is compared against different types of VC models, which is named as follows:

- **PPG/ASR-TTS**: text-dependent models including Phonetic Posteriorgram VC [64], concatenation of speech recognition (ASR) and text-to-speech (TTS) system , and leveraging TTS for VC methods [80]. Speech transcription is required to train these types of model.
- **AE**: Autoencoder based models including VQVAE, CycleVAE [81], AutoVC [51], and one-shot VC [82]. These types of VC models share the same paradigm as the proposed model in this study.

The results of naturalness MOS score and similarity score are summarized



(a) Naturalness MOS score

(b) Similarity MOS score

Figure 4.7: Average MOS score (top) and similarity score (bottom) with standard deviation of English listeners from all models (**All**), text-dependent models (**PPG/ASR-TTS**), autoencoder based models (**AE**), and the proposed model in this study (**Submitted**).

in Fig. 4.7. In both intra-lingual VC and cross-lingual VC tasks, the naturalness performance of the proposed model is significantly higher than the average of autoencoder based models and is comparable with the average of PPG/ASR-TTS based models. In terms of similarity performance, the proposed model still achieves a higher score than the average of autoencoder based VC in intra-lingual VC task. However, there is a decline in similarity score of the proposed model in cross-lingual VC task. This might be due to the lack of an explicit input F_0 information in the proposed VC model. Since the mean and the variance of F_0 is one of the important cues for speaker individuality [58], the speaker embedding may encode the F_0 statistics embedded in the mel-cepstrum. However, since different languages may have distinctive shape of F_0 contour which is reflected in F_0 statistics, the estimation of speaker embedding of foreign speaker will be biased. By providing the decoder with an explicit F_0 information, the speaker embedding will be freed from capturing F_0 mean and variance, hence increasing the accuracy of modeling speaker characteristics.

4.4.7 Voice quality test

A listening test is conducted to explore the physical meanings of speaker embedding regarding to the perceived speaker individuality. To describe the voice quality, 16 epithets (adjectives) derived from the study of *Hiroshi et al.* [83] are used. The epithets are divided into 8 pairs as listed in Table 4.2.

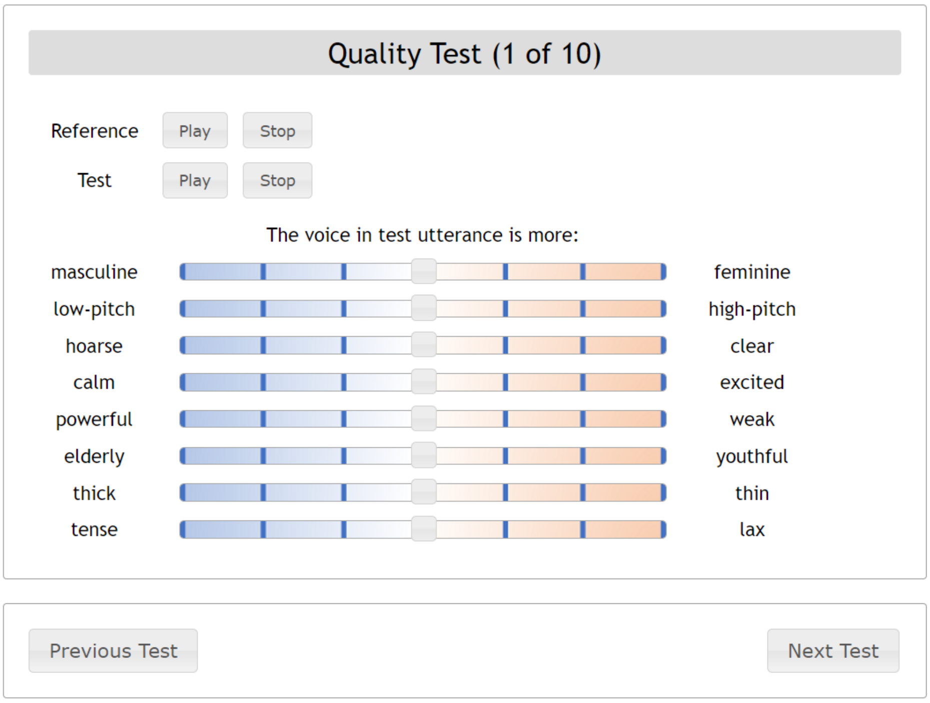
Table 4.2: Epithet pairs for describing speaker individuality from the study of *Hiroshi et al.*

Masculine	Feminine
High-pitched	Low-pitched
Hoarse	Clear
Calm	Excited
Powerful	Weak
Youthful	Elderly
Thick	Thin
Tense	Lax

In this test, the offset in voice quality between the reference voice and test voice is evaluated using epithet pairs. The reference and test utterances are generated from the voice conversion system by changing specific channel of PCA-projected speaker embedding. The amount of change is determined by the minimum and maximum value of a specific channel across speakers in training set. In particular, the reference utterance corresponds to lowest value of speaker embedding in a specific channel, while the test utterance is in opposite to the test utterance. Two Japanese speaker from the JVS database, one male and one female, are selected as the source speakers. The first 5 channels of the PCA-projected speaker embedding are modified to generate the reference and test utterances. Therefore, there are in total of 10 pairs of reference and test utterances in this test. Each epithet pair is evaluated using 7-point-scale score ranging from $[-3, 3]$. The test is conducted online[1] via website interface as illustrated in Fig. 4.8. There are a total of 7 participants joined the test. All participants evaluated the same 10 pairs of reference and test utterances.

The results of the listening test are shown in Fig. 4.9. Positive score indicates the positive correlation between the adjective and the speaker embedding channel. Analyzing the score that average across source speakers, the following points can be observed:

[1] http://www.jaist.ac.jp/~s1820029/quality_test/quality_test.html



Quality Test (1 of 10)

The voice in test utterance is more:

masculine	feminine
low-pitch	high-pitch
hoarse	clear
calm	excited
powerful	weak
elderly	youthful
thick	thin
tense	lax

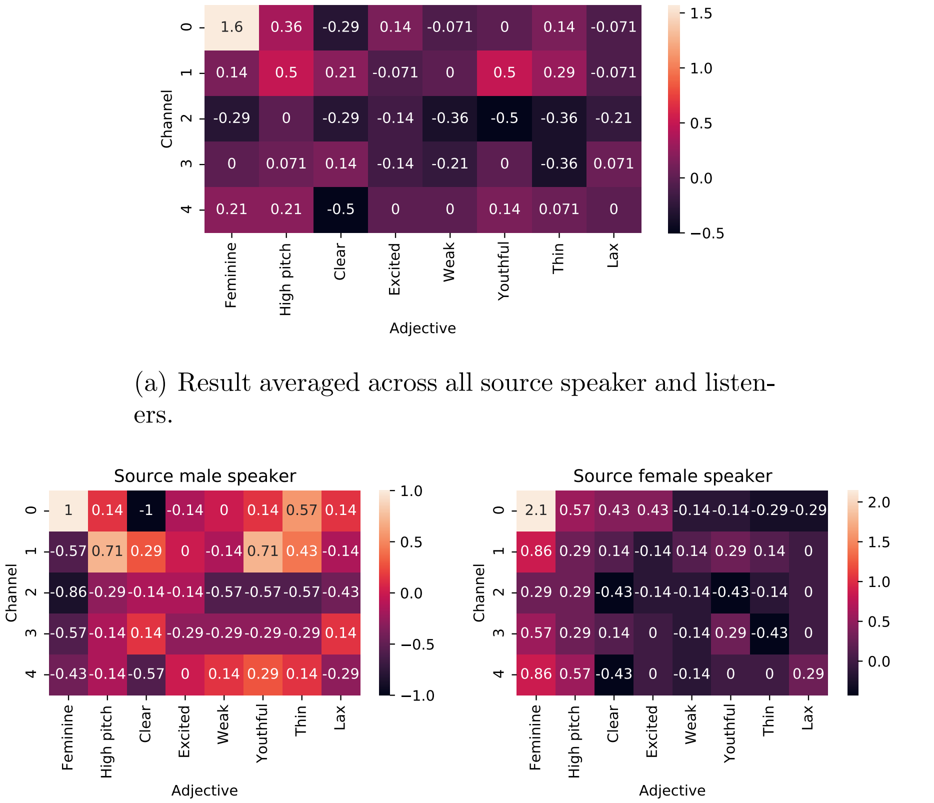
Figure 4.8: Interface of voice quality test

- The first channel mostly corresponds to the perceived gender of the voice. In particular, increase the first channel shift the voice quality toward feminine voice.
- Increase the second channel create more youthful/higher pitch voice.
- Increase the third channel increase the elderly voice quality, but not changing the pitch of the voice.
- Reduce the fifth channel create more clearer voice
- The forth channel shows no clear correspondence to the adjectives.

Moreover, comparing the result from source female and source male speaker, the following points can be derived:

- The first channel has stronger effect on female speaker.
- The third channel increase the perceived gender of the voice: increase the masculinity on male speaker and increase the femininity on female speaker.

In summary, it can be seen that some adjectives might be correlated to each other, i.e, feminine and high-pitched voice, youthful and clear voice, etc. Therefore, it is difficult to have a one-by-one relation between the channel of speaker embedding and the adjectives. In addition, the results for female and



(a) Result averaged across all source speaker and listeners.

(b) Results averaged across all listeners, broken down into source male speaker (left) and source female speaker (right).

Figure 4.9: Result of voice quality test. Positive value indicates positive proportion between the adjective and speaker embedding.

male source speakers are slightly different, which indicate that one adjective might correspond to different channels of speaker embedding. However, further investigation must be carried out to clarify the exact relationship.

4.5 Chapter Conclusion

This chapter has proposed a cross-lingual VC model with speaker individuality control based on VQVAE with a hierarchical latent structure. The experiment results show that the proposed model outperformed the vanilla VQVAE based VC model in both objective and subjective evaluation. Results from the official listening test in VCC2020 shown that the proposed HLE-VQVAE-3 model was comparable with the average performance of PPG/ASR-TTS models and superior to other autoencoder VC models in term of naturalness. The result of voice quality test reveal that the first channel of PCA-projected speaker embedding can control the perceived gender of converted speech. However, it is still difficult to obtain the one-by-one relation between the channel of PCA-projected speaker embedding and other voice adjectives. Based on these results, the sub-task of applying the speaker individuality control in cross-lingual task has been achieved. However, there are still rooms to improve the similarity performance of the proposed model. Therefore, in the next chapter, methods for improving the performance of cross-lingual voice conversion will be described.

Chapter 5

Method for improving cross-lingual voice conversion

In Chapter 3 and Chapter 4, a non-parallel voice conversion with speaker individuality control that can perform in both intra-lingual and cross-lingual voice conversion tasks has been described. However, there is still rooms to enhance the performance of the proposed cross-lingual model. As one of the sub-tasks mentioned in Section 1.3, this chapter proposes the methods for improving the performance of voice conversion. In particular, this chapter concentrates on enhancing the spectral features and prosody features of converted speech. To avoid the over-smoothing problem in the spectral features, the adversarial training scheme of the StarGAN is adopted to improve the training-objective function of the VAE in a CLVC task. An F_0 injection method is also introduced to enhance the F_0 modeling in the cross-lingual situation. In the first section, the introduction of this chapter is provided. Then, the proposed methods for improving the performance of non-parallel voice conversion is introduced in Section 5.2. The experiment procedure and the results are provided in Section 5.3 before the chapter conclusion in Section 5.4.

5.1 Introduction

In Chapter 3, a non-parallel VC model with controllable voice characteristics has been proposed. By using the Principal Component Analysis (PCA), the prominent properties of speaker individuality can be derived from the speaker embedding. However, the VAE-based VC model has three problems when applying for cross-lingual task. First, the learned speaker embedding encodes the speaker's nationality along with other voice characteristics, hence, linguistic information is also affected when modifying the speaker embedding. Second, the previous model is not capable of modeling the F_0 contour, which can be significantly different between languages. Finally, the training objective of VAE model does not implicitly guarantee that the out-

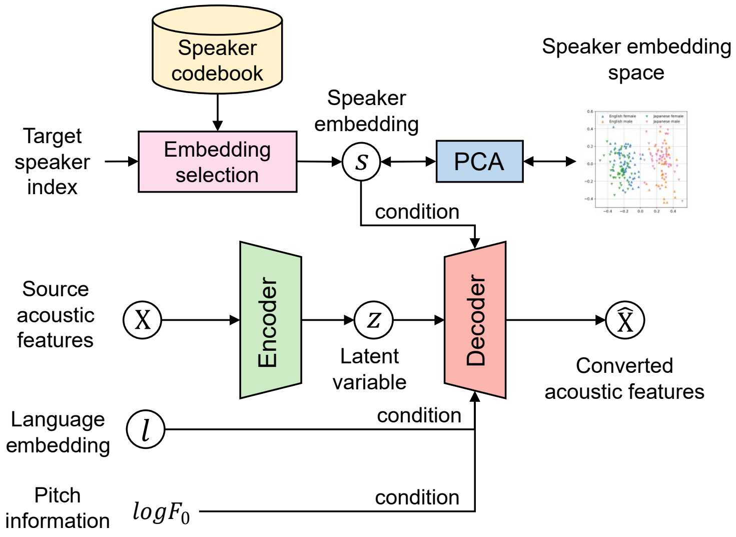


Figure 5.1: Overview of proposed VC system. Voice conversion is performed by selecting the target speaker embedding from the speaker codebook. Each voice characteristic can be independently controlled by means of PCA-projected speaker embedding.

put speech carries the desired voice characteristics corresponding to the input speaker embedding. This limitation reduces the speaker similarity between the converted speech and the target speech. Moreover, using element-wise mean square error in the reconstruction loss implies that the speaker features follow a normal distribution with no correlation across features. This over-simplified objective often leads to over-smoothing degradation, which results in speech that sounds muffled.

