High-Quality and Flexible Voice Conversion Techniques based on Statistical Spectral and Waveform Modeling

Department of Media Science
Graduate School of Information Science
Nagoya University

Patrick Lumbantobing

Contents

Abstract

This thesis presents an in-depth study for high-quality and flexible voice conversion (VC) using statistical spectral and waveform modeling techniques. Using VC, voice characteristics of a source speaker can be transformed into that of a target speaker while preserving the linguistic contents. To develop a VC, representations of vocal-tract resonance characteristics of the source speaker, such as vocal-tract spectrum parameterizations, have to be converted into that of the target speaker. Along with the source-excitation (pitch) characteristics, the converted speech waveform is generated using the transformed speech parameterizations. A high-quality and flexible VC can be beneficial for many speech applications in research and in daily life. In this work, the high-quality aspect is examined through the use of data-driven statistical waveform modeling, whereas the flexible aspect is examined through the use of data-driven statistical spectral modeling.

The statistical spectral modeling is developed to model the mapping function of the vocal-tract spectrum parameterizations, e.g., spectral envelope parameters, between the source and the target speakers. Another possible representation of resonance characteristics modeling is by the means of physical approach, such as through the use of articulatory (speech organs) configurations. The latter has the advantage of being more flexible in terms of direct control. Whereas, the former has the advantage on being more flexible in terms of system development, owing to the more straightforward methods to obtain spectral envelope parameters. On the other hand, the statistical waveform modeling is developed to model the speech waveform signal generation from speech parameterizations, such as spectral envelope and excitation parameters. Another way to generate speech waveform is to employ conventional rule-based approach for the generation procedure based on the source-filter theory of speech production (vocoder). The latter has the advantage of not requiring any model training, though the quality is limited. On the other hand, the former has the advantage of being able to produce natural sounding synthetic speech, thanks to the use of data-driven approach, such as with neural network model, i.e., neural vocoder. In this thesis, the high-quality aspect is emphasized by the usage of neural vocoder in VC, while the flexibility aspect is emphasized on the development of statistical model for spectral mapping.

To achieve the goal of this thesis, four main frameworks are elaborated, which correspond to the contribution for high-quality and flexible VC. First is the development of voice modification with articulatory manipulation, which enables flexible control of speech sounds by means of the intuitive representations of articulatory information. Second is the development of VC with neural-network (NN)-based spectral and waveform modeling, which uses spectral envelope parameters as vocal-tract representations. Third is the improvement of NN-based VC framework to achieve high-quality converted speech by performing fine-tuning of neural vocoder. Finally, in the fourth system, the flexibility in the development of VC is achieved by means of nonparallel spectral mapping model framework, which does not require parallel (paired) data between source and target speakers.

In order to develop a voice modification with articulatory manipulation system, Gaussian mixture model (GMM)-based statistical modeling is employed to perform both of the acoustic-to-articulatory (inversion) mapping and the articulatory-to-acoustic (production) mapping. A sequential mapping procedure between inversion and production mappings is developed to enable the manipulation of intermediate articulatory representations for performing speech modification. To yield higher quality of modified speech, the vocoder-based excitation generation is avoided through direct filtering of input speech waveform with the use of differential spectrum that is calculated between the input and the modified spectral parameters. The experimental results demonstrate that the system is capable of producing modified vowel sounds by manipulation of tongue positions, and the use of direct waveform modification yields significant quality improvements for varying modification of articulation efforts, i.e., hypo- and hyper-articulations.

In the second system, a VC framework is developed by means of NN-based modeling for spectral mapping. Further, an NN-based modeling is also used to directly model the speech waveform signal by conditioning on spectral and excitation parameters, such as a WaveNet vocoder. However, owing to the use of converted spectral parameters instead of natural parameters in the synthesis time, there exist mismatches between the spectral and the waveform modeling. To reduce these mismatches, a postprocessing method based on the direct waveform modification is used to obtain refined converted spectral envelope parameters. The experimental results demonstrate that the NN-based VC is capable of achieving higher quality and speaker similarity in cross-gender conversion compared to using conventional vocoder-based excitation generation and achieving higher speaker similarity in same-gender conversion compared to using direct waveform modification method.

In the third system, the NN-based VC is improved through the use of recurrent neural network (RNN)-based architecture for spectral mapping modeling and the fine-tuning of the WaveNet vocoder. This is because the use of postprocessing method does not directly address the mismatches within the WaveNet vocoder. However, it is not straightforward to achieve fine-tuning of WaveNet vocoder in VC due to the difference of temporal structure between source and target speakers. In other words, the converted spectral parameters from the source speaker cannot be used to fine-tune a WaveNet vocoder using the target speech waveform. To perform WaveNet fine-tuning in VC, a cyclic RNN structure is introduced, which can produce both of the converted source spectral parameters and the appraisal of estimated (oversmoothed) target spectral parameters. The oversmoothed target spectra obtained from the cyclic flow is used for WaveNet fine-tuning. The experimental results demonstrate that the CycleRNN-based spectral mapping model makes it possible to perform proper WaveNet fine-tuning in VC by significantly improving the quality and the speaker-similarity of the converted speech compared to the previous VC framework.

Finally, in the fourth system, the flexibility for developing voice conversion is achieved by means of nonparallel spectral modeling using variational autoencoder (VAE) framework. In the previous two systems, the statistical spectral models are developed with a parallel dataset between source and target speakers where they utter a same set of sentences. However, nonparallel speech datasets of source and target are more practical to be obtained, especially for speakers with different language. To achieve that, a VAE-based VC is employed, where the shared characteristics between speakers, such as phonetics, are to be captured within a latent space, and the speaker-dependent characteristics are to be determined by time-invariant speaker-coding features. However, owing to the inability to explicitly optimize converted spectral features (only reconstruction is considered), the performance of VAE-based VC is limited. To improve the VAE-based VC, a cyclic flow is introduced to recycle the converted spectra back into the system to obtain cyclic reconstructed spectra that can be directly optimized. The experimental results demonstrate that the CycleVAE-based VC gives significant improvements in quality and speaker-similarity of converted speech, especially for cross-gender conversions, as well as improvements in the disentanglement of speaker-independent and speaker-dependent traits from the latent space.

In summary, in this thesis, an investigation for achieving high-quality and flexible voice conversion system is conducted through the aforementioned four frameworks. The possibility of having flexible control is investigated through the development of articulatory controllable speech modification system, where speech sounds can be intuitively modified by manipulating the intermediate articulatory representations. To achieve VC in this way, in the future, it is necessary to model the physical vocal-tract shape with articulatory representations. Next, flexibility in the development of VC system is utilized through the use of NN-based spectral mapping model with parallel (paired) data, where spectral envelope parameters are more straightforward to be obtained compared to articulatory data. Then, the use of neural vocoder in VC is thoroughly investigated to achieve high-quality output, where the mismatches of spectral parameters are dealt with a spectral postprocessing method or with neural vocoder fine-tuning. Finally, to achieve flexible VC system for practical real-world applications, a nonparallel spectral modeling, which utilizes latent space to capture shared traits between different speakers, is presented. The last approach can be easily extended with the neural vocoder fine-tuning approach to ultimately achieve high-quality and flexible VC.

1 Introduction

1.1 Background

Speech is a signal produced by our speech organs (articulators) in such a way that when it is propagated through the ear organs and the brain of a listener, it would convey the corresponding intention of the speaker (speech producer) according to the phonetical and grammatical rules of the language and the comprehension level of the listener. It is therefore highly reasonable to take into account that speech is an essential component in daily life owing to its unavoidable use for communication.

Recently, a lot of progress has been made in the machine learning area for automatization and supports in daily life activities. This also includes many works on the advancements of technology for speech processing and automation. One such example, which is the main focus in this thesis, is a system that can flexibly modify the voice characteristics within the speech signal by harnessing machine learning techniques, i.e., voice conversion [1–3]. Using a voice conversion system, the voice characteristics of a source speaker can be transformed into that of a target speaker while preserving the linguistic contents of the speech. In practice, voice conversion can be used in a variety of speech applications, such as for singing voice conversion [4], recovery of impaired speech signal from handicapped people [5, 6], expressive speech synthesis [7, 8], and for body-conducted speech processing technology [9, 10]. Hence, it would be a fine contribution to work on the development of a dependable voice conversion framework.

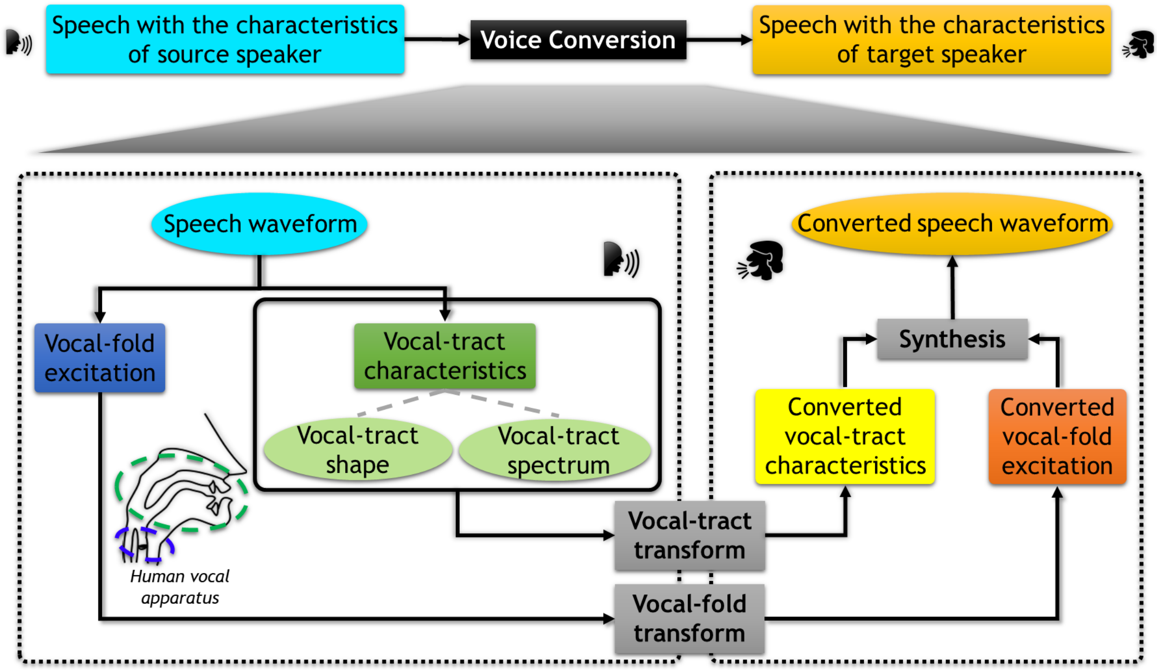

Speech with the characteristics of source speaker
Voice Conversion
Speech with the characteristics of target speaker
Speech waveform
Vocal-fold excitation
Vocal-tract characteristics
Vocal-tract shape
Vocal-tract spectrum
Human vocal apparatus
Vocal-tract transform
Vocal-fold transform
Converted speech waveform
Synthesis
Converted vocal-tract characteristics
Converted vocal-fold excitation

Figure 1.1: *General flow of a voice conversion system.*

In general, the flow of a voice conversion system can be depicted as in Fig. 1.1. The conversion of voice characteristics between source and target speaker can be achieved by transforming the vocal-tract resonance characteristics and the vocal-fold excitation. The converted speech can be generated from these transformed speech representations (parameters), such as with the use of a source-filter vocoder [11–13]. In this thesis, in order to develop a voice conversion framework, techniques for performing transformation of vocal-tract characteristics, such as for the mapping of vocal-tract spectrum, and techniques for synthesizing speech waveform signal are investigated.

In particular, in this work, the development of voice conversion system with the use of machine learning frameworks, such as data-driven statistical modeling techniques, is extensively studied. In a statistical voice conversion system, the conversion procedure between source and target speaker is performed with a statistical model that is trained

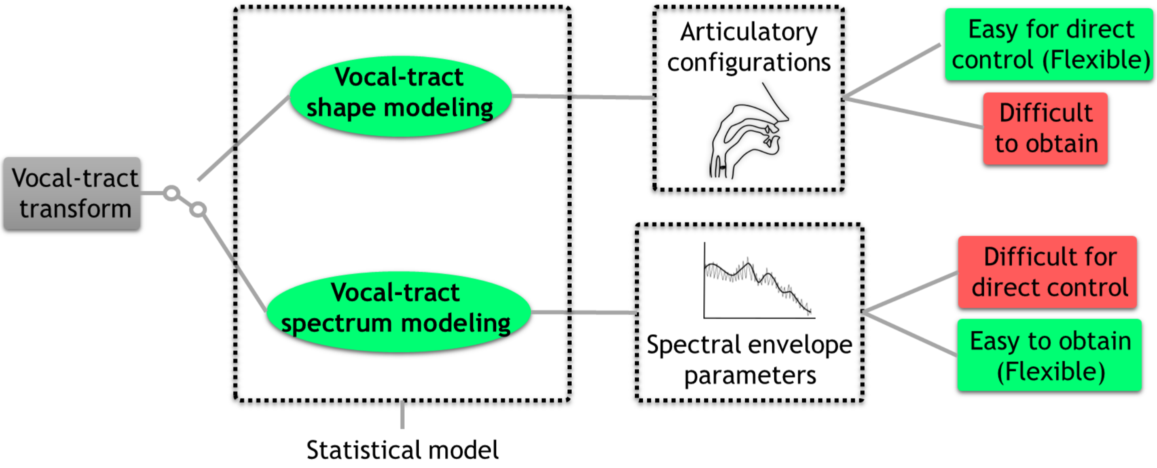

Vocal-tract transform
Vocal-tract shape modeling
Vocal-tract spectrum modeling
Statistical model
Articulatory configurations
Spectral envelope parameters
Easy for direct control (Flexible)
Difficult to obtain
Difficult for direct control
Easy to obtain (Flexible)

Figure 1.2: *Different approach to model vocal-tract transformation, i.e., through vocal-tract shape modeling (physical-based) or through vocal-tract spectrum modeling (extracted from speech signal).*

from available data, i.e., data-driven. This can include the creation of a statistical model for the mapping of the spectral parameters [3] and of a statistical model for synthesizing the speech waveform signal [14]. The use of statistical techniques for both of spectral and waveform modeling is performed with the goal in mind to achieve a high-quality and flexible voice conversion framework, which can be beneficial to speech applications in research and in daily life.