Recently, the StarGAN model [24] has been successfully applied for non-parallel multi-speaker voice conversion tasks [22]. The superiority of GAN over other deep generative models arises from its adversarial training scheme, where a generator and discriminator are simultaneously trained to compete with each other. The training process ends when the generator can generate samples indistinguishable from natural ones. This training scheme avoids the use of mean-square-error loss, hence reducing the over-smoothing problem usually found in other VC models. However, the training process of GAN is often very difficult and unstable, which may lead to degradation of the converted speech quality. Moreover, the lack of explicit latent modeling in GAN may discourage the disentanglement between speech content and speaker information, thus reducing the effectiveness of speaker embedding in controlling the voice characteristics.

Therefore, considering the pros and cons of previous studies, this study aims to design a model for text-independent CLVC that can both mimic voice and continuously control the voice characteristics of generated speech. To this end, the previous VAE-based VC model mentioned in Chapter 3 and 4 are improved in several significant ways:

- the additional language embedding is introduced to represent the language property of input speech. By this way, the language factor and speaker individuality factor can be disentangled.
- The $\log F_0$ is directly injected into the decoder to enhance the F_0 modeling and provide controllability over F_0 contour.
- The adversarial training scheme from the Star Generative Adversarial Network (StarGAN) [24] is adopted to improve the objective function of the cross-lingual VAE-based VC model.

Although combining the VAE and GAN has been proposed for non-parallel VC [8, 14], none of these studies focused on the controllability of speaker individuality. The proposed model specifically focuses on the many-to-many CLVC task with controllability of speaker individuality by combining the VAE and StarGAN. To take advantage of the high performance of the recent neural vocoder Parallel WaveGAN [77], the proposed model directly operates in the mel-spectrum domain. Even though continuous speaker embedding has been applied in some VC models [84, 85], they require a trained speaker-recognition model to extract the speaker embedding. In contrast, the proposed model can be trained in an end-to-end fashion by directly optimizing the speaker embedding during the training process. As shown in the next sections, the proposed model improves upon the performance of the previous VAE-based VC model and provides good controllability of speaker individuality by modifying the speaker embedding. Even though the proposed model shares a similar motivation with other VC model regarding F_0 conditioning, there are several differences between them. In general, the propsosed model focuses on cross-lingual VC settings. As different languages might have very different F_0 characteristics, F_0 conditioning helps eliminate the language-dependent factor in the speaker embedding. The previous VAE-based VC model described in Chapter 3 can still work well without F_0 conditioning in an intra-lingual setting [25].

5.2 Proposed method

In this section, the proposed methods for improving the quality of cross-lingual converted speech is described. To enhance the spectral features, the adversarial training scheme for cross-lingual voice conversion is described in

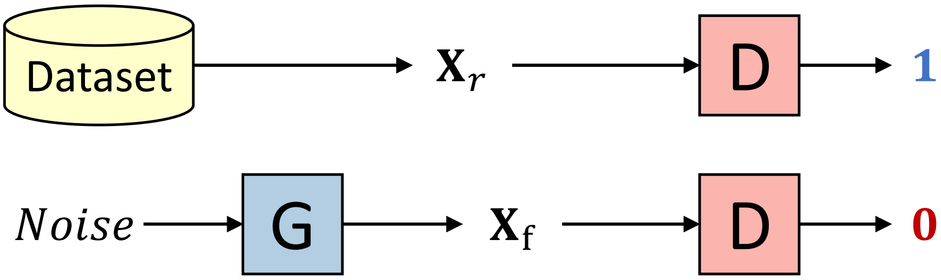


Figure 5.2: Example of GAN consisting of generator (G) and discriminator (D). D distinguishes real sample $\mathbf{X}_r$ and fake sample $\mathbf{X}_f$, which is generated from G. In contrast, G generates more realistic fake sample that can deceive D.

Section 5.2.1. Then, method to improve the prosody of converted speech is introduced in Section 5.2.2. The overview of the proposed CLVC framework is illustrated in Fig. 5.1. Finally, the investigation for language-independent speaker embedding is described in 5.2.3.

5.2.1 Improve spectral detail with Star-Generative Adversarial Network

5.2.1.1 Star-generative-adversarial-network-based voice conversion

A typical GAN consists of two networks, a generator G and discriminator D, which are alternatively trained to compete with each other in an adversarial scheme [86]. On one hand, D is trained to distinguish between the real sample from the training set and the fake sample from G. On the other hand, G is trained to generate samples that could deceive D. Figure 5.2 presents an overview of the conventional GAN structure. The model is converged when D exceeds its capability of classifying the generated samples from real samples. In such a situation, G is expected to generate highly realistic samples.

The conventional GAN can only convert data from one domain to another. To solve the problem of multi-domain generation, the StarGAN [24] was proposed. The goal with the StarGAN is to learn a single G that can map across multiple domains. To achieve this, G is trained to translate the input speech features $\mathbf{x}_r$ into output speech features $\mathbf{x}_f$ conditioned on the target domain label $\mathbf{y}_f$, such that $G(\mathbf{x}_r, \mathbf{x}_f) \rightarrow \mathbf{x}_f$. The target domain label is randomly generated to ensure that G can flexibly translate the input data to different target domains. Simultaneously, D is trained to estimate the probability $D(\mathbf{x}, \mathbf{y})$ of whether $\mathbf{x}$ is authentic, conditioned on $\mathbf{y}$ of the input data. Also, an auxiliary classifier C is trained to predict this label. Figure 5.3

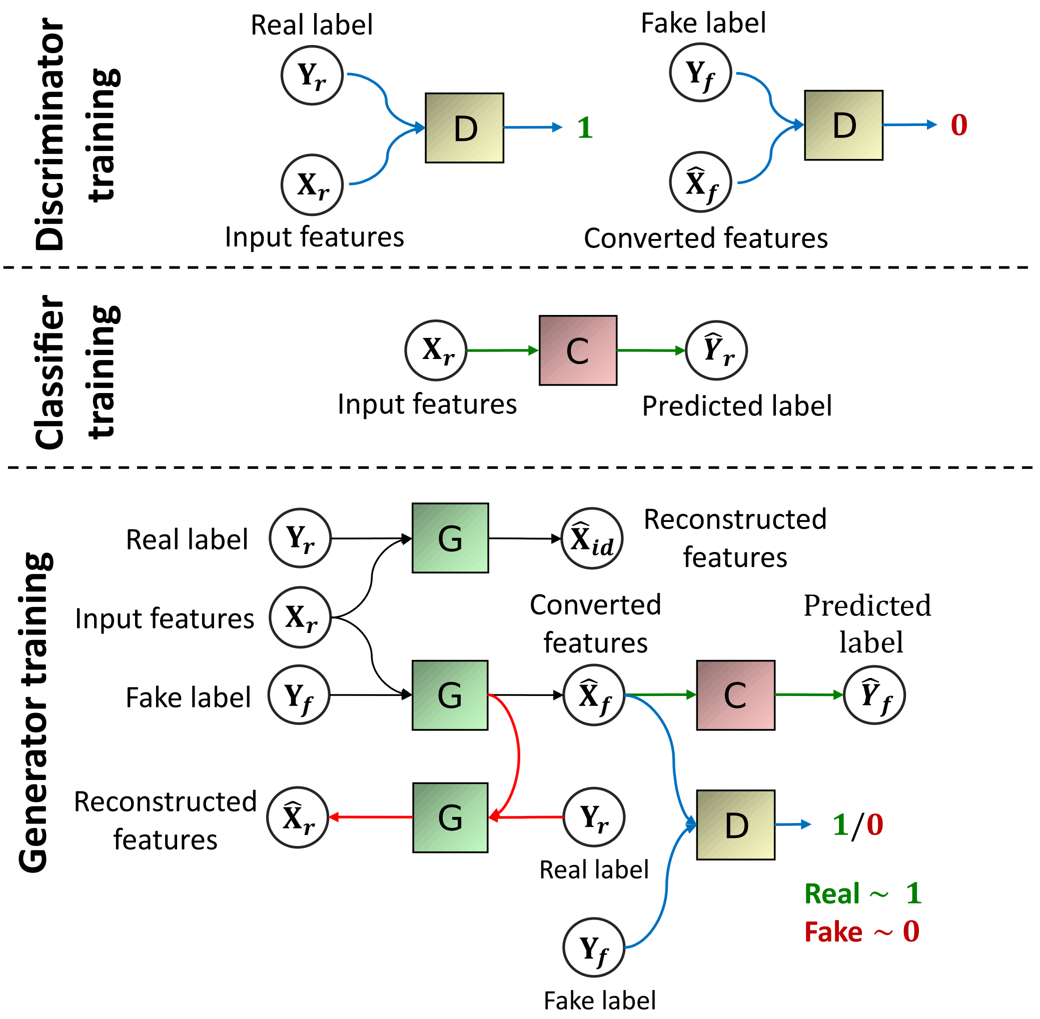


Figure 5.3: Flow chart of StarGAN training process

shows the training process of the StarGAN. The training objective consists of three loss functions, as detailed below.

- **Adversarial loss**: Adversarial loss encourages D to correctly classify real and fake samples while helping G to generate more realistic samples. The adversarial losses for D and G are respectively as follows:

$$\mathcal{L}_{adv}^{D} = -\mathbb{E}_{\mathbf{x}_r,\mathbf{y}_r}[\log D(\mathbf{x}_r, \mathbf{y}_r)] \\ - \mathbb{E}_{\mathbf{x}_r,\mathbf{y}_f}[\log(1 - D(G(\mathbf{x}_r, \mathbf{y}_f)), \mathbf{y}_f))], \quad (5.1)$$

$$\mathcal{L}_{adv}^{G} = -\mathbb{E}_{\mathbf{x}_r,\mathbf{y}_f}[\log(D(G(\mathbf{x}_r, \mathbf{y}_f), \mathbf{y}_f). \quad (5.2)$$

The $\mathcal{L}^D_{adv}$ is reduced when D can correctly classify real and fake samples, while $\mathcal{L}^G_{adv}$ is minimized when G can successfully deceive G.

- **Classification loss**: The C is trained for the speaker-classification task and helps G produce fake data with the correct target speaker voice. In particular, C outputs the probability p_C that $\mathbf{x}$ belong to speaker y. The losses for C and G are defined as

$$\mathcal{L}^C_{cls} = -\mathbb{E}_{\mathbf{x}_r, \mathbf{y}_r}[\log p_C(\mathbf{y}_r|\mathbf{x})], \tag{5.3}$$

$$\mathcal{L}^G_{cls} = -\mathbb{E}_{\mathbf{x}_r, \mathbf{y}_f}[\log p_C(\mathbf{y}_f|G(\mathbf{x}_r, \mathbf{y}_f))]. \tag{5.4}$$

  The $\mathcal{L}^C_{cls}$ is reduced when C can correctly classify to which target speaker the input speech belongs. The $\mathcal{L}^G_{cls}$ is minimized when the converted utterance has similar speaker individuality to the target speaker.

- **Reconstruction loss**: To preserve the linguistic content in the converted utterance, cycle-consistent loss is introduced to regularize G:

$$\mathcal{L}^G_{cyc} = \mathbb{E}_{\mathbf{x}_r, \mathbf{y}_r, \mathbf{y}_f}[||\mathbf{x}_r - G(G(\mathbf{x}_r, \mathbf{y}_f), \mathbf{y}_r)||^2_2], \tag{5.5}$$

  where $\mathbf{y}_r$ and $\mathbf{y}_f$ are the labels of arbitrary source and target speaker, respectively, $\mathbf{x}_r$ is the input speech feature belonging to $\mathbf{y}_r$, and $||\cdot||$ is the Euclidean distance.
  Identity loss is also introduced to keep the converted speech unchanged when the input speech already belongs to $\mathbf{y}_r$:

$$\mathcal{L}^G_{id} = \mathbb{E}_{\mathbf{x}_r, \mathbf{y}_r}[||\mathbf{x}_r - G(\mathbf{x}_r, \mathbf{y}_r)||^2_2]. \tag{5.6}$$

  In summary, the total loss for G is as follows:

$$\mathcal{L}^G = \mathcal{L}^G_{id} + \mathcal{L}^G_{cyc} + \lambda_{adv}\mathcal{L}^G_{adv} + \lambda_{cls}\mathcal{L}^G_{cls}, \tag{5.7}$$

  where λ_{adv} and λ_{cls} are the weighting factor for adversarial loss and classifier loss, respectively.

As seen in the training objective (5.7), the StarGAN does not completely rely on mean-squared-error loss to estimate the distribution of converted acoustic features, as in the VAE. In contrast, G uses feedback from D to produce the most likely sample that can deceive D. Therefore, to avoid over-smoothing in the VAE, the adversarial training scheme of the StarGAN can be adopted to replace the conventional mean-squared-error loss. However, the lack of an explicitly defined latent variable in the StarGAN might reduce the effect of speaker embedding on controlling speaker individuality because G might ignore the input speaker embedding. Hence, the combination of the VAE and StarGAN would alleviate the weakness of the other.