1.2 Issues to be considered in this Thesis

As illustrated in Figs. 1.1 and 1.2, the vocal-tract transformation can be performed by the means of vocal-tract shape modeling or vocal-tract spectrum modeling. In this thesis, the investigations are mainly done on the development of vocal-tract spectrum modeling, such as statistical model for mapping of spectral envelope parameters be-

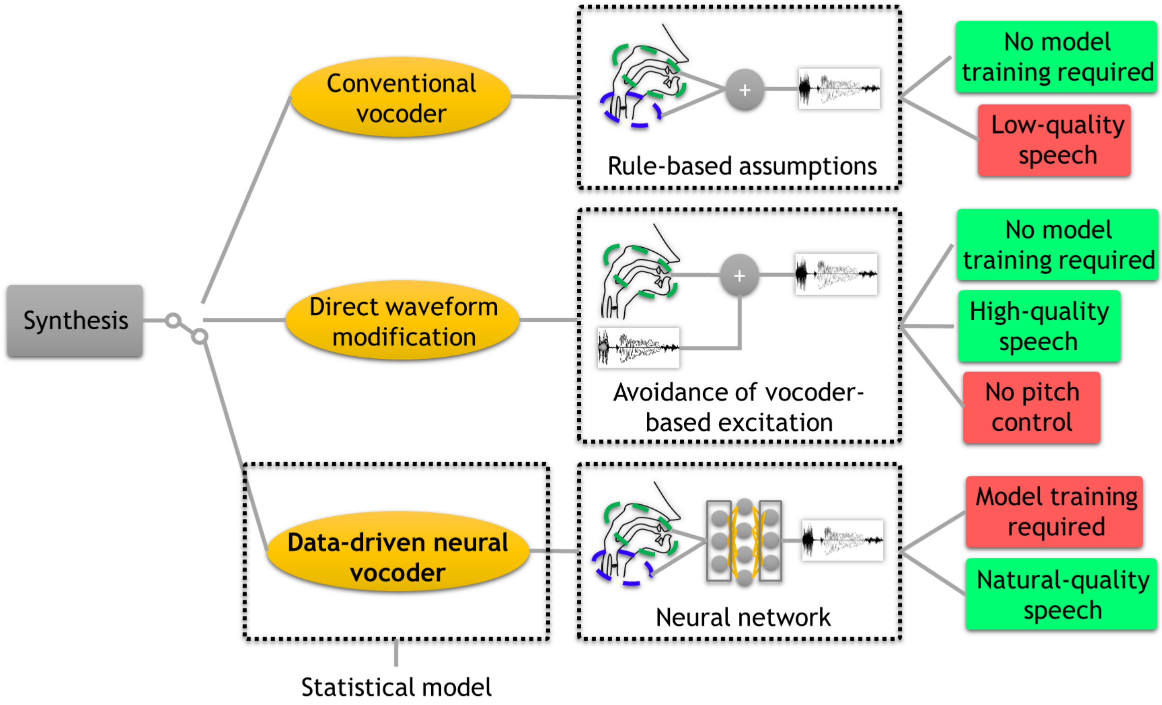


Figure 1.3: *Synthesis procedure with conventional vocoder, direct waveform modification, or data-driven neural vocoder.*

tween source and target speakers [3, 15]. On the other hand, another study on the use of articulatory (speech organs) configurations [16] is also conducted, which has a potential to be further developed for the vocal-tract shape modeling to achieve voice conversion. The latter approach has the advantage of being more flexible in terms of control, thanks to the intuitive representations of articulatory representations [17, 18], such as the positions of tongue or lips. However, it is not straightforward to obtain such articulatory data [19] for the development of statistical model. The flexibility in system development is achieved by the use of direct mapping of spectral parameterizations, such as spectral envelope parameters [20, 21], which can be obtained in a more straightforward manner [13, 22].

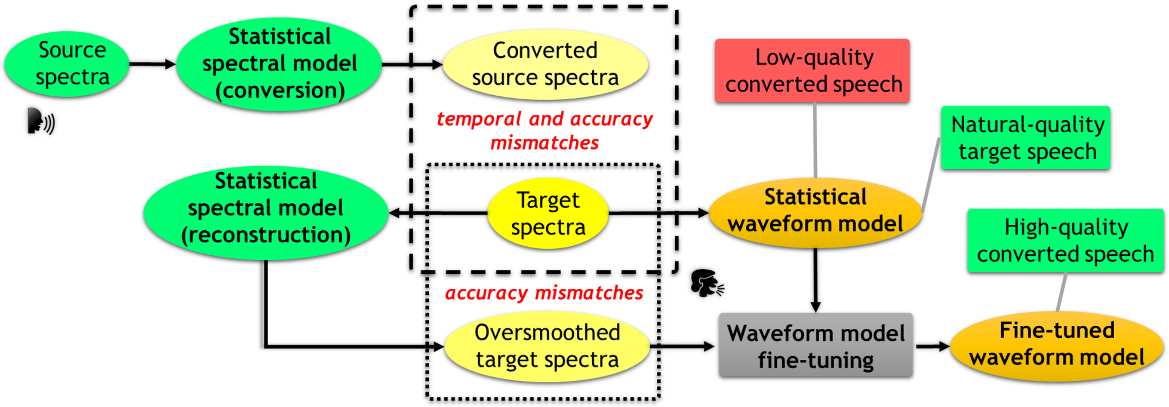

Source spectra
Statistical spectral model (conversion)
Converted source spectra
temporal and accuracy mismatches
Low-quality converted speech
Natural-quality target speech
Statistical spectral model (reconstruction)
Target spectra
Statistical waveform model
High-quality converted speech
accuracy mismatches
Oversmoothed target spectra
Waveform model fine-tuning
Fine-tuned waveform model

Figure 1.4: *Fine-tuning waveform model to achieve high-quality converted speech due to mismatches with natural speech features.*

Aside of the spectral modeling, another important aspect that needs to be taken into account is the synthesis module used in the voice conversion system. As has been briefly mentioned and illustrated in Fig. 1.3, the usual way is to use the conventional vocoder [11, 13, 22], which utilize the source-filter theory [11, 12] of speech production with predefined assumptions on the generation procedure using the vocal-fold excitation parameters (source) and the vocal-tract spectral parameters (filter). However, owing to those assumptions, the quality of the generated speech is limited compared to the natural speech. One possible way to alleviate the quality degradation is by avoiding the use of vocoder-based excitation generation through employing direct waveform modification technique [4], as depicted by the middle flow in Fig. 1.3. The use of direct waveform modification in VC has significantly improved the quality of converted speech, although its usage is limited for same-gender conversions, owing to the avoidance of source-excitation (pitch) parameterizations.

Another alternative way to achieve speech waveform generation is through the use of data-driven statistical model that directly models the speech waveform signal. As

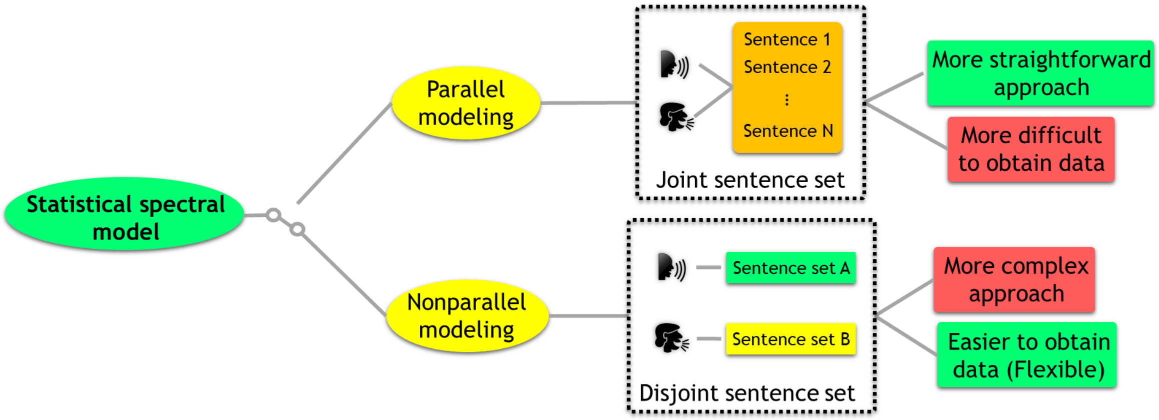


Figure 1.5: *Nonparallel spectral modeling for voice conversion, i.e., with different sentence set between speakers.*

illustrated in the bottom flow of Fig. 1.3, such model would also be conditioned on similar speech parameters, such as source-excitation and spectral. However, thanks to the data-driven approach, e.g., with neural network modeling [14, 23, 24], natural sounding synthetic speech can be generated. In this work, to achieve not only flexible, but also high-quality voice conversion, direct speech waveform modeling through the use of statistical neural network architecture, i.e., neural vocoder, is also studied. Specifically, owing to the mismatches that occur between the source spectral parameters converted by using the spectral modeling and the target spectral parameters, the quality of converted speech even if using neural vocoder will be degraded [25]. Hence, as depicted in Fig. 1.4, a fine-tuning approach of waveform model has to be done by considering the temporal and accuracy mismatches that might occur due to the use of statistical spectral model in VC [26].

Lastly, going back again to spectral mapping modeling, the use of nonparallel (unpaired) data between source and target speakers has to be considered as depicted in Fig. 1.5. This is because not all of the time paired data for parallel training procedure can be obtained. Therefore, to ultimately realize flexible VC development, nonparallel spectral modeling framework [27–29] is a must.

In summary, in this thesis, four main issues are investigated to achieve high-quality and flexible voice conversion system by using statistical techniques in spectral and waveform modelings. First is the issue of spectral modeling framework in general, where physical-based approach with intermediate articulatory representation can be used to provide flexibility in direct control, while a more straightforwad approach with spectral envelope features as parameterizations of vocal-tract spectrum is more flexible in terms of system development, owing to the flexibility of obtaining spectral envelope data compared to articulatory data. Second is the issue of speech generation procedure, where there are three possible framework to be chosen, i.e., conventional vocoder, direct waveform modification to avoid the vocoder-based excitation generation, and the statistical waveform model with neural network. Third is the issue of quality degradation in using statistical waveform modeling for VC and the possible approach in model fine-tuning procedure to achieve high-quality VC output while still allowing pitch conversion, such as for cross-gender cases. Finally, the fourth issue is the nonparallel spectral modeling problem to ultimately achieve flexible VC system development without the need of any paired (parallel) data between speakers. These issues are addressed by the proposed frameworks within the scope of this thesis, which are briefly overviewed in the next section.

1.3 Scope of this Thesis

This section gives the overview of the proposed frameworks within the scope of this thesis following the issues described in the previous section. Specifically, this thesis presents four frameworks in the investigation for achieving high-quality and flexible voice conversion technology. First is the statistical voice modification with articulatory mapping and manipulation, which makes it possible to perform flexible speech modification through manipulation of intuitive articulatory representations (flexible control). Second is the statistical voice conversion (VC) system based on neural network (NN) spectral and waveform modeling, which makes use of spectral envelope parameterizations that can be obtained in a more straightforward/flexible manner compared to articulatory representations (flexible system development and high-quality output). Third is the improvement of NN-based VC framework with fine-tuning of waveform modeling (neural vocoder) to achieve high-quality converted speech (higher-quality output). Fourth is the nonparallel spectral mapping model based on the use of variational autoencoder (VAE) framework to achieve flexible VC development with arbitrary/unpaired data between source and target speakers (more flexible system).

1.3.1 Statistical Voice Modification with Articulatory Mapping and Manipulation

In the first system, a statistical approach to use physical-based representations of vocal-tract characteristics is investigated. Specifically, this framework is capable of estimating articulatory representations from an input spectral envelope parameters (inverse mapping), and estimate modified spectral envelope parameters after performing manipulation of the articulatory representations (production mapping) [18]. Both of the aforementioned mappings are developed with Gaussian mixture model (GMM)-based technique [16], on which the intermediate articulatory representations are made to be available for intuitive manipulation of speech sounds. As depicted in Fig. 1.2, the use of articulatory information allows more flexible approach for direct control of speech signal, though, with the downside that it is not so straightforward to obtain articulatory data for the means of statistical model development. Nevertheless, the experimental results demonstrate that the system is capable of producing modified vowel sounds through manipulation of the tongue height positions. Further, it is also capable of producing high-quality modified speech sounds in varying articulation efforts, e.g., hypo- and hyper-articulations, by avoiding the use of vocoder-based assumptions of excitation generation, which limits the quality of synthetic speech from conventional vocoder, as depicted in Fig. 1.3. Future work from this topic includes the modeling of the vocal-tract shapes [30–34] to make it possible to perform voice conversion by transformation of the shape of the vocal-tract from the source into that of the target speaker.

1.3.2 Statistical Voice Conversion with Neural Network Spectral Mapping and WaveNet Vocoder

In the second system, a voice conversion framework based on neural network (NN) architecture is presented. Specifically, NN-based architectures are used for the statistical spectral mapping model [15] and for the statistical waveform model (WaveNet vocoder [14, 24]). In this case, as depicted in Fig. 1.2, the system development can be performed in a more flexible manner, owing to the use of spectral envelope parameters [20, 21] as representations of vocal-tract characteristics, which can be obtained in a more straightforward manner [13, 22] compared to articulatory data [19]. Further, the use of data-driven statistical waveform model, i.e., neural vocoder, specifically the WaveNet vocoder in this work, makes it possible to achieve higher-quality converted speech, as depicted in Fig. 1.3. Though, still, there exists mismatches between the spectral and the waveform model [25]. Hence, there is a need to use a postprocessing method for obtaining refined converted spectral envelope parameters to be used in generating the converted speech waveform. In the experimental evaluation, it has been demonstrated that the NN-based voice conversion framework is capable of achieving higher quality and speaker-similarity of converted speech in cross-gender conversions compared to conventional system with vocoder-based generation procedure and of achieving higher speaker-similarity in same-gender conversions compared to conventional system that avoids the use of vocoder-based excitation generation assumptions. Future work from this topic includes the handling of mismatches by performing fine-tuning of neural vocoder to achieve higher-quality output and the development of more flexible system with nonparallel (unpaired) data between source and target speakers.

1.3.3 Voice Conversion with Cyclic Recurrent Neural Network and Finely Tuned WaveNet Vocoder

In the third system, a framework to improve the NN-based voice conversion by performing fine-tuning of WaveNet vocoder is elaborated. In the development of WaveNet vocoder, the natural spectral parameters are used as conditioning features as depicted in Fig. 1.4. However, in the conversion phase, the converted spectra from source speaker is used to generate the converted speech, hence, the mismatches between spectral and waveform modeling occur, which degrade the converted speech quality. Thanks to the data-driven approach of WaveNet vocoder, it is possible to directly address these mismatches by fine-tuning a pretrained model with the estimated spectra of target speaker [35]. Though, as will be explained in more detail in the next chapter, it is not straightforward to obtain the estimated (oversmoothed) target spectra for fine-tuning, owing to the temporal differences between source and target speakers. This framework is developed to address these issues by utilization of cyclic recurrent neural network (CycleRNN) architecture for spectral mapping model [36], which can estimate both of the converted source-to-target spectra (from conversion flow) and the appraisal of oversmoothed target spectra (from cyclic flow). The experimental results demonstrate the effectiveness of the WaveNet fine-tuning using oversmoothed target spectra from cyclic flow in voice conversion with its significant improvements in both of quality and speaker-similarity of converted speech compared to the previous best voice conversion framework that utilizes postprocessing method for alleviating the mismatches issue between spectral and waveform models. Future work from this topic includes the use of neural vocoder fine-tuning in nonparallel spectral modeling system and the possibility of joint optimization between spectral and waveform models.

1.3.4 Non-Parallel Voice Conversion with Cyclic Variational Autoencoder

In the fourth system, a framework for the development of nonparallel spectral mapping modeling is presented. In previous spectral modeling systems, parallel speech dataset between source and target speakers is used, where they utter a same set of utterances, as depicted in Fig. 1.5. However, to enable a more flexible voice conversion development, it is necessary to employ the use of nonparallel spectral mapping model [28, 29], where the model optimization does not rely on the paired utterances between source and target speakers, e.g., speaking in different languages or different sets of utterances, which is more suitable for practical situations. In order to achieve that, in this framework, a nonparallel spectral mapping technique based on variational autoencoder (VAE) [37] is employed. As will be explained in more detail in the next chapter, a VAE-based VC [38] utilizes a latent feature space for capturing the shared representations between speakers, e.g., phonetics, while positioning the speaker-dependent characteristics with the use of speaker-coding features. The VAE-based VC is optimized by reconstruction losses of the spectral features and the regularization terms of the latent space. The performance of conventional VAE-based VC is significantly degraded due to the inability of using converted spectral features in model optimization. In this sytem, improvement is made by introducing cyclic flow (CycleVAE) [36], which makes it possible to recycle the converted spectral features back into the system, which can be optimized as the cyclic reconstructed spectra. The experimental results demonstrate that the CycleVAE-based VC significantly improves the quality and speaker-similarity of converted speech. Future work of this topic includes many-to-many VC, cross-language VC, and fine-tuning of neural vocoder for ultimate high-quality and flexible VC.

1.4 Thesis Overview

In summary, the relation of the aforementioned techniques with each chapter in this thesis to achieve high-quality and flexible voice conversion is given in Table 1.1. Hence, this thesis is organized as follows. In Chapter 2, related works on the speech and articulatory modeling, voice conversion, spectral mapping modeling and waveform modeling are presented. In Chapter 3, the framework for voice modification with articulatory manipulation is described to investigate the flexibility in direct control of speech sounds. Moreover, in Chapter 3, the use of direct waveform modification technique to avoid vocoder-based excitation is investigated for high-quality generation of modified speech. Chapter 4 presents the voice conversion with neural-network-based approach for spectral and waveform modeling, where spectral envelope parameters are used as they are more flexible to be obtained compared to articulatory data. Further, in the parallel VC system of Chapter 4, neural vocoder-based waveform generation is used along with the spectral postprocessing method to improve the quality of converted speech compared

Table 1.1: *A summary of relation of each chapter to the scope of aspect and techniques within this thesis. Wav-mod stands for direct waveform modification. Wav-gen stands for waveform generation technique (conventional or neural vocoder). FT wav-gen stands for fine-tuned neural vocoder.*

Aspect	**Technique**	**Chapter 3**	**Chapter 4**	**Chapter 5**	**Chapter 6**
Flexibility	**Control**	○			
	Parallel		○	○	
	Nonparallel				○
High quality	**Wav-mod**	○	○		
	Wav-gen	○	○	○	
	FT wav-gen			○	

to using conventional speech generation. In Chapter 5, an improvement of the voice conversion system that utilizes cyclic recurrent neural network (CycleRNN) to handle the mismatches between spectral mapping model and neural vocoder (WaveNet) is elaborated, where the WaveNet vocoder is fine-tuned by using the spectral features obtained from the CycleRNN to achieve high quality converted speech. In Chapter 6, a non-parallel voice conversion system based on cyclic variational autoencoder (CycleVAE) is given for realizing flexible VC system development without any paired data between speakers. Finally, in Chapter 7, the contributions of this thesis are summarized and future work is discussed.

2 Related work

2.1 General Overview

This section provides brief descriptions on several related works, which are utilized within this thesis. These include the following three topics: speech and articulatory mapping, spectral mapping modeling for voice conversion, and neural network (NN)-based vocoder (neural vocoder). The inclusion of these related works are conducted to achieve the goal of this thesis for high-quality and flexible voice conversion with statistical spectral and waveform modelings.

Specifically, the works on speech and articulatory mapping are utilized to make it possible in achieving flexible control of the speech signal with the use of intuitive representations of articulatory parameters. On the other hand, the works on spectral mapping modeling for voice conversion are utilized to make it possible in transforming the voice characteristics of a source speaker into that of a target speaker with the use of spectral envelope parameters that can be obtained in a more straightforward/flexible manner compared to articulatory data. Finally, the work on neural vocoder is deployed to definitely achieve high-quality converted speech output, thanks to the data-driven approach of NN-based architecture, which has significant advantage for improvements compared to conventional rule-based vocoder speech generation. Several issues that might arise in their use within this thesis, and possible solutions that are related to each corresponding topic are also briefly presented within this chapter.

2.2 Speech and Articulatory Mapping

This section briefly describes related statistical modeling techniques for the acoustic-to-articulatory (inversion) mapping and the articulatory-to-acoustic (production) mapping. In this thesis, the usage of articulatory information makes it possible to perform speech modification through manipulation of intermediate articulatory parameters within a sequential inversion and production mapping flows [18], which will be described in more detail within the next chapter.

2.2.1 Inversion mapping problem

As depicted in the top diagram of Fig. 2.1, the acoustic-to-articulatory (inversion) mapping is conducted to estimate articulatory representations, such as the positions of tongue and lips, from an input speech parameterizations, such as spectral envelope parameters extracted from speech signal. To realize the inversion mapping, several fundamental approaches based on mathematical functions (rule-based) exist [32, 34, 39, 40]. However, a vast number of approximations need to be considered to do so, which makes the inverse mapping solution is not straightforward to be implemented in various situations. Recent works based on statistical data-driven methods [41–49] for the inversion mapping problem have brought significant improvements in terms of its accuracy and its flexibility.

This thesis focuses on the use of Gaussian mixture model (GMM)-based technique [47] for modeling the mapping function from spectral characteristics of the speech signal onto the articulatory representations. In addition, in this thesis, additional technique is necessary for manipulating the articulatory representations estimated from the speech signal. This is because the changing of one articulatory configuration would naturally

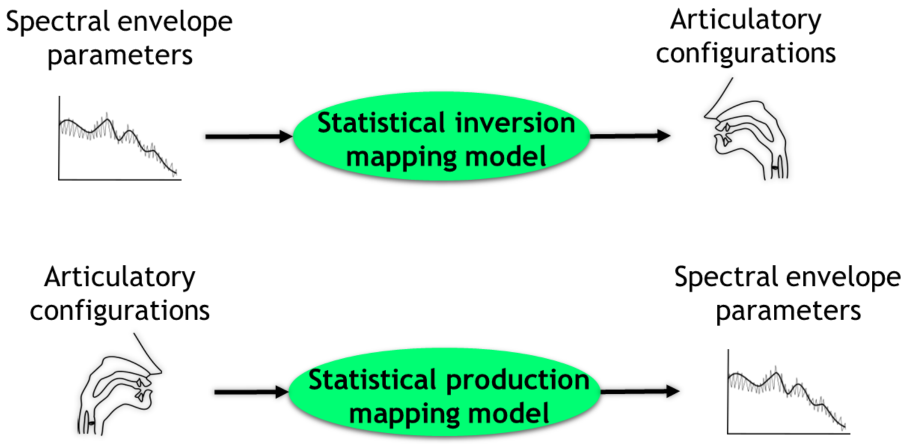

Spectral envelope parameters
Statistical inversion mapping model
Articulatory configurations
Articulatory configurations
Statistical production mapping model
Spectral envelope parameters

Figure 2.1: *Statistical inversion mapping model for acoustic-to-articulatory flow, and statistical production mapping model for articulatory-to-acoustic flow.*

affects the configurations of another. The technique to consider interdimensional correlation of articulatory parameters by utilizing statistical modeling of inversion mapping in the manipulation procedure [18] will be described within the Chapter 3.

2.2.2 Production mapping problem

On the other hand, conversely to that of the inversion mapping, as depicted in the bottom diagram of Fig. 2.1, the articulatory-to-acoustic production mapping is conducted to estimate spectral envelope parameters from input articulatory information. The difficulty in the conventional inversion mapping procedures is also faced with the techniques for realizing the forward/production mapping from vocal tract configurations to speech signal [30, 50]. In a similar way, due to the extensive needs of approximations within such methods, in this thesis we will focus on the use of statistical data-driven technique for production mapping, which have been proven to produce a reliable performance [51–55]. Specifically, the GMM-based articulatory-to-acoustic production mapping is used within this thesis [55].

In practice, to harness both of the inversion and production mappings, in this thesis, a sequential inversion and production mappings flow is adopted. This would make it possible to perform speech modification by manipulating the intermediate articulatory representations. Moreover, in order to be able to produce high-quality modified speech output, a technique that avoids the use of vocoder-based excitation generation assumptions [18], in a similar way as within the next section, is deployed. The detail of its implementation within the production mapping will be described within the Chapter 3.

2.2.3 GMM-based Statistical Inversion and Production Mapping Methods

Let $\boldsymbol{c}_t$, $\boldsymbol{s}_t$, and $\boldsymbol{x}_t$ be the spectral envelope parameters, i.e., mel-cepstral coefficients; the source excitation parameters, i.e., log-scaled F_0 and log-scaled waveform power; and the articulatory parameters at frame t, respectively. The time sequence vectors of these parameters over an utterance are respectively defined as $\boldsymbol{c} = [\boldsymbol{c}_1^\top, \dots, \boldsymbol{c}_T^\top]^\top$, $\boldsymbol{s} = [\boldsymbol{s}_1^\top, \dots, \boldsymbol{s}_T^\top]^\top$, and $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_T^\top]^\top$, where T denotes the number of frames and $\top$ denotes the transposition of the vector. Note that, the procedure of each of the inversion and production mappings is deliberately described because of their differences in the employment of respective source and target feature vectors. Mainly, wider contextual frames are needed in the inversion mapping, while the use of source excitation features is helpful in the production mapping [16].

GMM-based Acoustic-to-Articulatory Inversion Mapping

In the inversion mapping, spectral envelope parameters of an input speech signal are converted into their corresponding articulatory parameters. As the source feature, a mel-cepstral segment feature vector $\boldsymbol{O}_t$ is used at frame t, which is extracted from the mel-cepstral parameters $\boldsymbol{c}_t$ at multiple frames around the current frame t, as given by

$$\boldsymbol{O}_t = \boldsymbol{A}[\boldsymbol{c}_{t-L}^\top, \ldots, \boldsymbol{c}_t^\top, \ldots, \boldsymbol{c}_{t+L}^\top]^\top + \boldsymbol{b}, \tag{2.1}$$

where $\boldsymbol{A}$ and $\boldsymbol{b}$ denote the parameters for the linear transformation, which are calculated beforehand by principal component analysis using the training data. As the target feature, a joint static and dynamic feature vector of articulatory parameters, given by $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$, is used at frame t, where $\Delta\boldsymbol{x}_t$ is the dynamic feature vector of the articulatory parameters.