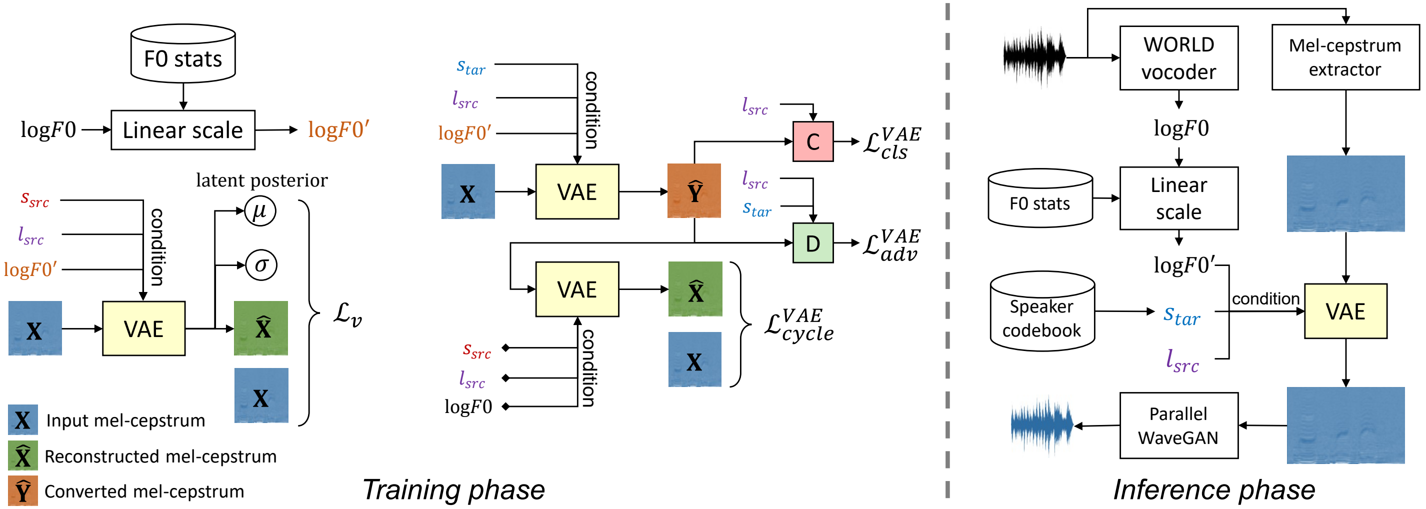


Figure 5.4: Overview of processing flow of proposed model. VAE acts as G of StarGAN. l_{src} refers to language embedding of input mel-cepstrum. s_{src} and s_{tar} are speaker embedding of source and target speakers.

5.2.1.2 Improving spectral details with StarGAN

The proposed model incorporates the StarGAN training scheme [24] as shown in Fig. 5.4. In this model, the VAE acts similarly to the G in the StarGAN. The D identifies whether the input speech is natural or converted given the speaker-identity label. The C learns to classify to which speaker the input speech belongs. Also, the converted voice is re-input to the VAE to convert it back to the source voice. Cycle-consistent loss minimizes the difference between the input features and re-converted features. With all these modifications, the new training objective for the VAE is to 1) generate converted speech to deceive D, 2) minimize the loss from C when inputting the converted speech, 3) minimize cycle-consistent loss, and 4) minimize reconstruction loss and D_{KL} loss.

- **Discriminator loss**: The D distinguishes real and converted speech samples, which are labeled as $\mathbf{1}$ and $-\mathbf{1}$, respectively. To improve the stability of the training process, the Wasserstein distance [87] is used instead of vanilla discriminator loss in (5.1). Therefore, discriminator loss is written as

$$\begin{split}\mathcal{L}_{adv}^{D} = \mathbb{E}_{\mathbf{x},\mathbf{s}_{src}}[1 - D(\mathbf{x}, \mathbf{s}_{src})] \\ + \mathbb{E}_{\mathbf{x},\mathbf{s}_{tar}}[1 + D(\text{VAE}(\mathbf{x}, \mathbf{s}_{tar}), \mathbf{s}_{tar})],\end{split} \tag{5.8}$$

  where $\mathbf{s}_{src}$ and $\mathbf{s}_{tar}$ is the speaker embedding of source and target speakers, respectively, and x is the input acoustic features belonging to the source speaker.

- **Classification loss** The C is trained with cross-entropy loss to identify the correct speaker identity conveyed in the input utterance. The loss for training C is as follows:

$$\mathcal{L}_{cls}^{C} = -\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log p_C(\mathbf{y}|\mathbf{x})] \tag{5.9}$$

  where $\log p_C(\mathbf{y}|\mathbf{x})$ is the output log likelihood that acoustic features x belongs to target speaker $\mathbf{y}$.

- **VAE loss**: In addition to variational loss, adversarial loss and classifier loss encourage the VAE to trick D and reduce the speaker dissimilarity between converted speech and natural speech. The adversarial loss and classifier loss for the VAE are expressed as

$$\begin{aligned}\mathcal{L}_{adv}^{\text{VAE}} &= -\mathbb{E}_{\mathbf{x},\mathbf{s}_{tar}}[D(\text{VAE}(\mathbf{x}, \mathbf{s}_{tar}), \mathbf{s}_{tar})], \\ \mathcal{L}_{cls}^{\text{VAE}} &= -\mathbb{E}_{\mathbf{x},\mathbf{s}_{tar}}[\log p_C(\text{VAE}(\mathbf{x}, \mathbf{s}_{tar}), \mathbf{s}_{tar})].\end{aligned}$$

Similar to the StarGAN training scheme, cycle-consistent loss is introduced to force the VAE to transform the converted features back to the original. This loss is written as

$$\mathcal{L}_{cycle}^{\text{VAE}} = \mathbb{E}_{\mathbf{x},\mathbf{s}_{tar},\mathbf{s}_{src}}[\|\mathbf{x} - \text{VAE}((\text{VAE}(\mathbf{x}, \mathbf{s}_{tar}), \mathbf{s}_{src})\|_2^2]. \tag{5.10}$$

Combined with the variational loss described in (3.3), the final training objective for the proposed model now becomes

$$\mathcal{L}_{obj}^{\text{VAE}} = \mathcal{L}_v + \mathcal{L}_{cycle}^{\text{VAE}} + \lambda_{adv}\mathcal{L}_{adv}^{\text{VAE}} + \lambda_{cls}\mathcal{L}_{cls}^{\text{VAE}}, \tag{5.11}$$

where λ_{adv} and λ_{cls} are the weight factor for each loss component. In empirical testing, $\lambda_{adv} = 0.0005$ and $\lambda_{cls} = 0.0001$ showed good results in this study.

5.2.2 Improving F_0 generation

Various high-performance vocoders based on deep neural networks have recently been proposed [77, 88, 89]. Most of these neural vocoders directly use mel-spectrum as the input feature. However, it is difficult to directly manipulate the F_0 information in mel-spectrogram, as it relates to the harmonic structure. In addition, different languages may have very different F_0 contours, which can degrade the cross-lingual converted speech with spurious pitch. To provide the controllability and stability of F_0 in converted speech, the mel-spectrogram generation is directly conditioned with $\log F_0$ input, as shown in Figs. 5.4 and 5.6.This method is referred as the *F0 injection* method. To generate fake samples during the training or inference phases, the source $\log F_0$ is linearly scaled to match the target F_0 mean-variance. Therefore, the statistics of the target F_0 must be pre-calculated for VC.

5.2.3 Controlling speaker individuality in cross-lingual setting

In conventional VAE-based VC, speaker identity is usually represented as a one-hot vector [9]. However, this type of encoding does not allow controllability of speaker individuality. Some studies have proposed using d-vector to represent speaker individuality, but this type of speaker representation requires an additional speaker-recognition network, which introduces more complexity to the VC model. The previous VAE-based VC model was developed for continuous learnable speaker embedding that can be jointly learned with other network parameters during the training process [25].

This model does not require any addition speaker-recognition network yet still achieves controllability of speaker individuality. Let $\mathbf{y}$ is the one-hot vector represents the speaker identity, the continuous speaker embedding $\mathbf{s}$ is calculated by using a simple linear transformation, as

$$\mathbf{s} = \mathbf{W}^\intercal \cdot \mathbf{y} + \mathbf{b}, \tag{5.12}$$

where $\mathbf{W}$ and $\mathbf{b}$ is a learnable kernel and bias in a fully-connected neural network layer. In this interpretation, the one-hot encoded vector $\mathbf{y}$ acts as a switch to select the correspond row vector in matrix $\mathbf{W}$. In the case of $\mathbf{b} = \mathbf{0}$, each row vector in the kernel matrix $\mathbf{W}$ can be seen as a speaker embedding. Fig. 5.5 illustrates the fist and second principal components of the learned speaker embeddings of the VCTK dataset [62]. As can be seen, the speaker are clearly clusterized based on the voice gender and input language, hence the speaker embedding can encode useful information about the speaker individuality. This behavior is undesirable because manipulating the speaker embedding would affect the linguistic content due to language differences. To avoid this problem, an additional language embedding is introduced to disentangle the language factor from speaker embedding. In this study, the language factor is simply represented by a one-hot encoded vector, which is concatenated with the speaker embedding along the channel dimension. The combined vector is then used to condition the decoder on generating the mel-spectrogram, as shown in Fig. 5.4.

5.3 Experiments

To evaluate the performance of the proposed model, the cross-lingual voice conversion between English and Japanese speakers is conducted using three models: the conventional VAE (**VAE**), StarGAN (**StarGAN**), and proposed model (**VAE-StarGAN**). To evaluate the effectiveness of F_0 injection, a VAE-based VC model trained without F_0 input is also implemented. This model is denoted as **VAE-noF0**. For a fair comparison, **VAE** and **VAE-StarGAN** had the same network structure. In addition, the classifiers of **StarGAN** and **VAE-StarGAN** had an identical structure.

To train the models, two open-source multi-speaker voice databases is used: the English VCTK corpus [62] and the Japanese Versatile Speech (JVS) corpus [90]. The training data included 100 speakers from the English VCTK dataset and 100 speakers from the JVS dataset. For each speaker, 100 utterances were randomly selected as training data and ten utterances as testing data. Each speaker was initially assigned to a random speaker

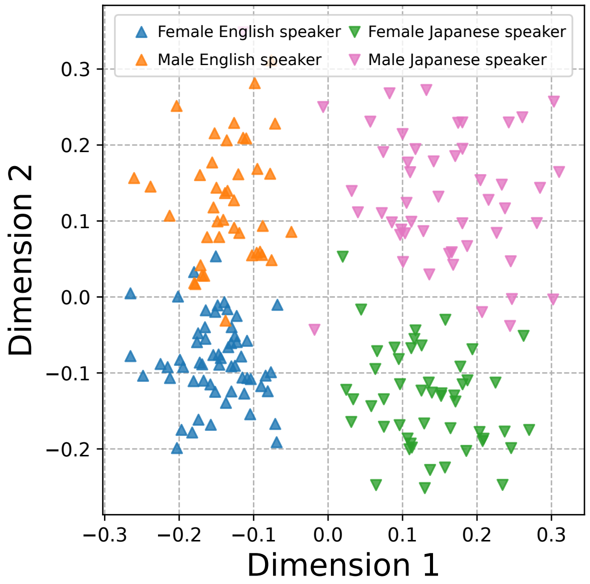


Figure 5.5: 2D visualization of speaker embedding learned using intra-lingual VAE-based VC model using PCA. Speaker embeddings are clustered on basis of voice gender and speaker language.

embedding. To condition the decoder on the language of the input mel-cepstrum, a one-hot embedding vector is used for language. Since there were two input languages (English and Japanese), the number of dimensions for language embedding was two.

5.3.1 Preprocessing

In the preprocessing step, the audio waveform was down-sampled to 24 kHz and normalized to the $[-1.0, 1.0]$ range. Then, an 80-dimensional mel-spectrogram was extracted using short-time Fourier transform (STFT) and mel-filterbank. The window length of STFT was set to 2048 and the hop-length was 300. The mel-filterbank spanned from 80 to $7600Hz$ to match the Parallel WaveGAN input. Then, the mel-spectrum was transformed into mel-cepstrum by applying inverse discrete Fourier transform on the log-magnitude mel-spectrum. Although some studies further normalized each mel channel by its mean and variance across the time dimension, preliminary experiment shows that this step degrades the quality of converted speech. Therefore, the raw mel-cepstrum value is directly used as the input feature. In addition to the mel-cepstrum feature, F_0 was extracted using the WORLD

Table 5.1: Network architecture of VAE encoder and decoder, D, and C.

Network	No. of WN cells	Dilation rate	Filters	Kernel size
Encoder	6	2×[1, 2, 4]	128	5
Decoder	16	4×[1, 2, 4, 8]	128	5
D	3	[1, 1, 1]	[128, 256, 512]	3
C	3	[1, 1, 1]	[128, 256, 512]	3

analysis system [63]. After extracting theF_0 from all utterances, the mean and variance of $\log F_0$ for each speaker are calculated for the linear scaling functions. The Parallel WaveGAN vocoder [77] trained on the VCTK dataset for 1000k iterations is used for waveform generation.

5.3.2 Network Architecture

Similar to the previous model in Chapter 3, the encoder and decoder of the VAE were constructed from a smaller network that resembles the WaveNet (WN) architecture [40]. Figure 5.6 shows the architecture of a WN cell. The input layer for the hidden variable h_n is the 1D dilated convolutional neural network [91], which expands the receptive field in the temporal dimension by dilation in the kernel. The details of the model parameters of the VAE encoder and decoder, D, and C are provided in Table 5.1.

The D and C share the same architecture, as illustrated in Fig. 5.7. Each WN cell is followed by a stride 1D convolution layer to reduce the temporal dimension by half after each stage. At the output, a fully connected layer consumes the compressed vector to produce the output vector. The speaker embedding and language embedding are represented as a one-hot vector. Both D and C are conditioned on both the speaker-embedding and language-embedding vectors, while C is conditioned only on the language embedding vector.

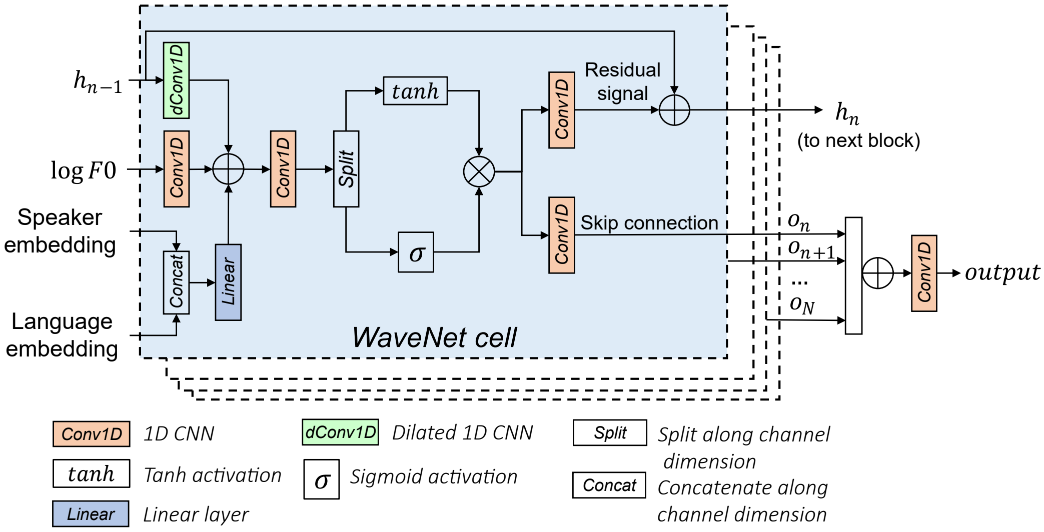
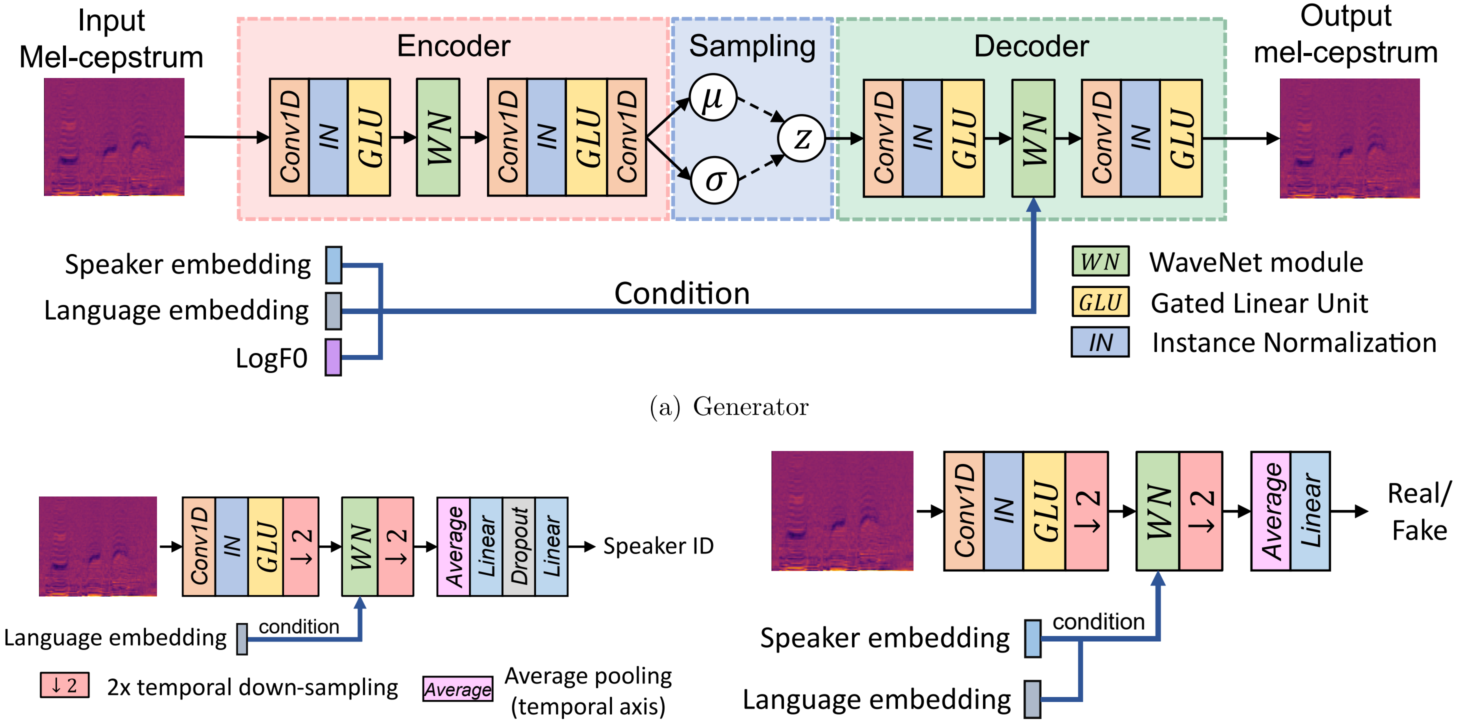


Figure 5.6: Stacks of WaveNet cells in WaveNet module.



(a) Generator

(b) Classifier

(c) Discriminator

Figure 5.7: Structure of (a) generator network, (b) classifier network, and (c) discriminator network.

Algorithm 1: VAE-StarGAN training procedure

```
Require: Functions $G$ (VAE model), $D$, $C$, $Scale$ ($\log F_0$ linear
         scale function), $\mathbf{X}$ (batch of source mel-cepstrum), $f0$
         (batch of source $\log F_0$ ), $\mathbf{s}_{src}$ (source speaker embedding),
         $\mathbf{s}_{tar}$ (target speaker embedding), $\mathbf{l}_{src}$ (source language
         embedding)
▷ Update the discriminator parameter $\theta_D$
$f0_f \leftarrow Scale(f0)$
$\mathbf{x}_f \leftarrow G(\mathbf{x}, f0_f, \mathbf{s}_{tar}, \mathbf{l}_{src})$
$\mathbf{d}_r \leftarrow max(0, 1 - D(\mathbf{x}, \mathbf{s}_{src}, \mathbf{l}_{src}))$
$\mathbf{d}_f \leftarrow max(0, 1 + D(\mathbf{x}_f, \mathbf{s}_{tar}, \mathbf{l}_{src}))$
$\mathcal{L}^{D}_{adv} \leftarrow \frac{\mathbf{d}_r + \mathbf{d}_f}{2}$
update $\theta_D$ to minimize $\mathcal{L}^{D}_{adv}$
▷ Update classifier parameter $\theta_C$
$\mathcal{L}^{C}_{cls} \leftarrow \text{CrossEntropy}(\mathbf{s}_{src}, C(\mathbf{x}, \mathbf{l}_{src}))$ update $\theta_C$ to minimize $\mathcal{L}^{C}_{cls}$
▷ Update VAE parameter $\theta_{\text{VAE}}$
$\mathbf{x}_{id}, \boldsymbol{\mu}_z, \boldsymbol{\sigma}_z \leftarrow G(\mathbf{x}, f0, \mathbf{s}_{src}, \mathbf{l}_{src})$
$\mathbf{x}_{cycle} \leftarrow G(\mathbf{x}_f, f0, \mathbf{s}_{src}, \mathbf{l}_{src})$
$\mathcal{L}^{VAE}_{adv} \leftarrow -D(\mathbf{x}_f, \mathbf{s}_{tar}, \mathbf{l}_{src})$
$\mathcal{L}^{VAE}_{cls} \leftarrow \text{CrossEntropy}(\mathbf{s}_{tar}, C(\mathbf{x}_f, \mathbf{l}_{src}))$
$\mathcal{L}^{VAE}_{cycle} \leftarrow \|\mathbf{x}_{cycle} - \mathbf{x}\|_2^2$
$\mathcal{L}_v \leftarrow \|\mathbf{x}_{id} - \mathbf{x}\|_2^2 - \frac{1}{2}(1 + \log \boldsymbol{\sigma}_z^2 - \boldsymbol{\mu}_z^2 - \boldsymbol{\sigma}_z^2)$
▷ Calculate 5.11
$\mathcal{L}^{\text{VAE}} \leftarrow \mathcal{L}_v + \mathcal{L}^{\text{VAE}}_{cycle} + \lambda_{adv}\mathcal{L}^{\text{VAE}}_{adv} + \lambda_{cls}\mathcal{L}^{\text{VAE}}_{cls}$
update $\theta_G$ to minimize $\mathcal{L}^{\text{VAE}}$
```

5.3.3 Training Procedure

All models were trained using the Adam optimizer [78] with 32 samples per batch. The mel-cepstrum is truncated or warp-padded to have 512 frames. The learning rate is initialized at 2×10^{-4} and gradually reduced to 1×10^{-4} for the first ten epochs. The training process was conducted using two Nvidia 2080Ti GPUs until the model converged, which took roughly two days for each model. The detailed training procedure for **StarGAN** and **VAE-StarGAN** is shown in Algorithm 1.

5.3.4 Visualizing speaker embedding

After the VC model was trained, the speaker-embedding space is visualized in Fig. 5.8 by analyzing the speaker codebook using PCA. Figure 5.8(a) illustrates the PCA-projected speaker embedding learned using the previous VAE-based VC model [25]. Without the input language embedding, it can be seen that the language of the speakers was separated on the first principal dimension. On the other hand, as shown in Fig. 5.8(b), only the speaker's sex was separated on the first principal dimension when the model was trained with language embedding input. Moreover, the clustering effect on language was removed, as there was no clear separation between Japanese speakers and English speakers. This result indicates that the speaker embedding can encode useful information from the speaker individuality while still remaining language-independent.

5.3.5 Objective Evaluation

Different objective measurements are selected to evaluate the performance of the proposed model. The objective evaluation set consists of cross-lingual converted utterances from English to Japanese and Japanese to English. Five male and five female speakers are selected from each language to form 200 conversion pairs, and each pair had ten converted samples. Therefore, the objective evaluation set consisted of 2000 converted utterances.

5.3.5.1 Modulation spectrum measurement

The modulation spectrum (MS) can provide hints about speech naturalness: a higher MS corresponds to better speech naturalness. Following the work of Takamichi et al. [79], the MS of the converted mel-cepstral sequence is calculated by taking the Fourier transform along the temporal dimension. Similar to a previous study [68], the MS was averaged for all modulation frequencies and all utterances as

$$MS = \frac{1}{N}\frac{1}{F}\sum_{n}^{N}\sum_{f}^{F}|DFT[\mathbf{X(n,f)}]|, \tag{5.13}$$

where $\mathbf{X}$ is a batch of test utterances, N is the number of utterances, $n \in [0, N)$ is the utterance index, F is the number of MS frequency bins, and $f \in [0, F)$ is the MS frequency bins. As shown in Fig. 5.9, **VAE-StarGAN** achieved a higher log-scaled MS on the lower mel-cepstral coefficients than the previous VAE-based VC model. These results indicate that the adversarial training scheme can lessen the over-smoothing of converted mel-cepstral coefficients. Figure 5.10 illustrates the mel-spectrogram generated from different models. It can be seen that the StarGAN and the VAE-StarGAN produced mel-spectrograms with a more detailed structure. Although the