In the training procedure of inversion mapping, a joint source and target feature vector $[\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top$ is developed at each frame t from all utterances in the training data. The joint probability density function of the source and target features is then modeled with a GMM of the inversion mapping as follows:

$$P(\boldsymbol{O}_t, \boldsymbol{X}_t | \boldsymbol{\lambda}^{(O,X)}) = \sum_{m=1}^{M} \alpha_m^{(O,X)} \mathcal{N}([\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top; \boldsymbol{\mu}_m^{(O,X)}, \boldsymbol{\Sigma}_m^{(O,X)}), \tag{2.2}$$

where $\mathcal{N}(; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The parameter set for the GMM of the inversion mapping is denoted as $\boldsymbol{\lambda}^{(O,X)}$, which consists of weights $\alpha_m^{(O,X)}$, the mean vector $\boldsymbol{\mu}_m^{(O,X)}$, and the covariance matrix $\boldsymbol{\Sigma}_m^{(O,X)}$ of individual mixture components. The mixture component index is m and the total number of mixture components is M [1]. These model parameters are trained with the expectation-maximization (EM) algorithm. The training scheme for the GMM of the inversion mapping is shown in the upper diagram of Fig. 2.2.

In the conversion procedure, given a time sequence of mel-cepstral segment feature vectors $\boldsymbol{O} = [\boldsymbol{O}_1^\top, \ldots, \boldsymbol{O}_T^\top]^\top$, a time sequence of articulatory feature vectors $\boldsymbol{x}$ is estimated by employing a conditional probability density function, which is analytically derived from the GMM of the inversion mapping given in Eq. (2.2). In this work, an approximation of the conditional probability density function is employed with the use of a single mixture component sequence $\boldsymbol{m} = \{m_1, \ldots, m_T\}$ [3], where m_t denotes the mixture component index at frame t. First, a suboptimum mixture component sequence $\hat{\boldsymbol{m}}^{(O)}$ is determined as

$$\hat{\boldsymbol{m}}^{(O)} = \underset{\boldsymbol{m}}{\operatorname{argmax}}\, P(\boldsymbol{m}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}). \tag{2.3}$$

Then, a time sequence of converted articulatory feature vectors $\hat{\boldsymbol{x}}$ is determined as follows:

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, P(\boldsymbol{X}|\boldsymbol{O}, \hat{\boldsymbol{m}}^{(O)}, \boldsymbol{\lambda}^{(O,X)}), \text{ subject to } \boldsymbol{X} = \boldsymbol{W}^{(x)}\boldsymbol{x}, \tag{2.4}$$

where $\boldsymbol{W}^{(\boldsymbol{x})}$ is the linear transformation matrix used to calculate the sequence of joint static and dynamic articulatory feature vectors $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top]^\top$ from a sequence of articulatory feature vectors $\boldsymbol{x}$. The conversion scheme using the GMM of the inversion mapping is shown in the upper diagram of Fig. 2.3.

GMM-based Articulatory-to-Acoustic Production Mapping

In the production mapping, articulatory parameters together with source excitation parameters are converted into their corresponding spectral envelope parameters. As the source feature, a joint static and dynamic feature vector of articulatory and source excitation parameters, given by $\boldsymbol{Y}_t = [\boldsymbol{x}_t^\top, \boldsymbol{s}_t^\top, \Delta\boldsymbol{x}_t^\top, \Delta\boldsymbol{s}_t^\top]^\top$, is used at frame t, where $\Delta\boldsymbol{s}_t$ is the dynamic feature vector of the source excitation parameters. As the target feature, a joint static and dynamic feature vector of mel-cepstral parameters, given by

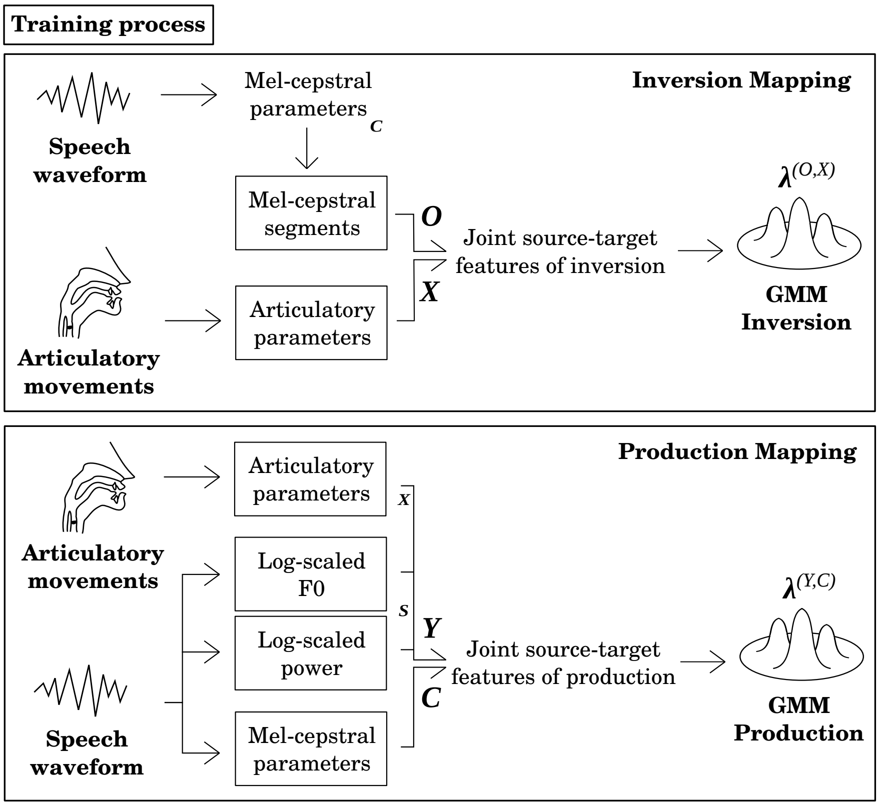


Figure 2.2: *Training scheme for GMM of the inversion mapping (top) and the production mapping (bottom).*

$\boldsymbol{C}_t = [\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top]^\top$, is used at frame t, where $\Delta\boldsymbol{c}_t$ is the dynamic feature vector of the mel-cepstral parameters.

In this training procedure, a joint source and target feature vector $[\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top]^\top$ is developed at each frame t. A joint probability density function is then modeled with a GMM of the production mapping as follows:

$$P(\boldsymbol{Y}_t, \boldsymbol{C}_t|\boldsymbol{\lambda}^{(Y,C)}) = \sum_{m=1}^{M} \alpha_m^{(Y,C)} \mathcal{N}([\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top]^\top; \boldsymbol{\mu}_m^{(Y,C)}, \boldsymbol{\Sigma}_m^{(Y,C)}), \tag{2.5}$$

where the parameter set for the GMM of the production mapping is denoted as $\boldsymbol{\lambda}^{(Y,C)}$, which consists of weights $\alpha_m^{(Y,C)}$, the mean vector $\boldsymbol{\mu}_m^{(Y,C)}$, and the covariance matrix

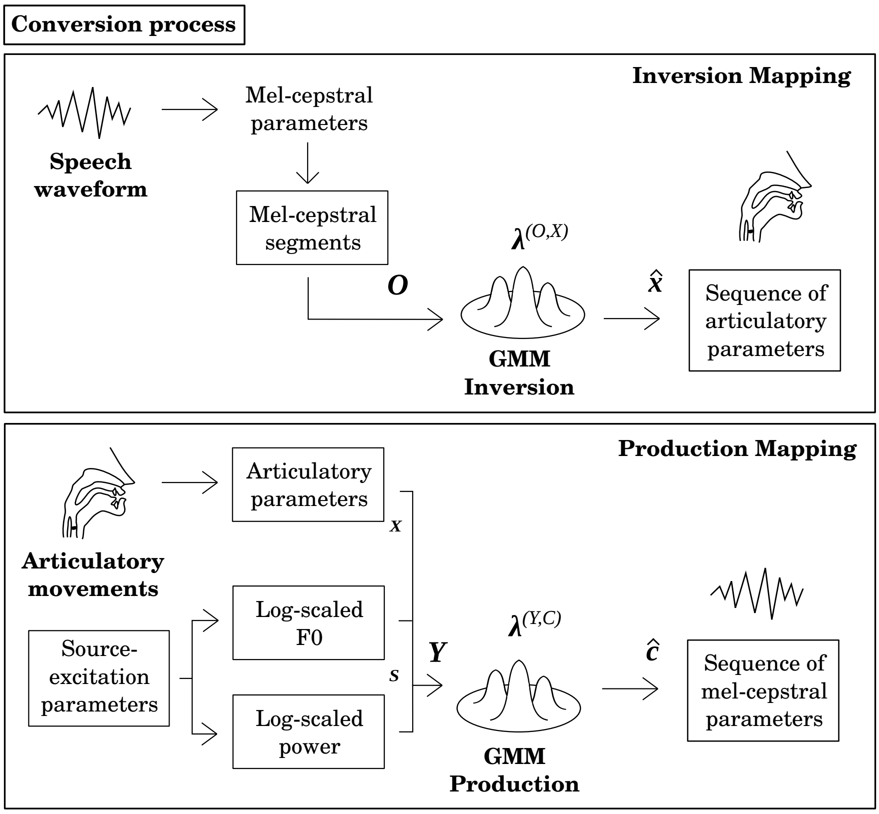


Figure 2.3: *Conversion scheme using GMM of the inversion mapping (top) and the production mapping (bottom).*

$\boldsymbol{\Sigma}_m^{(Y,C)}$ of individual mixture components. These model parameters are also trained with the EM algorithm. The training scheme for the GMM of the production mapping is shown in the lower diagram of Fig. 2.2.

The conversion procedure for the production mapping is also performed in a similar way to in the inversion mapping. Given a time sequence of joint static and dynamic articulatory and source excitation feature vectors $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_T^\top]^\top$, first, the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}^{(Y)}$ is determined as

$$\hat{\boldsymbol{m}}^{(Y)} = \operatorname*{argmax}_{\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}). \tag{2.6}$$

Then, a time sequence of converted mel-cepstral feature vectors $\hat{\boldsymbol{c}}$ is determined as follows:

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}} P(\boldsymbol{C}|\boldsymbol{Y}, \hat{\boldsymbol{m}}^{(Y)}, \boldsymbol{\lambda}^{(Y,C)}), \text{ subject to } \boldsymbol{C} = \boldsymbol{W}^{(c)}\boldsymbol{c}, \tag{2.7}$$

where $\boldsymbol{W}^{(\boldsymbol{c})}$ is the linear transformation matrix used to calculate the sequence of joint static and dynamic mel-cepstral feature vectors $\boldsymbol{C} = [\boldsymbol{C}_1^\top, \ldots, \boldsymbol{C}_T^\top]^\top$ from a sequence of mel-cepstral feature vectors $\boldsymbol{c}$. The conversion scheme using the GMM of the production mapping is shown in the lower diagram of Fig. 2.3.

2.3 Voice Conversion

Voice conversion [1,56] is a framework for transforming the speech characteristics of a source speaker into a particular target speaker while preserving the linguistic contents of the speech signal. A voice conversion system can be used in many speech applications, such as for augmenting speech database with various voice characteristics [1, 57], for singing voice conversion [4, 58], for recovery of impaired speech signal [5, 6, 59], for expressive speech synthesis [7,8], and for body-conducted speech processing [9,10,60]. Owing to its versatility in the development various speech applications, this thesis is written for achieving high-quality and flexible voice conversion, which would be beneficial in research and daily-life applications.

In previous section, related works on speech-articulatory mappings have been presented, which may prove useful for voice conversion if the modeling of vocal-tract shape can be realized. However, it is not straightforward to obtain articulatory data compared to spectral envelope parameters that can be extracted in a more flexible manner from a speech signal. In order to provide a flexible procedure in the development of voice conversion, statistical mapping modeling techniques for spectral envelope parameters are more emphasized within this thesis. Related to that, within this section, an inherent problem of parallel and nonparallel spectral mapping modeling is discussed, where ultimate flexibility of voice conversion development can be achieved with nonparallel modeling. In addition, several issues related to the problem of oversmoothing of generated spectral parameters and the degradation of speech quality using the conventional synthesis procedure are briefly discussed.

2.3.1 Spectral mapping problem of parallel and non-parallel speech data

To develop a voice conversion system, the main problem is located on the realization of a mapping function for the spectral characteristics, such as vocal tract spectrum, of the source speaker into that of the target speaker. One of the most convenient way to do so is through applying the use of data-driven/statistical mapping modeling, such as codebook-based mapping [1], or GMM-based methods [2, 3]. Recent advancements in neural network (NN)-based modeling have also proven to be successful in the implementations of voice conversion [61–63]. This thesis focuses on the use of NN-based architecture, such as conventional deep neural network (DNN), deep mixture density network (DMDN), or recurrent neural network (RNN), for the development of spectral mapping modeling. In particular, the use of DNN- and DMDN-based spectral mapping modeling techniques [15] are given in Chapter 4. On the other hand, to properly achieve high-quality converted speech output with fine-tuning of neural vocoder (briefly mentioned within the next section), a cyclic structure of RNN (CycleRNN)-based spectral mapping modeling [26, 36] is presented in Chapter 5.