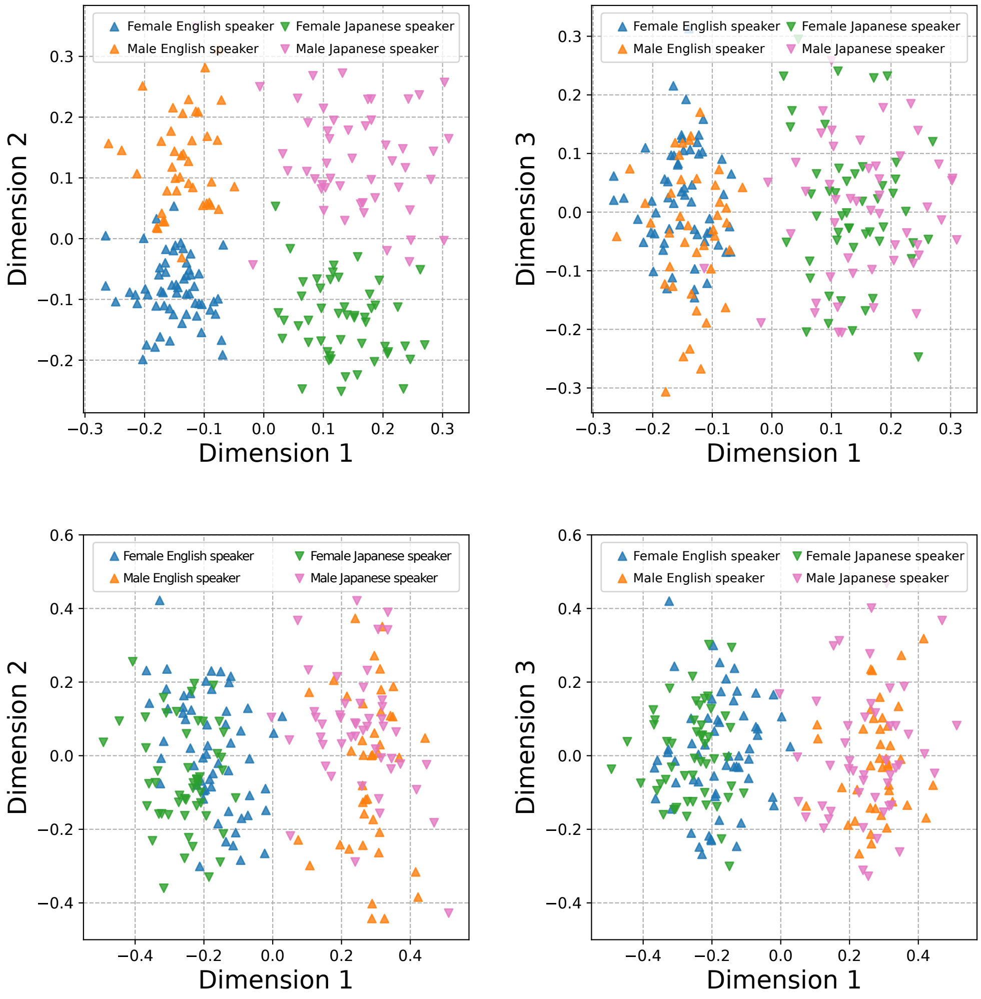
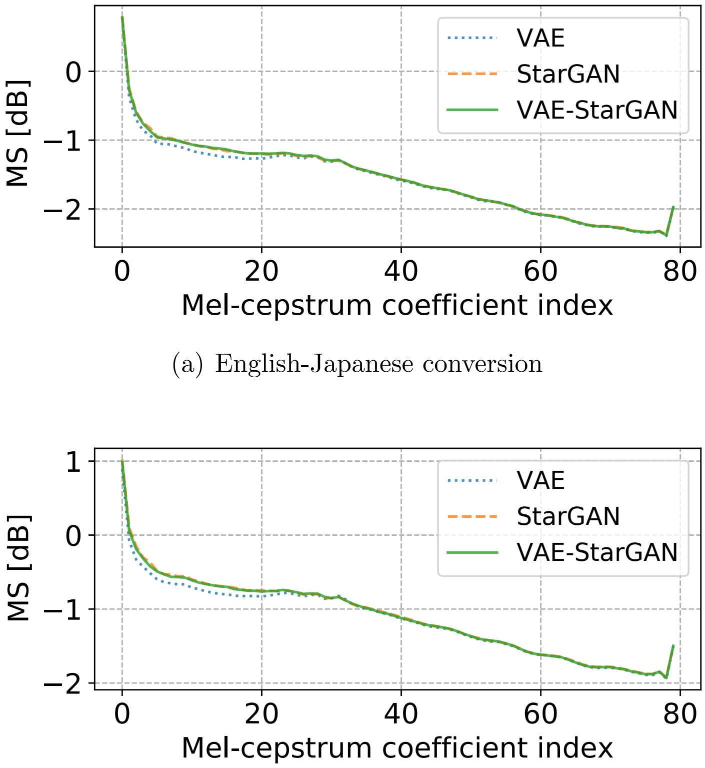


Figure 5.8: 2D PCA visualization of speaker embedding from model (a) without language embedding input and (b) with language embedding input. Speaker embedding from English and Japanese speakers are clearly separated into distinct clusters when language embedding is not used.



(a) English-Japanese conversion

(b) Japanese-English conversion

Figure 5.9: Log-scaled modulation spectrum of natural speech, reconstructed speech, and converted speech averaged over all utterances and modulation frequencies. **StarGAN** and **VAE-StarGAN** generated mel-spectrograms with higher MS than **VAE**.

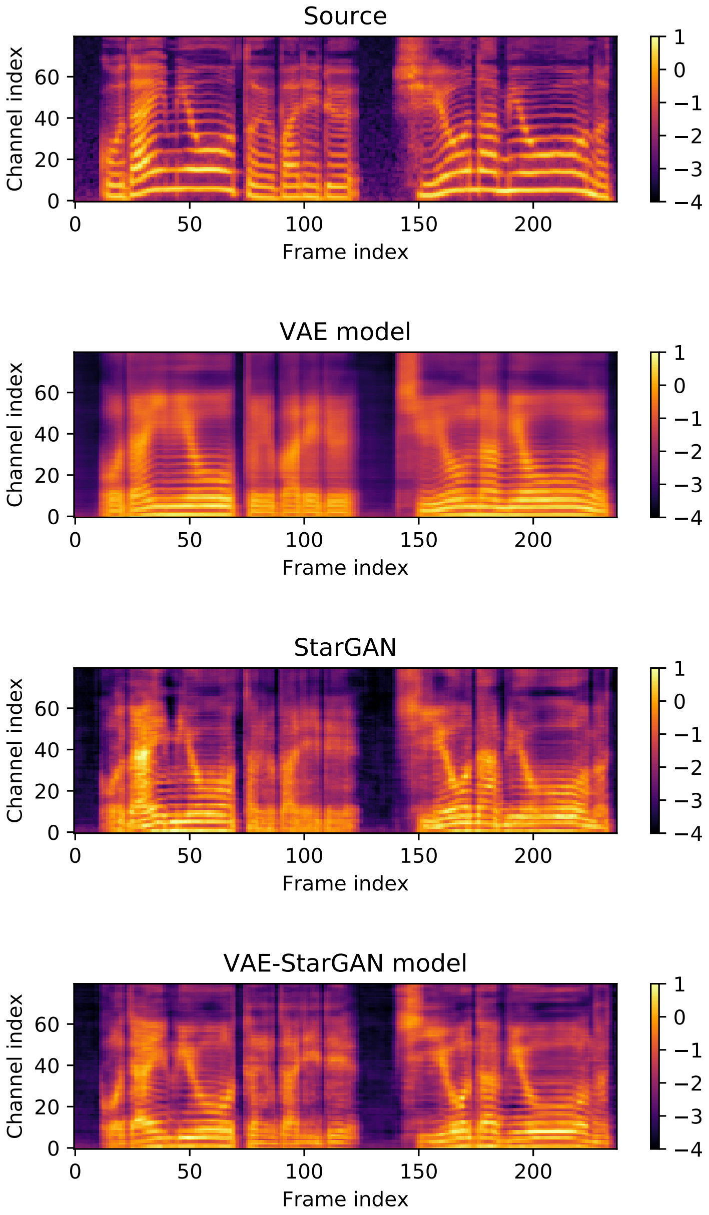


Figure 5.10: Mel-spectrogram of source voice and converted voice from different models. Mel-spectrogram generated from **StarGAN** and **VAE-StarGAN** clearly had more details than that generated from **VAE**.

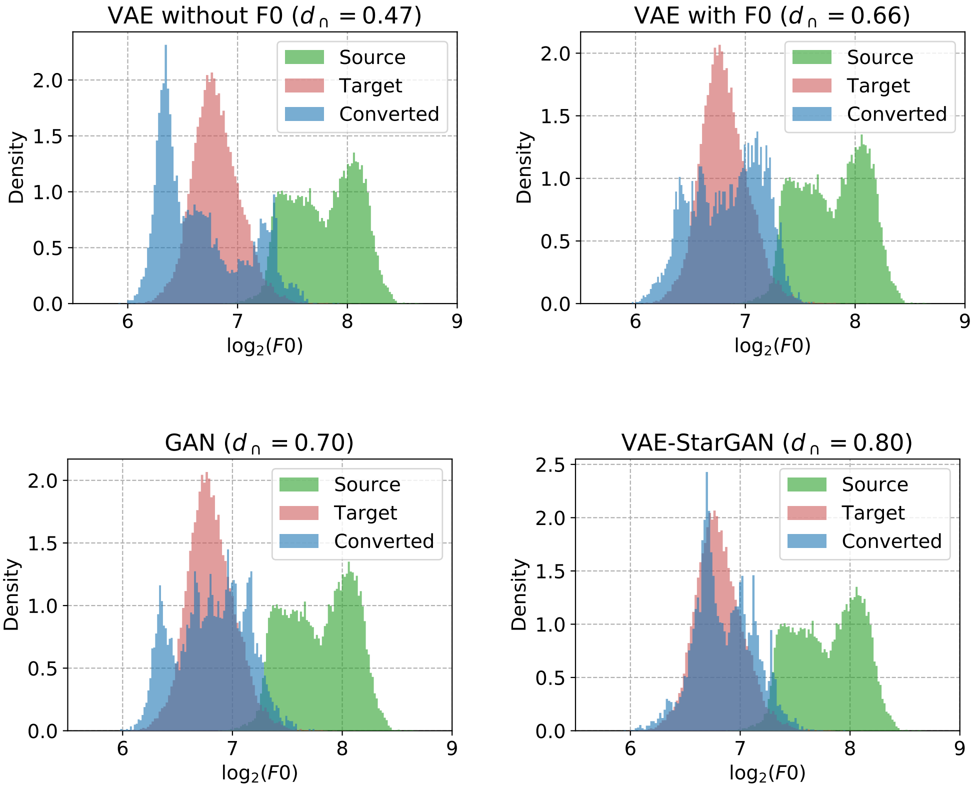
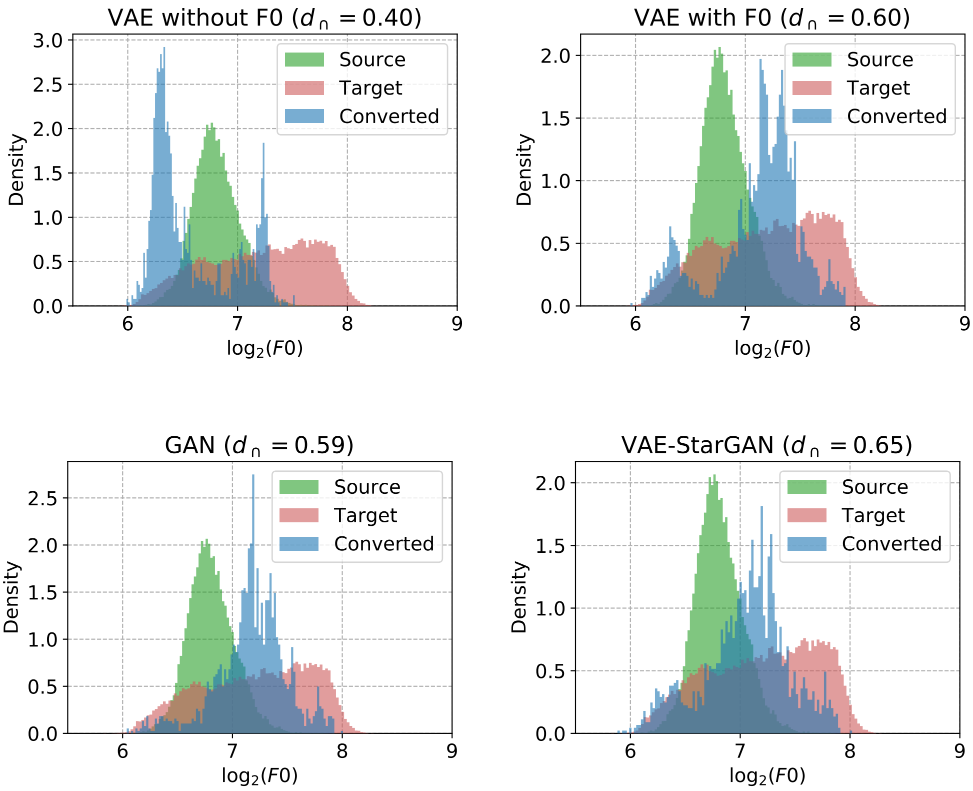


Figure 5.11: Distribution of $\log_2 F_0$ of source, target, and converted speech of Japanese-speaking female to English-speaking male conversion. Intersection index $d_\cap$ indicates amount of overlap between converted $\log_2 F_0$ and target $log_2 F_0$.

mel-spectrum of **VAE-StarGAN** was more refined than that of VAE, artifacts such as mispronunciation cannot be clearly shown on the mel-spectrum. Therefore, a listening test must be conducted to precisely compare the performances of different models.



(b)

Figure 5.12: Distribution of $\log_2 F_0$ of source, target, and converted speech of English-speaking male to Japanese-speaking male conversion. Intersection index $d_\cap$ indicates amount of overlap between converted $\log_2 F_0$ and target $log_2 F_0$.