Nevertheless, most of the aforementioned spectral mapping techniques still use parallel speech data of source and target speakers, i.e., they utter a same set of sentences. Hence, pairing alignment can be obtained in the development of statistical mapping

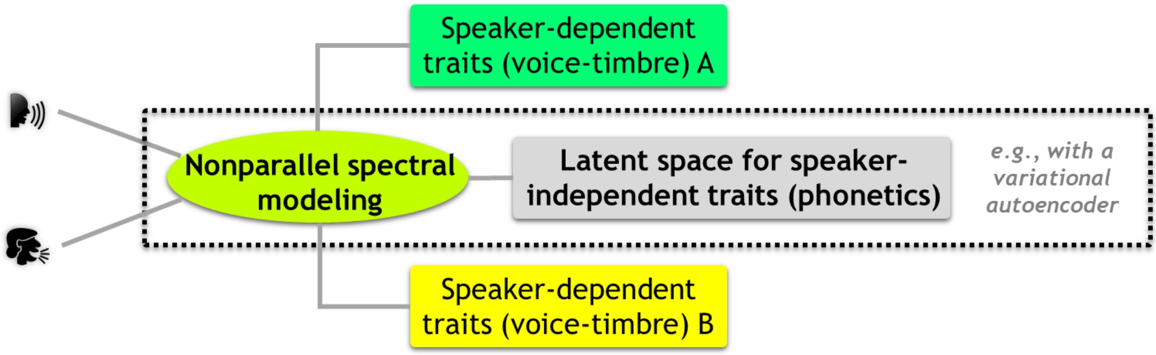

Speaker-dependent traits (voice-timbre) A
Nonparallel spectral modeling
Latent space for speaker-independent traits (phonetics)
e.g., with a variational autoencoder
Speaker-dependent traits (voice-timbre) B

Figure 2.4: *Possibility of nonparallel spectral modeling by means of latent space utilization to capture shared traits between speaker (speaker-independent), such as phonetics, which will be disentangled from the speaker-dependent traits, such as voice-timbre.*

models. However, obtaining such parallel speech dataset requires significant resources, and in most of the time non-parallel speech dataset, i.e., source and target speakers utter different sets of sentences, would certainly be more available. Moreover, it is also not possible to obtain parallel data for sentences in different languages. Related to this issue, several related works on non-parallel spectral modeling exist to realize non-parallel voice conversion, such as restricted Boltzmann machine [64], generative adversarial network [27, 28], and variational autoencoder (VAE) [38]. This thesis will focus on the improvement of VAE-based VC for the development of non-parallel voice conversion, owing to its utilization of latent space for capturing shared representations between source and target speakers, as illustrated in Fig. 2.4. Specifically, in the Chapter 6, a cyclic structure of VAE (CycleVAE) is presented which tackles the problem of its performance degradation by inclusion of converted spectral features in the training phase through the cyclic flow [29].

Parallel conversion model with deep neural network (DNN)

Let $\boldsymbol{x}_t = [x_t(1), x_t(2), \ldots, x_t(D)]^\top$ and $\boldsymbol{y}_t = [y_t(1), y_t(2), \ldots, y_t(D)]^\top$ be the D-dimensional spectral feature vector of the source speaker and that of the target speaker at frame t, respectively. The $2D$-dimensional joint static-delta feature vector of the source and that of the target are then respectively denoted as $\boldsymbol{X}_t = [\boldsymbol{x}_t, \Delta\boldsymbol{x}_t]^\top$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_t, \Delta\boldsymbol{y}_t]^\top$ at frame t, where the delta feature vectors are denoted as $\Delta\boldsymbol{x}_t$ and $\Delta\boldsymbol{y}_t$.

In the conventional DNN architecture, given an input source spectral feature vector $\boldsymbol{X}_t$ and the network parameters $\boldsymbol{\lambda}$, a conditional probability distribution function (pdf) of the target spectral feature vector $\boldsymbol{Y}_t$ on the network output layer is defined as follows:

$$P_s(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{Y}_t; f_\lambda(\boldsymbol{X}_t), \boldsymbol{D}), \tag{2.8}$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. In the above pdf, the network output is denoted as $f_\lambda(\boldsymbol{X}_t)$ and the diagonal covariance matrix of the target spectral feature vector is denoted as $\boldsymbol{D}$, which is inferred from training data. The DNN spectral conversion model is represented by the left graph in Fig. 4.1.

In the training phase, a set of updated network parameters $\hat{\boldsymbol{\lambda}}$ is estimated by back-propagating the following loss function:

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} -P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}), \tag{2.9}$$

where

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} P_s(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda}). \tag{2.10}$$

The spectral feature vector sequence of the source speaker and that of the target speaker are denoted as $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_T^\top]^\top$, respectively.

Note that in the training phase, a dynamic time warping (DTW) procedure is performed by aligning the length of the source spectral feature vector sequence with that of the target one to obtain a pair of time-aligned features.

In the conversion phase, given the source spectral feature vector sequence $\boldsymbol{X}$, the trajectory of the target spectral parameters $\hat{\boldsymbol{y}} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \dots, \boldsymbol{y}_T^\top]^\top$ is computed by the maximum likelihood parameter generation (MLPG) [65] procedure as follows:

$$\hat{\boldsymbol{y}} = (\boldsymbol{W}^\top \boldsymbol{U}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{U}^{-1} \boldsymbol{M}, \tag{2.11}$$

where $\boldsymbol{W}$ is a transformation matrix used to expand a static feature vector sequence into its joint static-delta feature vector sequence. The sequence of target mean vectors is denoted as $\boldsymbol{M} = [f_\lambda(\boldsymbol{X}_1)^\top, f_\lambda(\boldsymbol{X}_2)^\top, \dots, f_\lambda(\boldsymbol{X}_T)^\top]^\top$, whereas the sequence of diagonal covariance matrices is denoted as $\boldsymbol{U} = \boldsymbol{D} \otimes \boldsymbol{I}_{2D \times T}$ with $\otimes$ denoting the Kronecker delta product.

Non-parallel conversion model using VAE-based VC

Let $\boldsymbol{X}_t = [\boldsymbol{e}_t^{(x)^\top}, \boldsymbol{s}_t^{(x)^\top}]^\top$, $\boldsymbol{e}_t^{(x)} = [e_t^{(x)}(1), \dots, e_t^{(x)}(D_e)]^\top$, and $\boldsymbol{s}_t^{(x)} = [s_t^{(x)}(1), \dots, s_t^{(x)}(D_s)]^\top$ be the $D_e + D_s$, D_e, and D_s-dimensional feature vectors of the input, the excitation, and the spectra, respectively, at frame t. In the training phase, given a set of network parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$, a sequence of input features $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \dots, \boldsymbol{X}_T^\top]^\top$ and time-invariant D_c-dimensional source speaker-code features $\boldsymbol{c}^{(x)}$ [38], a set of updated network parameters $\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}\}$ is estimated by maximizing the variational lower bound function [37] as follows:

$$\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}\} = \underset{\boldsymbol{\theta}, \boldsymbol{\phi}}{\operatorname{argmax}} \sum_{t=1}^{T} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{X}_t, \boldsymbol{c}^{(x)}), \tag{2.12}$$

where

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{X}_t, \boldsymbol{c}^{(x)}) = &-D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|\boldsymbol{X}_t)||p_{\boldsymbol{\theta}}(\boldsymbol{z}_t)) \\ &+ \mathbb{E}_{q_{\boldsymbol{\phi}(\boldsymbol{z}_t|\boldsymbol{X}_t)}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_t^{(x)}|\boldsymbol{z}_t, \boldsymbol{c}^{(x)})], \qquad (2.13)\end{aligned}$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|\boldsymbol{X}_t) = \mathcal{N}(\boldsymbol{z}_t; f_{\boldsymbol{\phi}}^{(\mu)}(\boldsymbol{X}_t), \mathrm{diag}(f_{\boldsymbol{\phi}}^{(\sigma)}(\boldsymbol{X}_t)^2)), \qquad (2.14)$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{s}_t^{(x)}|\boldsymbol{z}_t, \boldsymbol{c}^{(x)}) \approx \mathcal{N}(\boldsymbol{s}_t^{(x)}; g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_t^{(x)}, \boldsymbol{c}^{(x)}), \boldsymbol{I}), \qquad (2.15)$$

$$\hat{\boldsymbol{z}}_t^{(x)} = f_{\boldsymbol{\phi}}^{(\mu)}(\boldsymbol{X}_t) + f_{\boldsymbol{\phi}}^{(\sigma)}(\boldsymbol{X}_t) \odot \boldsymbol{\epsilon} \quad \text{s. t.} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \qquad (2.16)$$

$\boldsymbol{z}_t$ denotes a D_z-dimensional latent feature vector, $f_{\boldsymbol{\phi}}(\cdot)$ denotes an encoder network, $g_{\boldsymbol{\theta}}(\cdot)$ denotes a decoder network, $\odot$ denotes an element-wise product, and $\mathcal{N}(;\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is for a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Therefore, the reconstructed source spectra feature vector $\hat{\boldsymbol{s}}_t^{(x)}$, i.e., estimated spectra with the same speaker characteristics as the input source speaker, is given by

$$\hat{\boldsymbol{s}}_t^{(x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_t^{(x)}, \boldsymbol{c}^{(x)}). \qquad (2.17)$$

On the other hand, the converted source-to-target spectra $\hat{\boldsymbol{s}}_t^{(y|x)}$, i.e., estimated spectra with the voice characteristics of a desired target speaker, is given by

$$\hat{\boldsymbol{s}}_t^{(y|x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_t^{(x)}, \boldsymbol{c}^{(y)}), \qquad (2.18)$$

where $\boldsymbol{c}^{(y)}$ denotes the time-invariant D_c-dimensional target speaker-code features [38]. In this work, not only source speakers, but also target speakers are used as input in training. In order to use the corresponding target speaker as the input speaker, i.e., optimization of reconstructed target spectra and/or performing target-to-source conversion, the notations of x and y, in Eqs. (2.12)–(2.18), are swapped with each other. Though, the performance of VAE-based VC is noticeably insufficient because the converted features are not considered in the parameter optimization.

2.3.2 Oversmoothing problem of spectral trajectory and global variance postfilter solution

It is well known that the maximum likelihood (ML)-based optimization in a statistical modeling would cause oversmoothing problem in the parameter estimation. This oversmoothing problem is shown by the lack of variance structure within the trajectory of the estimated spectral parameters in a voice conversion system [3]. The lack of detailed structure in the spectral trajectory would cause speech quality degradation, where the sounds of the speech tends to be muffled. In order to alleviate this problem, a global variance (GV) postfilter solution has been proposed in [3], which can take into account the variance of the natural spectral trajectory of the training data to enhance the detailed structure of the estimated spectral trajectory. Such method has been proven to significantly improve the quality of the converted speech waveform. However, due to the use of vocoder-based excitation generation for generating the synthetic speech, such as in [13,22], the quality of the converted speech is still limited. Two possible approaches can be adopted to address the issue of quality limitation. First is the by avoiding the use of vocoder-based excitation generation assumptions with direct waveform modification as briefly explained within the next subsecion. Second is by the use of neural vocoder to ultimately achieve higher-quality of converted speech in a data-driven manner as will be explained within the next section. Nevertheless, the global variance postfilter is still useful for adopting a straightforward method in alleviating the quality degradation, which is also used in the voice conversion framework elaborated in Chapter 4.

GV postfiltering procedure for converted spectra

Given a converted spectral trajectory $\hat{\boldsymbol{y}}$, the GV [3] postprocessed spectra $\hat{\boldsymbol{y}}'$ is calculated as follows:

$$\hat{y}_t'(d) = \beta \left(\sqrt{\frac{\mu_v(d)}{v(d)}} (\hat{y}_t(d) - \overline{\hat{y}}(d)) + \overline{\hat{y}}(d) \right) + (1-\beta)\hat{y}_t(d), \tag{2.19}$$

where the GV of the d-th dimension of the target spectral trajectory, computed beforehand using whole training data, is given by

$$\mu_v(d) = \frac{1}{T}\sum_{t=1}^{T}(y_t(d) - \overline{y}(d))^2, \tag{2.20}$$

the mean of target spectral trajectory is computed as

$$\overline{y}(d) = \frac{1}{T}\sum_{t=1}^{T} y_t(d), \tag{2.21}$$

and β denotes the weighting coefficient. The GV of converted spectral trajectory $\boldsymbol{v} = [v(1), \ldots, v(D)]^\top$ is also computed beforehand, with respect to the training data, in a similar manner as in Eq. (2.20), whereas the mean of converted spectral trajectory $\overline{\hat{\boldsymbol{y}}} = [\overline{\hat{y}}(1), \ldots, \overline{\hat{y}}(D)]^\top$ is computed with respect to the testing source utterance that is to be converted. Hence, the oversmoothing of converted spectral trajectory can be alleviated by enlarging the movements of the spectral trajectory to match the GV statistics of the target.

2.3.3 Overcoming speech quality degradation through direct waveform modification

As has been mentioned beforehand, the use of vocoder-based excitation generation assumptions causes significant quality degradations in a voice conversion system [4]. In

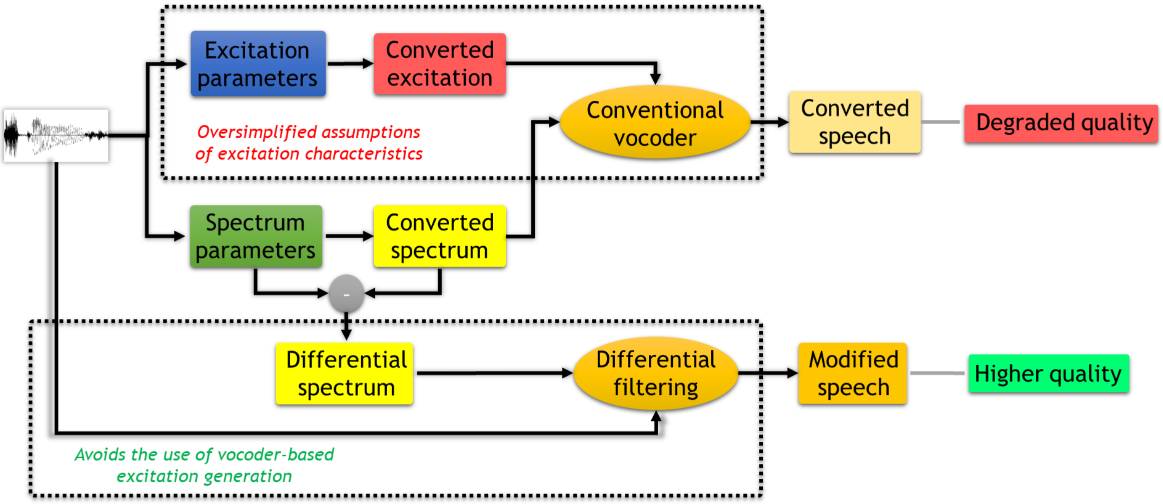


Figure 2.5: *Direct waveform modification with differential spectrum in voice conversion to avoid the use of vocoder-based excitation generation.*

order to alleviate this problem, a direct waveform modification method that avoids the use of vocoder-based excitation generation has been proposed [4]. Specifically, as illustrated in Fig. 2.5, this technique employs the use of spectrum differential parameters that are computed by taking the differences between the converted spectral parameters and the natural input spectral parameters. The resulting parameter coefficients are used to directly filter the input speech waveform signal to produce a modified speech waveform that would have higher-quality compared to using conventional vocoder-based excitation generation [13, 22]. In this thesis, this direct waveform modification method is also employed to improve the quality of the modified speech signal within the framework of the voice modification with articulatory mapping and manipulation, as in Chapter 3. However, in a practical voice conversion framework, the conversion of excitation features, such as fundamental frequency, might still be needed for example in cross-gender conversions (female-to-male speakers or male-to-female speakers). In such a case, it is not straightforward to employ the direct waveform modification method. Hence, this technique serves more as a way to perform postprocessing [15] to obtain refined spectral parameters to be used for a neural vocoder in voice conversion as in Chapter 4. Definite quality improvement will be achieved with fine-tuning procedure of neural vocoder instead of direct waveform modification as described within the next section an as used in Chapter 5.

Direct waveform modification procedure

The direct waveform modification technique [4] utilizes the MLSA filter [20], on which its synthesis function for a waveform signal of the source speaker $\boldsymbol{x} = [x_1, \ldots, x_T]^\top$ is as follows:

$$X(z) = H_t^{(x)}(z)E(z), \tag{2.22}$$

where $E(z)$ is the transfer function of the excitation signal, and the filter transfer function $H_t^{(x)}(z)$ is defined as

$$H_t^{(x)}(z) = \exp \sum_{m=0}^{M} c_{\alpha,t}^{(x)}[m] z_\alpha^{-m}, \tag{2.23}$$

where the all-pass filter z_α^{-m} is given by

$$z_\alpha^{-m} = \frac{z^{-m} - \alpha}{1 - \alpha z^{-m}}. \tag{2.24}$$

α denotes the frequency warping parameter, $c_{\alpha,t}^{(x)}[m]$ denotes the m-th mel-cepstrum coefficient of the mel-cepstrum feature vector $\boldsymbol{c}_{\alpha,t} = [c_{\alpha,t}^{(x)}[1], \ldots, c_{\alpha,t}^{(x)}[M]]^\top$ at time t, and the number of mel-cepstrum coefficients is M.

By retaining the excitation signal of the source speaker, the synthesis function of the waveform signal of the target speaker $\boldsymbol{y} = [y_1, \ldots, y_T]^\top$ can be defined in a similar way as in Eq. (2.22) as follows:

$$Y(z) = \frac{H_t^{(y)}(z)}{H_t^{(x)}(z)} X(z), \tag{2.25}$$

because

$$E(z) = \frac{X(z)}{H_t^{(x)}(z)}, \tag{2.26}$$

and

$$Y(z) \simeq H_t^{(y)}(z) E(z). \tag{2.27}$$

Hence, given the m-th mel-cepstrum differential $d_{\alpha,t}[m]$ that is computed as

$$d_{\alpha,t}[m] = c_{\alpha,t}^{(y)}[m] - c_{\alpha,t}^{(x)}[m], \tag{2.28}$$

the transfer function of the differential filter $H_t^{(y/x)}(z)$ is given by

$$H_t^{(y/x)}(z) = \frac{H_t^{(y)}(z)}{H_t^{(x)}(z)} = \exp \sum_{m=0}^{M} d_{\alpha,t}[m] z_\alpha^{-m}, \tag{2.29}$$

because of by following Eq. (2.23)

$$\frac{H_t^{(y)}(z)}{H_t^{(x)}(z)} = \frac{\exp \sum_{m=0}^{M} c_{\alpha,t}^{(y)}[m] z_\alpha^{-m}}{\exp \sum_{m=0}^{M} c_{\alpha,t}^{(x)}[m] z_\alpha^{-m}}. \tag{2.30}$$

2.4 Neural Vocoder

Recently, the development of NN-based architectures for the modeling of speech signal waveform, i.e., neural vocoder, has been growing rapidly. This includes the emergence of Wavenet [14,23], SampleRNN [66,67], WaveRNN [68,69], FFTNet [70,71], LPCNet [72,73], parallel WaveNet [74,75], WaveGlow [76], and neural-source-filter [77]. The capability of neural vocoder frameworks to generate high-quality synthetic speech has been proven, and has surpassed that of the conventional rule-based vocoder. Further, thanks to the data-driven concept, a neural vocoder can be additionally adapted (fine-tuned) to accomodate the condition of distorted speech features, such as spectral parameters estimated from a spectral mapping modeling. Hence, in order to achieve high-quality converted speech output in voice conversion, in this thesis, the autoregressive WaveNet vocoder [14, 23] will be used and briefly discussed within this section to elaborate its possible limitation, owing to its use in voice conversion.

2.4.1 Autoregressive WaveNet vocoder for achieving high-quality synthetic speech and its possible limitation

WaveNet consists of a stack of dilated causal convolutional layers with residual blocks that can effectively models the long-term causal relationship of the speech waveform samples [14]. A WaveNet vocoder is capable of generating meaningful speech waveform signal by conditioning the network to a set of speech parameters, such as spectral and excitation features [23]. In the synthesis time, the speech waveform is generated sample-by-sample, i.e., in an autoregressive manner. A well trained WaveNet vocoder is able to generate high-quality synthetic speech that is indistinguishable from a natural speech. However, when it is used in voice conversion, there exist mismatches between the estimated speech features, such as converted spectral parameters, and the natural speech features used in the WaveNet training. In this thesis, this problem is solved by means of either a postprocessing method for the converted spectral features or by directly addressing the mismatches problem through fine-tuning the WaveNet vocoder.

The first approach to address the limitation of WaveNet vocoder in voice conversion, i.e., the mismatches between spectral and waveform modeling, is to use a postprocessing method based on direct waveform modification [15], which has been briefly explained

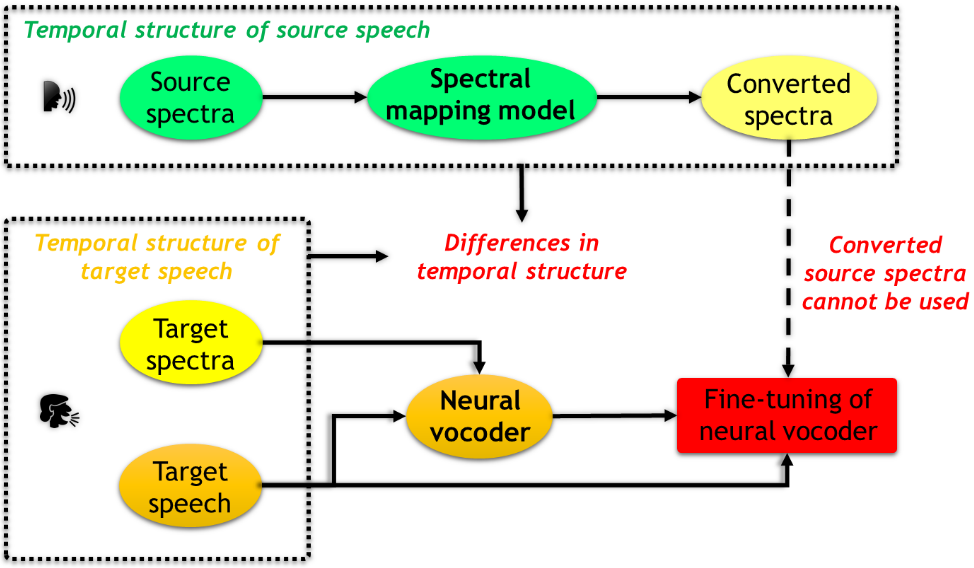

Temporal structure of source speech
Source spectra
Spectral mapping model
Converted spectra
Temporal structure of target speech
Differences in temporal structure
Converted source spectra cannot be used
Target spectra
Neural vocoder
Fine-tuning of neural vocoder
Target speech

Figure 2.6: *Problem of neural vocoder fine-tuning in voice conversion due to the differences in temporal structure between source speaker and target speaker.*

in the previous section. This technique will be further elaborated in the usage of a voice conversion in Chapter 4. Secondly, to directly address the mismatches within the WaveNet vocoder, instead of using a spectral postprocessing method, a fine-tuning procedure for WaveNet vocoder in voice conversion will be utilized [35, 36]. As illustrated in Fig. 2.6, it is not straightforward to perform fine-tuning of a neural vocoder in voice conversion using converted spectral features, owing to the differences of temporal structure between source and target speakers. This limitation/problem is addressed through the use of a cyclic structure of recurrent neural network (CycleRNN) for spectral mapping modeling that can estimate oversmoothed target spectra with the same temporal structures as the target speech [36]. This technique, which makes it possible to achieve high-quality voice conversion will be described in more detail in Chapter 5.

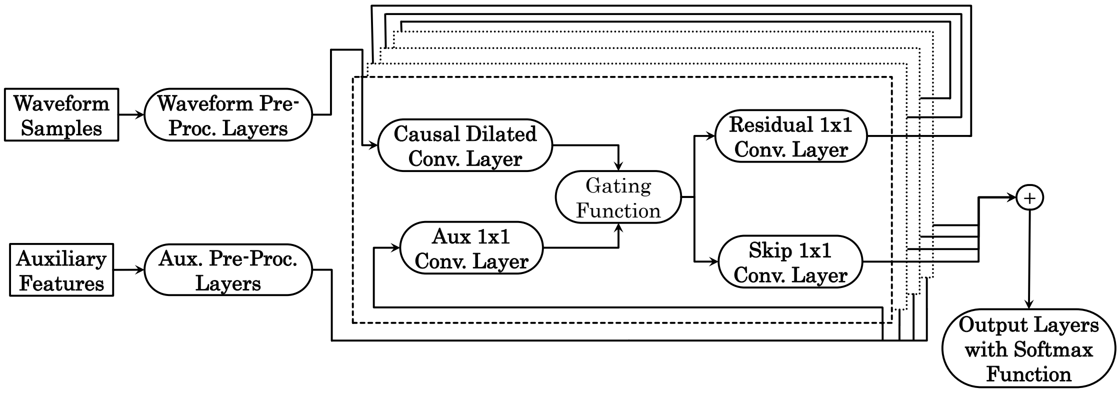


Figure 2.7: *Architecture of the WaveNet vocoder using dilated causal convolution within layers of residual blocks.*

2.4.2 WaveNet Vocoder

WaveNet [14], as illustrated in Fig. 2.7, is a deep AR waveform generation network that is capable of efficiently modeling waveform samples based on their corresponding previous samples through the use of a stack of dilated convolutional layers. When it is conditioned with auxiliary speech parameters [23, 24], such as spectral and excitation features, a well-developed WaveNet vocoder is capable of producing meaningful speech waveform signals with natural quality. Given a sequence of speech waveform samples $\boldsymbol{s} = [s_1, s_2, \ldots, s_t, \ldots, s_T]^\top$, the likelihood function of the WaveNet vocoder is defined as

$$P(\boldsymbol{s}|\boldsymbol{h}, \boldsymbol{\theta}) = \prod_{t=1}^{T} P(s_t|\boldsymbol{s}_{t-p}, \boldsymbol{h}'_t, \boldsymbol{\theta}), \tag{2.31}$$

where $\boldsymbol{h}$ denotes a sequence of auxiliary feature vectors and $\boldsymbol{h}'_t$ is the upsampled auxiliary feature vector at time t. For a waveform sample s_t at time t, the sequence of its p previous samples is denoted as $\boldsymbol{s}_{t-p}$. The set of WaveNet parameters is denoted as $\boldsymbol{\theta}$.

Details of the WaveNet architecture, as illustrated in Fig. 2.7, are as follows. An input sequence of auxiliary speech feature vectors is fed through convolutional input layers, which are the same as those used for the RNN-based spectral mapping, described in Section 5.2. Then, a trainable upsampling layer is used to match the resolution of the speech parameters to that of the speech waveform samples. On the other hand, the waveform samples are first discretized into 256 categorical values using the μ-law algorithm, which gives a 256-dimensional one-hot vector for each time t. The sample feature vectors are then processed through a causal (2×1) convolutional input layer. Following these preprocessing steps, both the sample feature vectors and the auxiliary feature vectors are fed through a stack of dilated convolutional layers with residual blocks. Specifically, for each residual block, a 2×1 convolution layer with double the dilation size of the residual block one step deeper is used to process the sample feature vectors. The sequence of doubled dilation sizes is repeated several times, in other words, after the end of the doubled dilation sequence, the sequence starts again from 1. On the other hand, a 1×1 convolution layer is used to process the auxiliary feature vectors. For the kth residual block at time t, the output of the 2×1 convolution $\tilde{\boldsymbol{s}}_{t,k}$ and that of the 1×1 convolution $\tilde{\boldsymbol{h}'}_{t,k}$ are fed to a gating function to produce the output of the residual block as follows:

$$\sigma(\boldsymbol{W}_{f,k}\tilde{\boldsymbol{s}}_{t,k} + \boldsymbol{V}_{f,k}\tilde{\boldsymbol{h}}'_{t,k}) \odot \tanh(\boldsymbol{W}_{g,k}\tilde{\boldsymbol{s}}_{t,k} + \boldsymbol{V}_{g,k}\tilde{\boldsymbol{h}}'_{t,k}), \tag{2.32}$$

where $\odot$ denotes the elementwise product, f and g denote the filter and gate, respectively, and the corresponding convolution parameters are denoted by $\boldsymbol{W}$ and $\boldsymbol{V}$. The output of the residual block is then passed to both a skip layer connection and the next residual block. The collection of output feature vectors from the skip connections are summed and fed to the final output layers, which employ the softmax function to treat the WaveNet optimization as a classification problem, i.e., with cross-entropy loss. During the generation of a speech waveform signal, a sampling procedure is simply performed using the softmax distribution to estimate the speech waveform sample by sample.