Table 5.2: Average scores and standard deviations of F_0 analysis results from different models. For mean F_0 error and voice/unvoiced error rate (v/uv) error, lower is better. For histogram intersection, higher is better.

Test / model	All			English to Japanese			Japanese to English		
	Mean F0 error	v/uv error (%)	Histogram intersection	Mean F_0 error	v/uv error (%)	Histogram intersection	Mean F0 error	v/uv error (%)	Histogram intersection
VAE-noF0	0.286 ± 0.20	0.191 ± 0.03	0.506 ± 0.08	0.419 ± 0.17	0.198 ± 0.03	0.481 ± 0.07	0.151 ± 0.12	0.184 ± 0.03	0.532 ± 0.08
VAE	0.132 ± 0.09	0.152 ± 0.03	0.563 ± 0.13	0.173 ± 0.07	0.160 ± 0.03	0.564 ± 0.12	0.090 ± 0.09	0.143 ± 0.02	0.562 ± 0.13
StarGAN	**0.082** ± 0.06	**0.141** ± 0.02	0.568 ± 0.11	**0.080** ± 0.05	**0.135** ± 0.02	0.567 ± 0.08	0.084 ± 0.07	0.147 ± 0.02	0.568 ± 0.13
VAE-StarGAN	0.128 ± 0.09	0.148 ± 0.02	**0.580** ± 0.12	0.173 ± 0.08	0.154 ± 0.02	**0.581** ± 0.09	**0.083** ± 0.08	**0.143** ± 0.01	**0.578** ± 0.14

5.3.5.2 F0 injection

To measure the effectiveness of F_0 injection method, the F_0 histogram intersection [92] is measured between converted speech and target speech. The histogram intersection can indicate the amount of similarity between two distributions. Given the histogram of converted speech P and that of target speech Q, where each one contains n bins, the histogram intersection is defined as follows:

$$d_{\cap}(P, Q) = \frac{\sum_j^n min(P_j, Q_j)}{\sum_j^n Q_j}. \tag{5.14}$$

The maximum histogram intersection $d_{\cap max} = 1$ is achieved when P and Q are completely identical. In the experiments, the number of bins is set to $n = 1000$ and the range of $\log_2 F_0$ is between 5.5 and 9.0. Figure 5.3.5.1 and 5.3.5.1 shows a comparison of the $\log_2 F_0$ distribution between source, target, and converted utterances from different models. It can be seen that the $\log_2 F_0$ distribution does not always follow the Gaussian shape. Therefore, simply performing linear transformation of F_0 extracted from a parametric vocoder (e.g., WORLD or STRAIGHT [22,85,93]) cannot ensure the correct shape of F_0 distribution.

In addition to histogram intersection, the average error between the mean of converted $\log_2 F_0$ and that of target $\log_2 F_0$ is measured. The voice/unvoiced error rate between converted F_0 and source F_0 is also calculated. The results are summarized in Table 5.2. It can be seen that the models with F_0 injection had a significantly higher histogram intersection, lower v/uv error rate, and lower mean F_0 error than the model without. The two-tailed t-test showed that the effect of using the F_0 injection method is statistically significant. These results indicate that the F_0 injection method can improve the performance of VC models for controlling the F_0 in the converted utterance.

5.3.6 Subjective Evaluation

Listening tests are conducted to evaluate the speech naturalness and speaker similarity of the converted utterances. one male and one female speaker from each language are selected, for a total of four speakers in the evaluation set. Since only CLVC was carried out, there were eight combinations from the selected speakers. Japanese-to-English conversion is denoted as "**SJ-TE**" and English-to-Japanese conversion is denoted as "**TE-SJ**". Two sentences were selected from each source-target pair to create the listening test set. Therefore, the listening test set consisted of 48 pairs of converted utterances

(2 sentences × 8 source-target speaker pairs × 3 model pairs). For reference stimuli in the ABX similarity test, the original utterances of the target speakers are randomly selected from the training set. Nine individuals with normal listening ability participated in both listening tests. All participants had a basic level of using Japanese/English even if Japanese/English was not their first language. Each participant rated 24 random pairs of converted utterances for each test via an online interface.

To measure speaker similarity, the ABX test scheme was used to compare the performance of **VAE-StarGAN**, **StarGAN**, and **VAE**. Listeners were asked to select the closest utterance ("A" or "B") to the reference utterance X or choose *Same* if there was no difference. The X is the natural speech of the target speaker selected from the test set, while utterances "A" and "B" are generated from different models. For speech naturalness, the AB test scheme is applied, in which listeners were asked to determine the more natural utterance ("A" or "B") or choose *Same* if there was no difference. The generated utterance from both models was presented in random order (AB or BA) to avoid any bias. To analyze the results, the one-way ANOVA test is used with alpha value of 0.05.

As shown in Figs. 5.13 and 5.14, **VAE-StarGAN** outperformed **StarGAN** for both naturalness and similarity in all cases. Except for the similarity score of SE-TJ conversion, these differences are statistical significant. When comparing with the **VAE**, the one-way ANOVA test and the post-hoc two-tailed t-test determined that **VAE-StarGAN** had a statistically better similarity score than **VAE** in SJ-TE conversion. However, no significant difference was observed between these two models in other cases. **VAE** had better naturalness and similarity scores than **StarGAN** in most cases except for the SE-TJ similarity score. The reason might be that although the converted speech from **StarGAN** sounded less muffled than that from **VAE**, artifacts such as mispronunciation severely affected the perceived speech naturalness. The low preference score of **VAE-StarGAN** for speaker similarity indicates that the speaker embedding of **StarGAN** has less controllability on speaker individuality than **VAE** and **VAE-StarGAN**. This behavior may be due to the lack of explicit latent modeling in **StarGAN**, which discourages the disentanglement between speech content and speaker information.

5.3.7 Fictitious Speaker

To evaluate the controllability of speaker individuality with **VAE-StarGAN**, 11 test utterances were generated by linearly interpolating the speaker embedding between the source and target speaker embeddings. The source speaker was a female Japanese speaker and the target speaker was a male

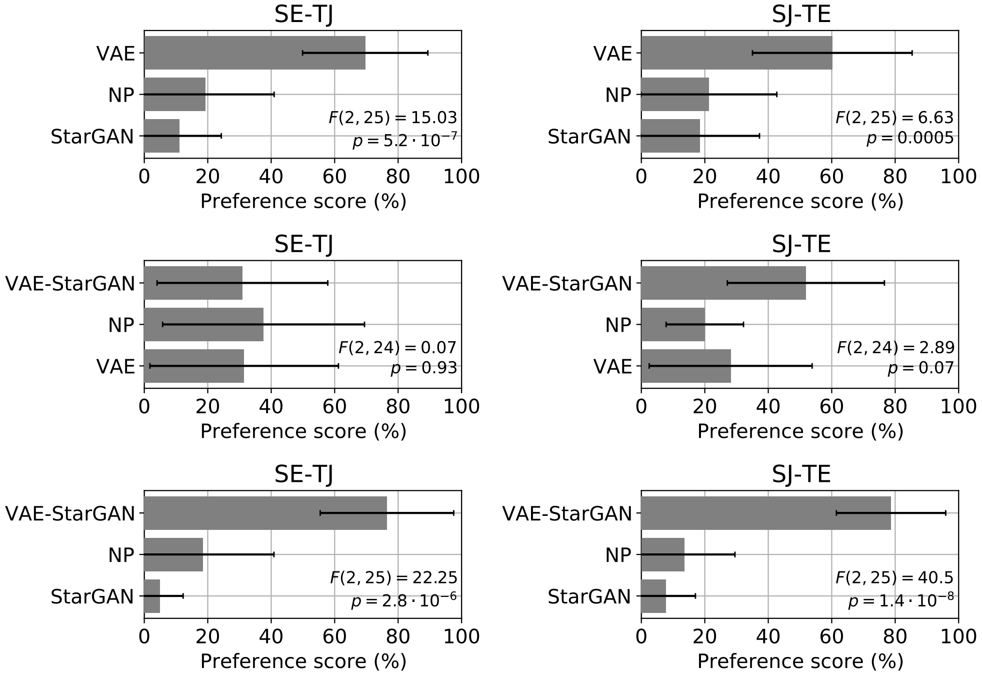


Figure 5.13: Preference scores of AB naturalness test with 95-percent confidence interval and results from one-way ANOVA test. **NP** means no preference.

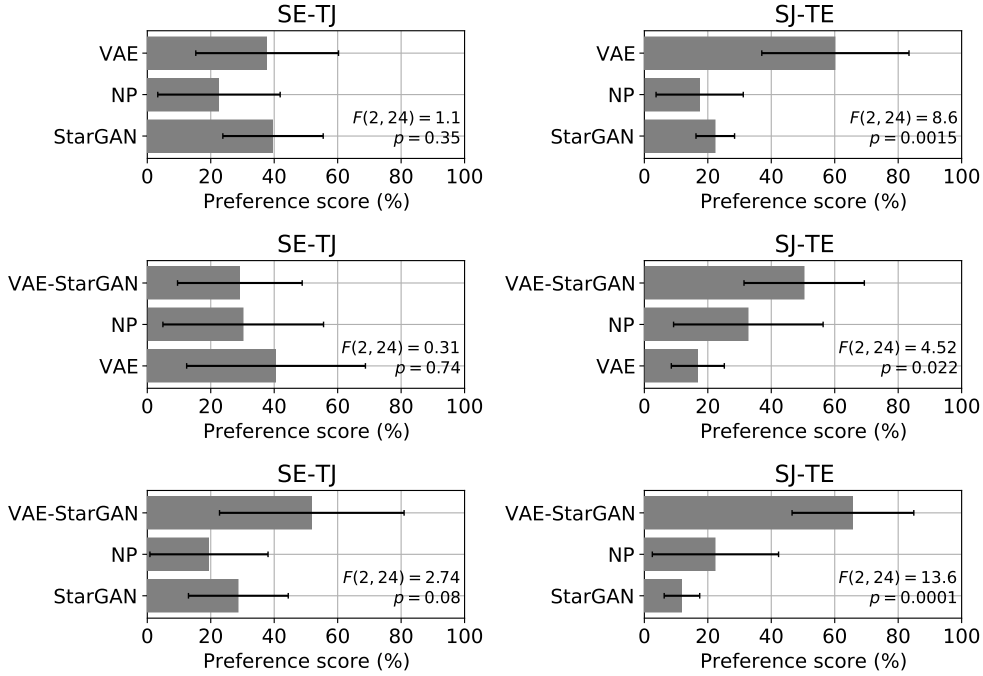


Figure 5.14: Preference scores of ABX speaker similarity test with 95-percent confidence interval and results from one-way ANOVA test

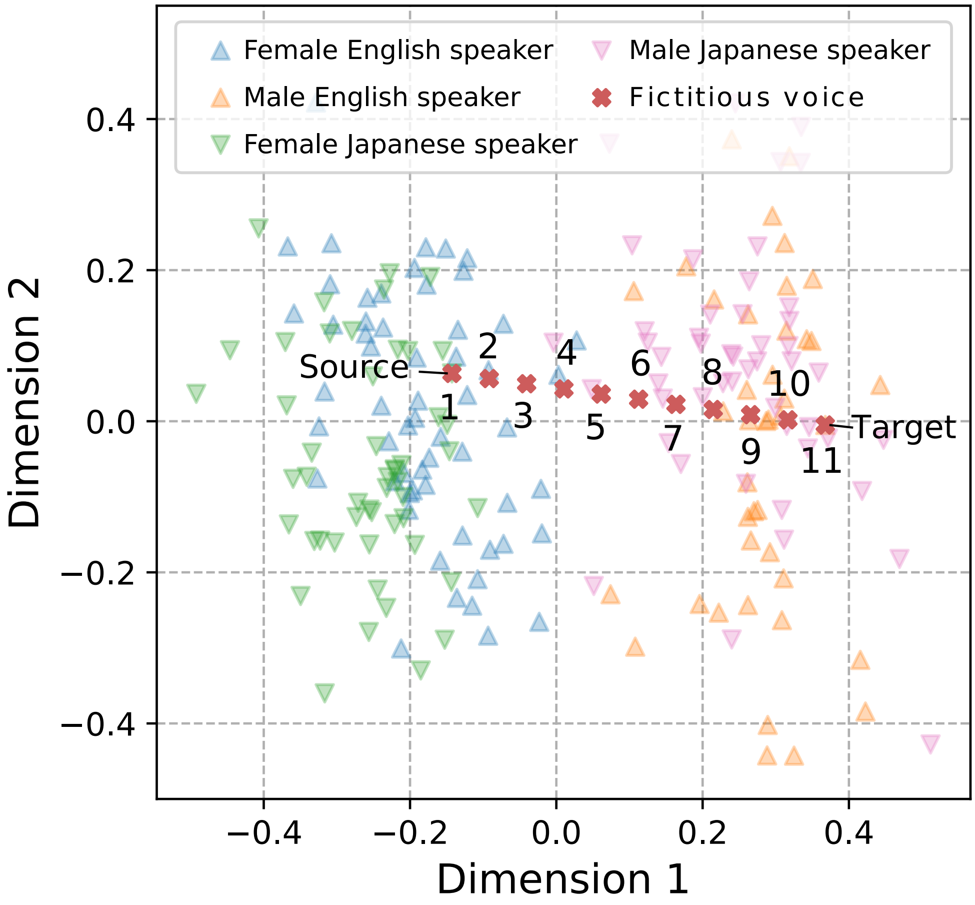


Figure 5.15: Position of linearly interpolated speaker embedding between source female Japanese speaker and target male English speaker. Index of each converted utterance is marked from 1 to 11.