2.4.3 Collapsed Waveform Detection

The WaveNet-based vocoder is capable of generating much more natural-sounding waveforms than a conventional vocoder. However, sometimes the converted waveform generated by the WaveNet vocoder incorporates collapsed segments. In VC, this is most likely caused by the mismatch between the converted auxiliary features and the original features used in training the model.

In [78], a power-based detector is employed to automatically detect collapsed segments in WaveNet-generated waveforms. From the spectrum of a generated waveform, frame-based summation is performed using the power spectrum of all frequency bins $\boldsymbol{P}$ and that of the Nyquist frequency components $\boldsymbol{L}$. Let $\boldsymbol{P}^{(W)} = [P_1^{(W)}, \ldots, P_T^{(W)}]$ and $\boldsymbol{P}^{(C)} = [P_1^{(C)}, \ldots, P_T^{(C)}]$ denote the power summation sequence from all frequency bins of a WaveNet-generated waveform and that of a conventional vocoder, respectively. The power summation sequences from the Nyquist frequency components are respectively denoted as $\boldsymbol{L}^{(W)} = [L_1^{(W)}, \ldots, L_T^{(W)}]$ and $\boldsymbol{L}^{(C)} = [L_1^{(C)}, \ldots, L_T^{(C)}]$. In the detection, the differences in the maximum power between the WaveNet-generated waveform and that generated the conventional vocoder are computed as follows:

$$\Delta\boldsymbol{P} = \max(\boldsymbol{P}^{(W)}) - \max(\boldsymbol{P}^{(C)}) \tag{2.33}$$

$$\Delta\boldsymbol{L} = \max(\boldsymbol{L}^{(W)}) - \max(\boldsymbol{L}^{(C)}). \tag{2.34}$$

The system selects the best waveform generation flow through the comparison of $\Delta\boldsymbol{P}$ and $\Delta\boldsymbol{L}$ with an empirical threshold, where both values will be higher than the threshold for a low-quality waveform.

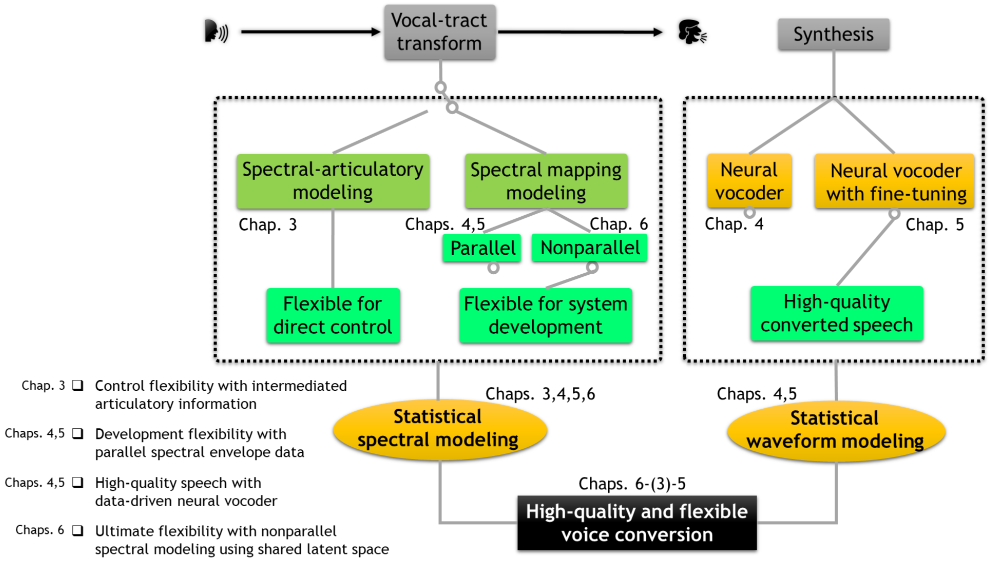

Vocal-tract transform
Synthesis
Spectral-articulatory modeling
Spectral mapping modeling
Neural vocoder
Neural vocoder with fine-tuning
Chap. 3
Chaps. 4,5
Chap. 6
Chap. 4
Chap. 5
Parallel
Nonparallel
Flexible for direct control
Flexible for system development
High-quality converted speech
Chap. 3 □ Control flexibility with intermediated articulatory information
Chaps. 4,5 □ Development flexibility with parallel spectral envelope data
Chaps. 4,5 □ High-quality speech with data-driven neural vocoder
Chaps. 6 □ Ultimate flexibility with nonparallel spectral modeling using shared latent space
Chaps. 3,4,5,6
Statistical spectral modeling
Chaps. 4,5
Statistical waveform modeling
Chaps. 6-(3)-5
High-quality and flexible voice conversion

Figure 2.8: *Relation with thesis scope to achieve high-quality and flexible voice conversion system.*

2.5 Relations with Thesis Scope for High-Quality and Flexible Voice Conversion

In this thesis, to comprehensively study the development of high-quality and flexible voice conversion with statistical spectral and waveform modeling techniques, the aforementioned related works are utilized throughout this work.

In the first part, a voice modification system is realized by the use of mapping approach between physical configurations of the vocal tract (articulators) and speech signal. The so-called articulatory controllable speech modification system integrates the GMM-based inversion and production mappings in a sequential manner while allowing the manipulation of intuitive representations of articulatory positions, such as to modify phonemic sounds or to modify the articulation effort. High-quality modified speech signal can also be synthesized by harnessing the use of direct waveform modification technique that avoids the use of vocoder-based excitation generation assumptions (direct waveform modification). As illustrated in Fig. 2.8, such framework would provide a flexible way for directly controlling the speech signal, thanks to the use of intuitive articulatory representations. However, obtaining articulatory data for flexible statistical model development is not straightforward, and more advanced technique is needed to fully model the vocal-tract shape for achieving voice conversion procedure with physical-based representations.

In the second part, the voice conversion framework is realized by implementing the mapping function of spectral envelope parameters between that of the source and that of the target speakers with the use of neural-network (NN)-based architectures. Further, to improve the quality of the converted speech waveform, statistical waveform modeling, i.e., neural vocoder, is deployed, specifically, the WaveNet vocoder. To address the mismatches that exist between the estimated spectral parameters from the spectral mapping model and the trained WaveNet vocoder, a postprocessing method based on the signal processing technique on direct waveform modification is employed. As depicted in Fig. 2.8, the use of statistical spectral mapping modeling, particularly for spectral envelope parameters, allows more flexible approach in system development compared to using articulatory data.

In this thesis, to facilitate further improvement of voice conversion framework with neural vocoder, a fine-tuning approach that directly deals the mismatches of the spectral parameters between spectral mapping model and WaveNet vocoder is fully elaborated. The fine-tuning approach is based on a cyclic structure of RNN (CycleRNN) spectral mapping model that can estimate the oversmoothed spectral parameters to be used in the fine-tuning of a WaveNet vocoder. Finally, the problem of non-parallel spectral modeling in a voice conversion framework is also tackled within this thesis with the use of variational autoencoder (VAE)-based VC system, which uses a regularized latent space for the shared/common representations between different speakers. Specifically, in this work, a cyclic framework of VAE (CycleVAE) is presented which is capable of taking into account the converted spectral parameters in the optimization of the network, where, in the conventional VAE, it can only uses reconstruction aspect of the spectral features. As also depicted in Fig. 2.8, the use of neural vocoder, especially with fine-tuning procedure makes it possible to achieve high-quality converted speech output within the voice conversion framework. Moreover, improvements of nonparallel spectral modeling will ultimately allows a flexible voice conversion development procedure, where arbitrary speech dataset can be used between different speakers.

3 Statistical Voice Modification with Articulatory Mapping and Manipulation

3.1 Introduction

During speech production, both of the vocal folds and the articulators are used to achieve the so-called source-filter combination in the generation of speech signals [11, 12]. To accomplish this, the air pressure must be increased in the lungs. Then, the corresponding air flow is channeled in the trachea through the vocal folds. A particular characteristic of the excitation signal is then determined by the configuration of the vocal folds while the air is flowing, for example, a periodic signal is produced by constant vibrations of the vocal folds. Subsequently, this source-excitation signal is modulated within the vocal tract by the articulatory organs, including the tongue, teeth, and velum. Hence, a certain configuration of articulators appropriately determines the resonance/filter characteristics of the vocal tract, which, in turn naturally regulates the traits of the generated phonemic sounds.

This intimate relationship between speech and articulatory organs appears to be in contrast with their corresponding attributes. While producing speech sounds, the movements of the articulators in fact vary much more slowly than their counterparts in the speech signal [34, 79, 80], such as the trajectory of the vocal tract spectrum.

Undeniably, a broad range of possibilities in the development of speech technologies would be viable through the utilization of slowly varying articulatory representations, such as articulatory parameters. Indeed, researchers have been extensively studying the use of articulatory parameters in speech technologies for several decades. Several notable comprehensive works have been reported on applications to low-bit-rate speech coding [40, 79], speech analysis and synthesis [79, 81, 82], speech recognition [19, 83, 84], and speech visualization [85, 86].

To establish a relationship between speech and vocal tract composition, it is widely known that the fundamental approach is based on mathematical functions [32, 34, 39, 40]. Unfortunately, the nature of the speech production mechanism itself does not provide a straightforward procedure for doing this. This is shown by the fact that there is no one-to-one mapping between speech signals and configurations of articulators. Such a peculiarity is observed in the so-called inverse mapping from acoustic to articulatory parameters [31, 33] as well as in the forward/production mapping from articulatory configurations to the vocal tract spectrum [30, 50]. A vast number of approximations must be considered when examining the affiliation between speech and articulators.

Recently, researchers have considered the use of statistical approaches. This has been made possible by the availability in parallel recording data of speech and articulatory movements [87–90]. Indeed, the elegance of capturing statistical traits within the accessible data has led to many notable works on the advancement of statistical data-driven methods for both acoustic-to-articulatory inversion and articulatory-to-acoustic production mappings. The statistical approach for the acoustic-to-articulatory inversion mapping was first introduced with the use of a codebook-based method [41]. Later, in [42], it was reported that by incorporating a constraint on the articulation dynamics, an improvement in the accuracy of inversion mapping can be achieved. The utilization of a neural-network method in inversion mapping was reported in [43, 44]; in [44], it was found that by employing multiple mixtures to model the density of articulatory features, significant improvement in the estimation accuracy can be obtained. Then, in [45,46], phonetic information was used to improve the mapping effectiveness. Meanwhile, in [47], a Gaussian mixture model (GMM)-based mapping approach was shown to be capable of preserving the effectiveness while allowing independence from textual input features. For the articulatory-to-acoustic production mapping, the progress can be described in a similar manner to that for the inversion mapping, where it was first used with a codebook-based method in [51]. In [52, 53], production mapping based on a neural network was reported, and phonetic information was shown to enhance the production mapping performance in [54]. Similarly to above, a GMM-based production mapping was reported to perform effectively in [55].

In this work, the GMM-based statistical method for both the acoustic-to-articulatory inversion and articulatory-to-acoustic production mappings [16] is used, where the GMM concept itself is widely used in voice conversion systems [2]. The GMM-based statistical feature mapping technique essentially has three notable advantages. First, it provides non-black-box procedures in both modeling and estimation mechanisms with low resources. Second, it allows the possibility of developing language-independent systems owing to its independence of textual input features. Third, its low computational complexity opens a wide range of possibilities for implementation, particularly for real-time processing. Thus, in this work, to maximize the potential of GMM-based inversion and production mappings, an attempt to utilize the close relationship between speech sounds and articulatory movements has to be made. One effective means of achieving this is by developing a system capable of producing modified speech sounds through adjustments of unobserved articulatory features, which is closely related to a system using an HMM-based technique [17]. A system that is capable of performing the aforementioned scheme will offer immense opportunities in the development of various speech applications, such as acoustic and/or articulatory visualization feedback for speech therapy [91, 92], language learning/pronunciation training [93, 94].

To make it possible to take advantage of the use of articulatory parameters in the above speech applications, in this chapter, a voice modification system with articulatory manipulation that employs the GMM-based statistical inversion and production mappings is presented. This system allows one to modify an input speech signal through manipulation of the unobserved articulatory movements. In a more advanced development to adjust for various speech applications, one can conveniently adapt the manipulation of articulatory parameters for different procedures. Further, for the use in a voice conversion system, articulatory parameters have to be utilized for the estimation and mapping of the physical vocal-tract shape between the source and the target speakers. In this system, however, an integration of the statistical inversion and production mappings is performed into a single sequential mapping procedure, which allows one to adjust the unobserved articulatory parameters. These unobserved movements of the articulators can be conveniently modified with an advanced manipulation procedure, which considers the intercorrelation between articulatory parameters [95]. Additionally, high-quality modified speech sounds can be generated with the implementation of direct waveform modification method, capable of avoiding vocoder-based waveform generation by straightforwardly filtering an input speech signal with spectrum differential parameters [96, 97]. The experimental evaluation results suggest that the system makes it possible to produce more accurate spectral parameters, generate natural trajectories of modified articulatory movements, yield high-quality modified speech sounds, and control appropriate articulatory configurations for the modification of several vowel sounds.