English speaker. The positions of the interpolated speaker embedding **s** are shown in Fig. 5.15. The input F_0 was also transformed using the linearly interpolated mean and standard deviation between the source and target F0.

Each test utterance was marked from 1 to 11 with respect to its position on the speaker-embedding map. In this test, the participants listened to the test stimuli in random order to avoid any bias then were asked to judge the similarity between test stimuli and the reference utterance on a scale from 0 to 100. Figure 5.16 shows the average similarity score of each test utterance. The Pearson correlation coefficient is used to evaluate the linear relationship between average similarity scores and expected similarity scores, which is calculated as

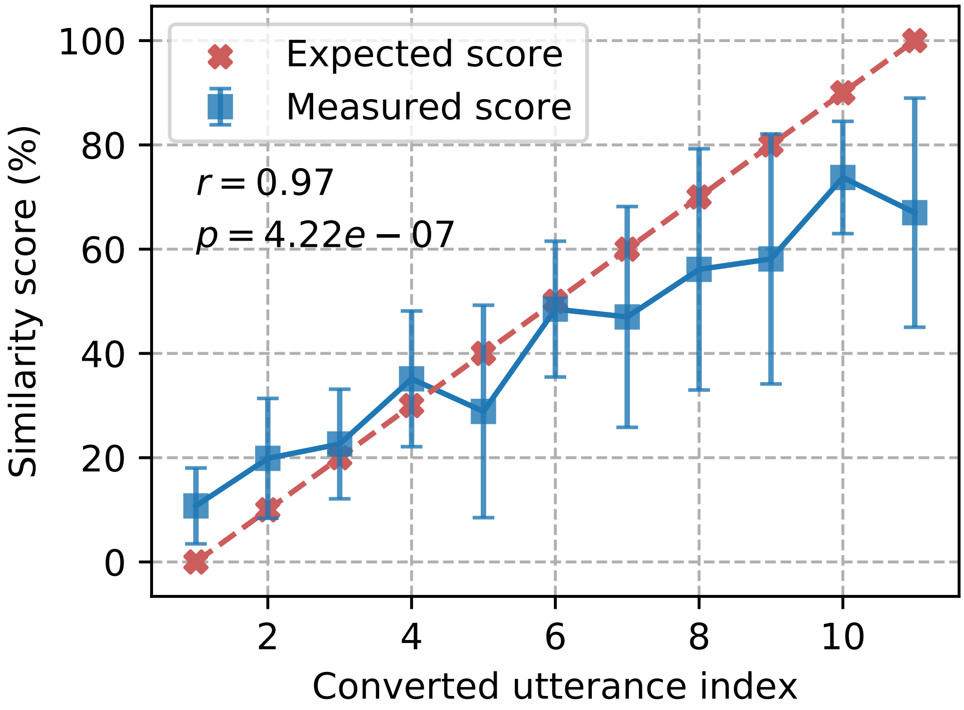


Figure 5.16: Similarity scores of interpolated speaker embedding with standard deviation. Dotted line denotes expected similarity score that linearly increased from 0 to 100. r and p indicate Pearson correlation and p-value, respectively.

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}, \tag{5.15}$$

where m_x is the mean of vector x and m_y is the mean of vector y. The correlations of $+1$ or -1 suggest an exact linear relationship. The measured correlation was $r = 0.97$ and the p-value was $p = 4.22 \times 10^{-7}$, which indicates that the average similarity scores have a strong positive correlation with the expected similarity score, thus statistically sufficient.

5.4 Chapter Conclusion

In this chapter, improvement methods for cross-lingual voice conversion system based on the VAE-StarGAN model are proposed. The results from this chapter showed that the proposed VC model, which is trained solely on acoustic features, can effectively control the speaker individuality in a cross-lingual setting via the speaker embedding. In terms of the over-smoothing degradation problem, the objective results showed that the proposed adversarial training scheme can effectively enhance the fine-structure in the converted mel-spectrogram. The results from the subjective test show that the improvement in SJ-TE conversion is statistical significant. With the additional language embedding, the nationality factor can be disentangled from the speaker embedding, hence avoiding the effect on linguistic information when converting voice. Moreover, the results from objective measurements indicated that the F_0 injection can improve the F_0 modeling in a cross-lingual voice conversion scheme, which suggests the potential of using modern neural vocoders in the VC system to enhance the quality of converted speech. Moreover, the high correlation between the average similarity score of the fictitious voice and the expected similarity score is evidence for a strong linear relation between speaker embedding and perceptual speaker similarity, which indicates the controllability of speaker individuality of the proposed model. Based on these results, the proposed methods in this chapter have effectively improved the performance of cross-lingual voice conversion, therefore, the goal of the third sub-task defined in Section 1.4 has been achieved.

Chapter 6

Conclusion and Future Direction

6.1 Summary of the thesis

In the previous chapters, a speech features disentanglement framework for intra-lingual and cross-lingual voice conversion has been considered. To control the speaker individuality, a variational autoencoder-based non-parallel voice conversion with continuous-controllable speaker embedding has been proposed. As mentioned in Section 1.3, the challenges of this framework are: 1) the difficulty of modelling speaker individuality in a continuous domain and 2) the performance of conversion function in non-parallel data condition. To cope with these challenges, the three corresponding sub-tasks have been defined Section 1.4. To deal with the first sub-task, a method to represent speaker individuality via speaker embedding and principal component analysis has been proposed in Chapter 3. Chapter 4 has focused on the second sub-task by applied the proposed framework in Chapter 3 into cross-lingual voice conversion domain using on the vector-quantized variational autoencoder. Finally, to achieve the last sub-task, Chapter 5 has proposed methods focused on enhancing the spectral features and prosody features to improve the quality of converted speech.

With the purpose to control the speaker individuality in the converted speech, Chapter 3 proposes the non-parallel text-independent voice conversion framework continuous speaker embedding. To represent the speaker individuality on a continuous plane, a speaker codebook is learned via backpropagation during the training process. After training, the speaker codebook is analysed using principal component analysis (PCA) to reveal the voice characteristics in the speaker embedding. It is observed that the first components of the PCA-projected speaker embedding mostly corresponds to the gender attribute of the speaker, while the second speaker embedding corresponds to the voice quality. Experiment results show that the proposed model performs better than the baseline parallel GMM-based voice conversion model with the same amount of target data. Moreover, the converted speech with fictitious voice identity has comparable naturalness as those with real target voice identity, with the average MOS score of 3.1. These results show that the proposed model is capable of generating new speaker individuality without any sacrifice in performance. With these results, the first sub-task on controlling the speaker individuality for non-parallel voice conversion has been achieved.

Chapter 4 focuses on the cross-lingual voice conversion using the same framework proposed in Chapter 3. The proposed cross-lingual voice conversion system includes the neural vocoder, Parallel WaveGAN [77], to improve the quality of synthesized speech waveform. This chapter proposes the hierarchical structure and discrete latent representation to better capture various linguistic information at different temporal scale. Adaptation to cross-lingual speaker is performed using backpropagation during the fine-tuning process, similar to the procedure in Chapter 3. The experiment results indicate that the proposed hierarchical structure leads to the significant improvement in speech naturalness and speaker similarity in both intra-lingual and cross-lingual situations. Official results in the Voice Conversion Challenge 2020 reveal that the proposed model ranks highest among text-independent non-parallel approaches in term of speech naturalness, with the average naturalness score of 3.2 MOS. These results show that the proposed framework can be well applied for both intra-lingual and cross-lingual tasks, satisfying the second sub

To achieve the last sub-task, Chapter 5 concentrates on methods for improving the performance of cross-lingual voice conversion. The proposed methods include the adversarial training scheme to enhance the spectral detail, the F_0-injection method to improve the pitch modeling, and a language-independent speaker embedding to avoid unnatural pronunciation. Conventional training loss function of the voice conversion model consists of the mean-square-error loss for reconstruction of speech features. However, such type of loss function causes the over-smoothing problem, which affect the naturalness of converted speech. The adversarial training scheme mitigates this problem by indirectly estimate the likelihood of the data via an additional discriminator network. The experiment result show that the generated mel-spectrogram with adversarial training scheme has finer detail structure than the conventional training scheme, hence, better naturalness is obtained. With the use of neural vocoder to synthesize speech waveform, voice conversion model operates on mel-spectrogram often suffers from the pitch-instability in the converted speech. To alleviate this problem, the F_0-injection approach is applied by directly conditioned the generation of mel-spectrogram on input $\log_2 F_0$. At run-time conversion phase, the conditioning

$\log_2 F_0$ is the linearly scaled version of source $\log_2 F_0$ to match the target distribution. Objective results show that the converted speech with F_0-injection method has more similar F_0 distribution and lower F_0 error to the target F_0. By visualizing the speaker embedding via principal component analysis, the language-independent property of the proposed speaker embedding is confirmed. To evaluate the effectiveness of the language-independent speaker embedding, 11 converted utterances are generated from the speaker embedding linearly interpolated between source and target ones. The listening test results indicate a strong correlation between the average similarity score of the fictitious voice and the expected similarity scores. This result indicate the effectiveness in controlling the speaker individuality using the proposed speaker embedding.

With the above results, a semi-supervised non-parallel voice conversion with speaker individuality control has been achieved. The proposed model can both perform voice mimicry and speaker individuality control with good speech naturalness in intra- and cross-lingual task. However, the quality of speaker similarity in cross-lingual conversion still need to be improved in the future.

6.2 Future works

The purpose of this thesis is to develop a flexible voice conversion model that can both passively mimicking voice and actively generate new speaker individuality using only acoustic data. Several techniques to control the speaker individuality in the non-parallel and cross-lingual situations have been developed. However, the current study mainly focuses on the spectral features conversion. To improve the performance of voice conversion, especially in the cross-lingual situation, it is necessary to include both the prosody features and spectral features for conversion.

6.2.1 Prosody conversion

In Chapter 5, the proposed F_0-injection method give some sort of controlling the pitch in converted speech. During conversion, the F_0 is simply linearly scale to match the distribution of target F_0. To transform the prosody of speech, it is needed to considered the characteristics of the F_0 contour. One potential solution is to parameterized the F_0 contour using Fujisaki model [94], which introduces the critical damping model to model the components of F_0 contour. The Fujisaki model describes the F_0 contour by three components: a asymptotic baseline F_b, a slowly varying phrase component and a local accent component.

$$
\begin{aligned}
\ln F_0(t) =& \ln F_b + \sum_{i=1}^{I} A_{p_i} G_p(t - T_{0_i}) \\
&+ \sum_{j=1}^{J} A_{a_j} \left\{ G_a(t - T_{1_j}) - G_a(t - T_{2_j}) \right\}, \\
G_p(t) =& \begin{cases} \alpha^2 t e^{-\alpha t}, & t \geq 0, \\ 0, & t < 0, \end{cases} \\
G_a(t) =& \begin{cases} \min\left\{1 - (1 + \beta t) e^{-\beta t}, \gamma\right\}, & t \geq 0, \\ 0, & t < 0, \end{cases}
\end{aligned}
\tag{6.1}
$$

The phrase components are the responses of the second-order critical-damping system described by the impulse response $G_p(t)$, whereas each phrase component is characterised by the impulse magnitude A_{p_i}. The accent components are the responses of the second order critical-damping model described by the impulse response $G_a(t)$, and each accent component is characterized by the impulse magnitude A_{a_i}. Previous study [58] has shown that the speaker speaker individuality in F_0 contour mostly corresponds to three parameters F_b, A_{a_i}, and A_{p_i}. If the voice conversion can extract and modify these parameters, the speaker individuality conveyed in the speech prosody can therefore be transformed. The proposed framework in this thesis can be extended with prosody modeling to boost the performance of voice conversion in both intra- and cross-lingual tasks.

6.2.2 Urgency voice conversion

Urgency voice conversion is the task of adding the sense of urgency in the neutral speech so that the listener can perceive the certain degree of danger. This voice conversion is very useful in public announcement system, especially in the event of disaster. The degree of urgency should be controlled depends on the imminence of danger. For example, warning for tsunami or fire should be at the highest level of urgency, whereas the warning for heavy rain should at lower level of urgency. This type of voice conversion can be regarded as one application of expressive voice conversion. However, most studies on expressive voice conversion focus on the categorical emotion, hence, controlling the degree of expressiveness is not possible. In the pioneering study of Xue et al. [95], an expressive voice conversion with controllable degree of expressiveness has been proposed. This study based on the three-layer model to describe the relation between emotions and acoustic features via semantic primitives [96]. To control the degree of emotion, a two-dimensional space (Valence-Activation) is used to represent the emotion on a continuous scale. An adaptive neuro-fuzzy inference system (ANFIS) is used to estimate the acoustic features from the valence-activation value. It is noticed that the idea of controlling the degree of emotion is parallel with the proposed voice conversion system in this thesis. Moreover, deep learning approaches have not become very popular in expressive voice conversion with controllable degree of emotion. Therefore, the approach in this thesis can be extended for urgency voice conversion to improve the flexibility of conversion model and enhance the quality of converted speech

References

[1] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Emotional speech recognition and synthesis in multiple languages toward affective speech-to-speech translation system," in *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE, 2014, pp. 574–577.

[2] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z.-H. Ling, and T. Toda, "Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 80–98. [Online]. Available: http://dx.doi.org/10.21437/VCC_BC.2020-14

[3] D. Suendermann-Oeft, H. Höge, A. Bonafonte, H. Ney, A. Black, and S. S. Narayanan, "Text-independent voice conversion based on unit selection," *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, pp. I–I, 2006.

[4] D. Erro, A. Moreno, and A. Bonafonte, "Inca algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 944–953, 2010.

[5] J.-X. Zhang, L.-J. Liu, Y.-N. Chen, Y.-J. Hu, Y. Jiang, Z.-H. Ling, and L.-R. Dai, "Voice Conversion by Cascading Automatic Speech Recognition and Text-to-Speech Synthesis with Prosody Transfer," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 121–125. [Online]. Available: http://dx.doi.org/10.21437/VCC_BC.2020-16

[6] W.-C. Huang, T. Hayashi, S. Watanabe, and T. Toda, "The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 160–164. [Online]. Available: http://dx.doi.org/10.21437/VCC_BC.2020-24

[7] W.-C. Huang, P. L. Tobing, Y.-C. Wu, K. Kobayashi, and T. Toda, "The NU Voice Conversion System for the Voice Conversion Challenge

2020: On the Effectiveness of Sequence-to-sequence Models and Autoregressive Neural Vocoders," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 165–169. [Online]. Available: http://dx.doi.org/10.21437/VCC_BC.2020-25

[8] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *INTERSPEECH*, 2017.

[9] ——, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016.

[10] T. V. Ho and M. Akagi, "Non-parallel voice conversion based on hierarchical latent embedding vector quantized variational autoencoder," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 140–144. [Online]. Available: http://dx.doi.org/10.21437/VCC_BC.2020-20

[11] D. Sundermann, H. Ney, and H. Hoge, "Vtln-based cross-language voice conversion," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 676–681, 2003.

[12] Y. Qian, J. Xu, and F. Soong, "A frame mapping based hmm approach to cross-lingual voice transformation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5120–5123, 2011.

[13] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6790–6794.

[14] B. Sisman, M. Zhang, M. Dong, and H. Li, "On the study of generative adversarial networks for cross-lingual voice conversion," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 144–151.

[15] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," in *ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*. ISCA, 2020.

[16] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on gaussian mixture model," in *INTERSPEECH*, 2006.

[17] ——, "One-to-many and many-to-one voice conversion based on eigenvoices," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4. IEEE, 2007, pp. IV–1249.

[18] Y. Stylianou, O. Cappé, and É. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 131–142, 1998.

[19] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, vol. 1, pp. 285–288 vol.1, 1998.

[20] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, pp. 2222–2235, 2007.

[21] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2100–2104, 2018.

[22] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[23] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion," in *Proc. Interspeech 2019*, 2019, pp. 679–683. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2236

[24] Y. Choi, M.-J. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8789–8797, 2018.

[25] T. V. Ho and M. Akagi, "Non-parallel voice conversion with controllable speaker individuality using variational autoencoder," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 106–111.

[26] ——, “A unified framework for non-parallel voice conversion and voice synthesis using variational autoencoder,” in *Proceedings of the Autumn Meeting of the Acoustical Society of Japan*, 2019.

[27] ——, “Non-parallel voice conversion with controllable speaker individuality using variational autoencoder,” in *Proceedings of the Acoustic Symposium*, 2019.

[28] ——, “Cross-lingual voice conversion with multi-codebook hierarchical vector-quantized variational autoencoder,” in *Proceedings of the Autumn Meeting of the Acoustical Society of Japan*, 2020.

[29] ——, “Improving spectral detail and f0 modelling for vae-based cross-lingual voice conversion with adversarial training,” in *Proceedings of the Spring Meeting of the Acoustical Society of Japan*, 2021.

[30] ——, “Cross-lingual voice conversion with controllable speaker individuality using variational autoencoder and star generative adversarial network,” *IEEE Access*, vol. 9, pp. 47 503–47 515, 2021.

[31] V. Dellwo, M. Huckvale, and M. Ashby, *How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 1–20. [Online]. Available: https://doi.org/10.1007/978-3-540-74200-5_1

[32] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations.* de Gruyter, 1971.

[33] T. Kitamura and M. Akagi, *Speaker Individualities in Speech Spectral Envelopes and Fundamental Frequency Contours*, C. Müller, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[34] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, “Restructuring speech representations using a 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.

[35] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, “The hmm-based speech synthesis system (hts) version 2.0,” in *The 6th ISCA Workshop on Speech Synthesis*, 2007.

[36] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 202–207. [Online]. Available: http://dx.doi.org/10.21437/SSW.2016-33

[37] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 841–844.

[38] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion based on gmm and straight," in *INTERSPEECH*, 2001.

[39] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on gmm with straight mixed excitation," in *INTERSPEECH*, 2006.

[40] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016.

[41] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[42] K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," in *1991 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1991, pp. 594–597 vol.1.

[43] A. B. Kain, "High resolution voice transformation," *PhD Thesis*, 2001.

[44] Y. Stylianou, O. Cappé, and É. Moulines, "Statistical methods for voice quality transformation," in *EUROSPEECH*, 1995.

[45] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 952–963, 2006.

[46] C.-H. Wu, C.-C. Hsia, T.-H. Liu, and J.-F. Wang, "Voice conversion using duration-embedded bi-hmms for expressive speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1109–1116, 2006.

[47] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[48] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME).* IEEE, 2016, pp. 1–6.

[49] X. Tian, J. Wang, H. Xu, E. S. Chng, and H. Li, "Average modeling approach to voice conversion with non-parallel data." in *Odyssey*, vol. 2018, 2018, pp. 227–232.

[50] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "Wavenet vocoder with limited training data for voice conversion." in *Interspeech*, 2018, pp. 1983–1987.

[51] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," 2019.

[52] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 5, pp. 816–824.

[53] T. Toda and P. Dymarski, "Modeling of speech parameter sequence considering global variance for hmm-based speech synthesis," in *Hidden Markov Models, Theory and Applications.* InTech, 2011, pp. 131–150.

[54] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in hmm-based speech synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 290–294.

[55] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[56] O. Turk and L. M. Arslan, "Voice conversion methods for vocal tract and pitch contour modification," in *Proceedings of Eurospeech*, 2003.

[57] J. M. Gutierrez-Arriola, Y.-S. Hsiao, J. M. Montero, J. M. Pardo, and D. G. Childers, "Voice conversion based on parameter transformation," in *Fifth International Conference on Spoken Language Processing*, 1998.

[58] M. Akagi and T. Ienaga, "Speaker individuality in fundamental frequency contours and its control," *J. Acoust. Soc. Jpn. (E)*, vol. 18, no. 4, pp. 73–80, 1997.

[59] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[60] M. Rolinek, D. Zietlow, and G. Martius, "Variational autoencoders pursue pca directions (by accident)," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 398–12 407, 2019.

[61] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modified postfilter to recover modulation spectrum for hmm-based speech synthesis," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014, pp. 547–551.

[62] C. Veaux, J. Yamagishi, and K. Macdonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017.

[63] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99-D, pp. 1877–1884, 2016.

[64] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "Wavenet vocoder with limited training data for voice conversion," in *INTERSPEECH*, 2018, pp. 1983–1987. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1190

[65] W.-C. Huang, T. Hayashi, S. Watanabe, and T. Toda, "The sequence-to-sequence baseline for the voice conversion challenge 2020: Cascading asr and tts," in *ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*. ISCA, 2020.

[66] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," 2016, pp. 1–6.

[67] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z.-H. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Odyssey*, 2018.

[68] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100–2104.

[69] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *Proc. Interspeech*, 2017.

[70] T. Dinh, A. Kain, and K. Tjaden, "Using a manifold vocoder for spectral voice and style conversion." in *INTERSPEECH*, 2019, pp. 1388–1392.

[71] M. Rolinek, D. Zietlow, and G. Martius, "Variational autoencoders pursue pca directions (by accident)," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 406–12 415.

[72] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[73] X. Wang, S. Takaki, J. Yamagishi, S. King, and K. Tokuda, "A vector quantized variational autoencoder (vq-vae) autoregressive neural f_0 model for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 157–170, 2020.

[74] B. van Niekerk, L. Nortje, and H. Kamper, "Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge," *arXiv preprint arXiv:2005.09409*, 2020.

[75] S. Ding and R. Gutierrez-Osuna, "Group latent embedding for vector quantized variational autoencoder in non-parallel voice conversion." in *INTERSPEECH*, 2019, pp. 724–728.

[76] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 866–14 876.

[77] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[78] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[79] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 755–767, 2016.

[80] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," *ArXiv*, vol. abs/1912.06813, 2019.

[81] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-Parallel Voice Conversion with Cyclic Variational Autoencoder," in *INTERSPEECH*, 2019, pp. 674–678.

[82] J. chieh Chou and H.-Y. Lee, "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," in *Proc. Interspeech 2019*, 2019, pp. 664–668. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2663

[83] H. Kido and H. Kasuya, "Representation of voice quality features associated with talker individuality," in *Fifth International Conference on Spoken Language Processing*, 1998.

[84] J.-X. Zhang, Z. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 540–552, 2020.

[85] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5274–5278, 2018.

[86] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27, 2014, pp. 2672–2680.

[87] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. ICML'17, vol. 70, 2017, p. 214^^e2^^80^^93223.

[88] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617–3621, 2019.

[89] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel wavenet: Fast high-fidelity speech synthesis," pp. 3918–3926, 2018.

[90] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "Jvs corpus: free japanese multi-speaker voice corpus," in *Information Processing Society of Japan Research Report (2019-SLP-129)*, vol. 4, 2019, pp. 1–4.

[91] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *International Conference on Learning Representations (ICLR)*, 2016.

[92] M. Swain and D. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 2004.

[93] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin," *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, vol. 4, pp. 410–414, 2007.

[94] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," in *Speech Prosody 2004, International Conference*, 2004.

[95] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Communication*, vol. 102, pp. 54–67, 2018.

[96] R. Elbarougy and M. Akagi, "Improving speech emotion dimensions estimation using a three-layer model of human perception," *Acoustical science and technology*, vol. 35, no. 2, pp. 86–98, 2014.

Publications

Journal Paper

[1] **T. V. Ho** and M. Akagi, "Cross-lingual voice conversion with controllable speaker individuality using variational autoencoder and star generative adversarial network," IEEE Access, vol. 9, pp. 47503-47515, 2021.

[2] **T. V. Ho** and M. Akagi, "Nonparallel dictionary-based voice conversion using variational autoencoder with modulation-spectrum-constrained training," Journal of Signal Processing, vol. 22, no. 4, pp. 189-192, 2018.

International Conference

[3] **T. V. Ho** and M. Akagi, "Non-parallel voice conversion based on hierarchical latent embedding vector quantized variational autoencoder," in Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge, 2020, pp. 140-144.

[4] T. V. Ngo, **T. V. Ho**, M. Unoki, R. Kubo, and M. Akagi, "Enhancement of Speech Intelligibility under noisy reverberant conditions based on modulation spectrum concept," in Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2020, pp. 753-758.

[5] **T. V. Ho** and M. Akagi, "Non-parallel voice conversion with controllable speaker individuality using variational autoencoder," in Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 106-111.

[6] **T. V. Ho** and M. Akagi, "Non-parallel Training Dictionary-based Voice Conversion with Variational Autoencoder," in Proc. of International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'18), Hawaii, USA, 2018, pp. 695-698.

Domestic Conference

[7] **T. V. Ho** and M. Akagi, "Improving spectral detail and F0 modelling for VAE-based cross-lingual voice conversion with adversarial training," in Proceedings of the Spring Meeting of the Acoustical Society of Japan, March 2021, Japan.

[8] **T. V. Ho** and M. Akagi, "Cross-lingual voice conversion with multi-codebook hierarchical vector-quantized variational autoencoder," in Proceedings of the Autumn Meeting of the Acoustical Society of Japan, September 2020, Japan.

[9] T. Hirai, **T. V. Ho**, and I. Setiawan, "A study on Japanese end-to-end speech synthesis using romanized input text with accent symbols," in Proceedings of the Spring Meeting of the Acoustical Society of Japan, March 2020, Japan.

[10] **T. V. Ho** and M. Akagi, "A Unified Framework for Non-parallel Voice Conversion and Voice Synthesis using Variational Autoencoder," in Proceedings of the Autumn Meeting of the Acoustical Society of Japan, September 2019, Japan.

[11] **T. V. Ho** and M. Akagi, "Non-parallel Voice Conversion with Controllable Speaker Individuality using Variational Autoencoder," in Proceedings of the Acoustic Symposium, 2019, Japan.

[12] **T. V. Ho** and M. Akagi, "Speech Accent and Gender Recognition using Dilated Convolution Neural Network with Skip and Residual Connection," in Proceedings of the Spring Meeting of the Acoustical Society of Japan, March 2019, Tokyo, Japan.

[13] **T. V. Ho** and M. Akagi, "Non-parallel voice conversion using Convolutional Variational Autoencoder," in Proceedings of the Autumn Meeting of the Acoustical Society of Japan, September 2018, Oita, Japan.