3.2 Acoustic-Articulatory Speech Data

In this work, the Multichannel Articulatory Database (MOCHA) [98] is used as the acoustic-articulatory data, which is provided by the Centre for Speech Technology Research (CSTR), University of Edinburgh. The MOCHA database consists of speech and articulatory movement data, which were simultaneously recorded at Queen Margaret University College. It contains two sets of speaker data from one male speaker (msak0) and one female speaker (fsew0), with both speakers having a Southern England accent. There are a total of 460 British TIMIT sentences uttered by each speaker.

The speech data were recorded with a sampling rate of 16 kHz. For the articulatory movement data, an electromagnetic articulograph (EMA) device was used to record the positions of articulators while speaking. The EMA data provide recorded measurements of seven articulators: the upper lip, lower lip, lower incisor, tongue tip, tongue body, tongue dorsum, and velum. The locations of the articulators are measured as x- and y-coordinates in the mid-sagittal plane, where the bridge of the nose and the upper incisor are chosen as points of reference. The articulatory movement data were recorded with a sampling rate of 500 kHz. Preprocessing procedures were performed [16] to reduce the effect of noise from measurement errors and normalize data values into Z-scores.

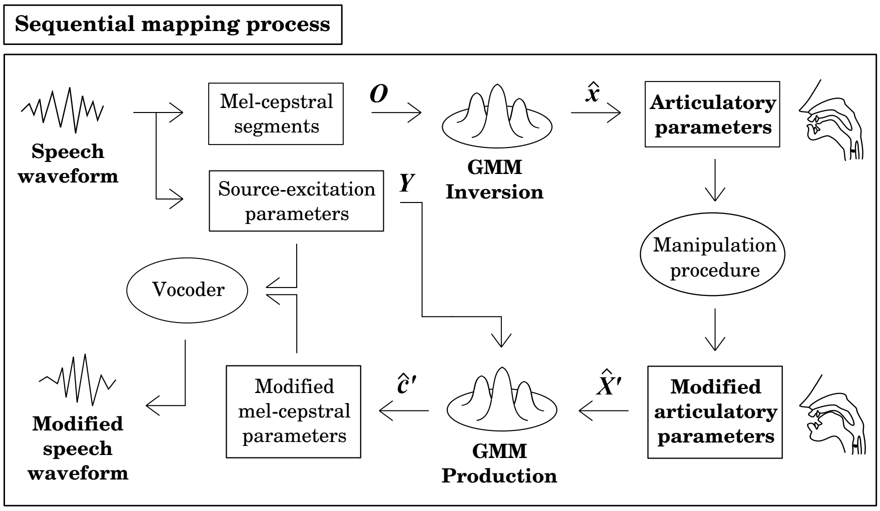


Figure 3.1: *Flow of the articulatory controllable speech modification system using the sequential inversion and production mapping procedure.*

3.3 Articulatory Controllable Speech Modification using GMM-based Models

3.3.1 Sequential Procedure of Inversion and Production Mappings

In this work, to harness the use of articulatory parameters, an articulatory controllable speech modification system based on a sequential mapping process using both the GMM of the inversion mapping and that of the production mapping is presented. These two GMMs are trained by the procedures described in Sections 2.2.3 and 2.2.3, respectively. By performing a sequence of inversion and production mappings, an input speech signal can be conveniently modified through manipulation of the unobserved articulatory movements. The methods for manipulating the articulatory parameters are elaborated in Section 3.3.2.

The flow of the sequential mapping is shown in Fig. 3.1. First, given an input speech signal, its spectral envelope parameters $\boldsymbol{c}$, i.e., mel-cepstral coefficients, and its source excitation parameters $\boldsymbol{s}$, i.e., log-scaled F_0 and log-scaled waveform power, are extracted. Then, the corresponding articulatory parameters $\hat{\boldsymbol{x}}$ are estimated from the mel-cepstral segments $\boldsymbol{O}$ by using the GMM of the inversion mapping as described in Section 2.2.3. To modify the spectral characteristics of the input speech signal, these estimated articulatory parameters $\hat{\boldsymbol{x}}$ are manually manipulated to produce a set of modified articulatory parameters $\hat{\boldsymbol{x}}'$. Then, the corresponding spectral envelope parameters $\hat{\boldsymbol{c}}'$ are estimated from the joint features $\boldsymbol{Y}$ of the modified articulatory parameters $\hat{\boldsymbol{x}}'$ and the source excitation parameters $\boldsymbol{s}$ by using the GMM of the production mapping as described in Section 2.2.3. Finally, the modified speech signal is generated from the modified spectral envelope parameters $\hat{\boldsymbol{c}}'$ and the source excitation parameters $\boldsymbol{s}$ by using a vocoder-based waveform generation procedure.

3.3.2 Methods for Manipulating Articulatory Parameters

To modify the unobserved articulatory movements, two methods for manipulating the articulatory parameters are presented: a simple manipulation method and a smoothing method to take into account the intercorrelation of articulatory parameters.

Simple manipulation method

Let $\hat{\boldsymbol{x}}_t$ be the estimated articulatory feature vector at frame t. A manipulated articulatory feature vector $\hat{\boldsymbol{x}}'_t$ is then given by the following linear transformation:

$$\hat{\boldsymbol{x}}'_t = \boldsymbol{\Lambda}_t \hat{\boldsymbol{x}}_t + \boldsymbol{\psi}_t, \tag{3.1}$$

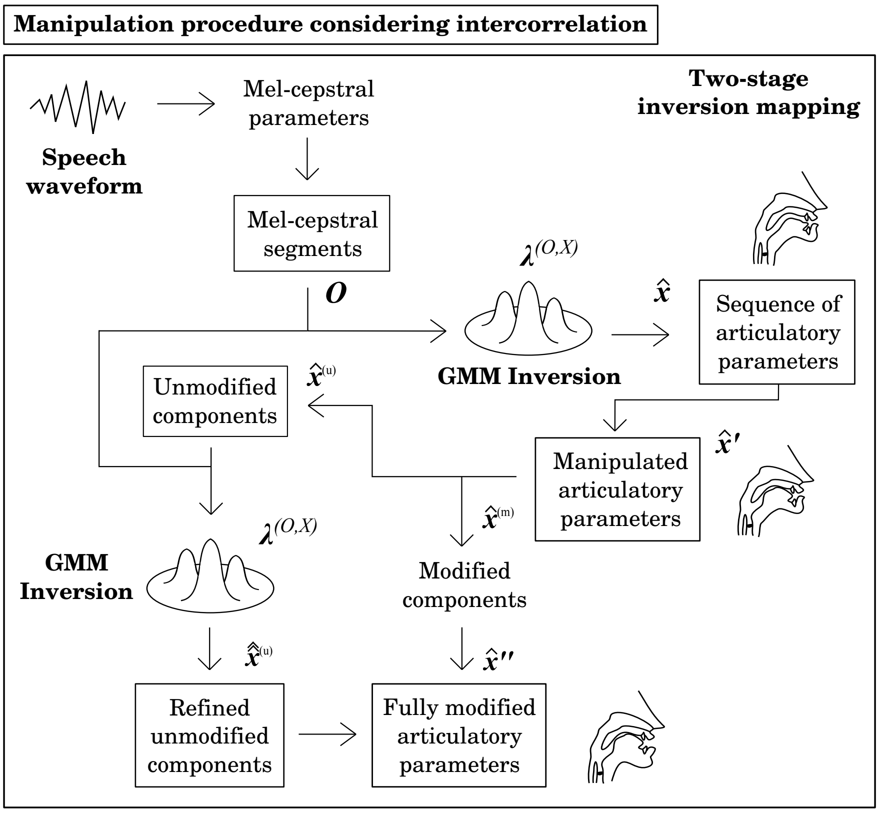

Manipulation procedure considering intercorrelation
Speech waveform
Mel-cepstral parameters
Mel-cepstral segments
Two-stage inversion mapping
$\boldsymbol{\lambda}^{(O,X)}$
$\boldsymbol{O}$
$\hat{\boldsymbol{x}}$
Sequence of articulatory parameters
GMM Inversion
Unmodified components
$\hat{\boldsymbol{x}}^{(u)}$
Manipulated articulatory parameters
$\hat{\boldsymbol{x}}'$
$\hat{\boldsymbol{x}}^{(m)}$
$\boldsymbol{\lambda}^{(O,X)}$
GMM Inversion
Modified components
$\hat{\hat{\boldsymbol{x}}}^{(u)}$
$\hat{\boldsymbol{x}}''$
Refined unmodified components
Fully modified articulatory parameters

Figure 3.2: *Flow of the articulatory manipulation procedure that considers the intercorrelation of articulatory parameters by performing a two-stage inversion mapping.*

where the scaling matrix $\boldsymbol{\Lambda}_t$ and the shifting vector $\boldsymbol{\psi}_t$ are respectively written as

$$\boldsymbol{\Lambda}_t = \operatorname{diag}\left[\Lambda_t(1), \ldots, \Lambda_t(d), \ldots, \Lambda_t(D_x)\right], \tag{3.2}$$

$$\boldsymbol{\psi}_t = \left[\psi_t(1), \ldots, \psi_t(d), \ldots, \psi_t(D_x)\right]^\top. \tag{3.3}$$

Through the use of scaling factors in $\boldsymbol{\Lambda}_t$ at each frame t, the dilation or contraction of articulatory movements can be managed since Z-scores are used, i.e., the mean of the articulatory trajectory is not changed. On the other hand, the positions of individual articulators can be conveniently altered by using the shifting factors in $\boldsymbol{\psi}_t$ at each frame t.

By using the above linear transformation, the unobserved articulatory movements can be modified by manipulating the parameters of individual articulators with a set of scaling and shifting factors. However, it is known that these articulators have a certain degree of correlation between each other [99], for example, the movements of the tongue tip strongly affect those of the tongue body. Therefore, considering this fact, the manual manipulation of particular articulators must be compensated by the other articulators [33] by considering the degree of their correlation. Hence, unnatural articulatory movements are likely to be generated from this simple manipulation method.

Manipulation procedure considering intercorrelation of articulatory parameters

To generate more natural trajectories of modified articulatory movements, a manipulation procedure that takes into account the intercorrelation of articulatory parameters is presented. To achieve this, specifically, a scheme that can be called a two-stage inversion mapping strategy is used. In the first stage, a sequence of articulatory parameters is estimated given the corresponding sequence of mel-cepstral segments. Then, a simple linear transformation is performed to manipulate these articulatory parameters as described in Section 3.3.2. In the second stage, the modified components of the articulatory parameters are appended onto the input mel-cepstral segments. Then, a set of refined parameters corresponding to the unmodified articulatory components is estimated by utilizing the intercorrelation of articulatory parameters embedded within the GMM of the inversion mapping. Finally, a set of fully modified articulatory parameters is constructed from the modified components and the refined unmodified components. The flow of this manipulation procedure is depicted in Fig. 3.2.

Let $\hat{\boldsymbol{x}}_t^{(d)}$ and $\hat{\boldsymbol{x}}_t^{(u)}$ be the articulatory feature vectors of the modified components and the unmodified components, respectively, at frame t. Their joint static and dynamic feature vectors are then respectively given by $\hat{\boldsymbol{X}}_t^{(d)} = [\hat{\boldsymbol{x}}_t^{(d)^\top}, \Delta\hat{\boldsymbol{x}}_t^{(d)^\top}]^\top$ and $\hat{\boldsymbol{X}}_t^{(u)} = [\hat{\boldsymbol{x}}_t^{(u)^\top}, \Delta\hat{\boldsymbol{x}}_t^{(u)^\top}]^\top$ at frame t. Thus, by using the second stage of the inversion mapping, a sequence of refined unmodified components $\hat{\hat{\boldsymbol{x}}}^{(u)}$ can be estimated as follows:

$$\begin{aligned}\hat{\hat{\boldsymbol{x}}}^{(u)} &= \mathop{\mathrm{argmax}}_{\hat{\boldsymbol{x}}^{(u)}} P\left(\hat{\boldsymbol{X}}^{(u)}|\boldsymbol{O}, \hat{\boldsymbol{X}}^{(d)}, \hat{\boldsymbol{m}}^{(O)}, \boldsymbol{\lambda}^{(O,X)}\right), \\ &\qquad \text{subject to } \hat{\boldsymbol{X}}^{(u)} = \boldsymbol{W}^{(x^{(u)})}\hat{\boldsymbol{x}}^{(u)}, \end{aligned} \tag{3.4}$$

where the suboptimum mixture component sequence $\hat{\boldsymbol{m}}^{(O)}$ is determined by Eq. (2.3). The transformation matrix used to expand the dynamic features of the unmodified components is denoted as $\boldsymbol{W}^{(x^{(u)})}$. The corresponding sequences of articulatory feature vectors are written as $\hat{\boldsymbol{X}}^{(d)} = [\hat{\boldsymbol{X}}_1^{(d)^\top}, \ldots, \hat{\boldsymbol{X}}_T^{(d)^\top}]^\top$, $\hat{\boldsymbol{X}}^{(u)} = [\hat{\boldsymbol{X}}_1^{(u)^\top}, \ldots, \hat{\boldsymbol{X}}_T^{(u)^\top}]^\top$, and $\hat{\hat{\boldsymbol{x}}}^{(u)} = [\hat{\hat{\boldsymbol{x}}}_1^{(u)^\top}, \ldots, \hat{\hat{\boldsymbol{x}}}_T^{(u)^\top}]^\top$.

To capture the intercorrelation of articulatory parameters, first, the interdimensional correlation is taken into account with the use of mixture-dependent full-covariance matrices within the conditional pdf of the inversion mapping. Thus, the modified components of the articulatory parameters implicitly govern the change in the unmodified components through their statistical correspondence. Second, the interframe correlation of articulatory parameters is also considered owing to the use of a trajectory-based conversion framework [16], which employs an explicit relationship between the static and dynamic features [65]. Therefore, this manipulation procedure should be capable of generating natural movements of the articulatory parameters.

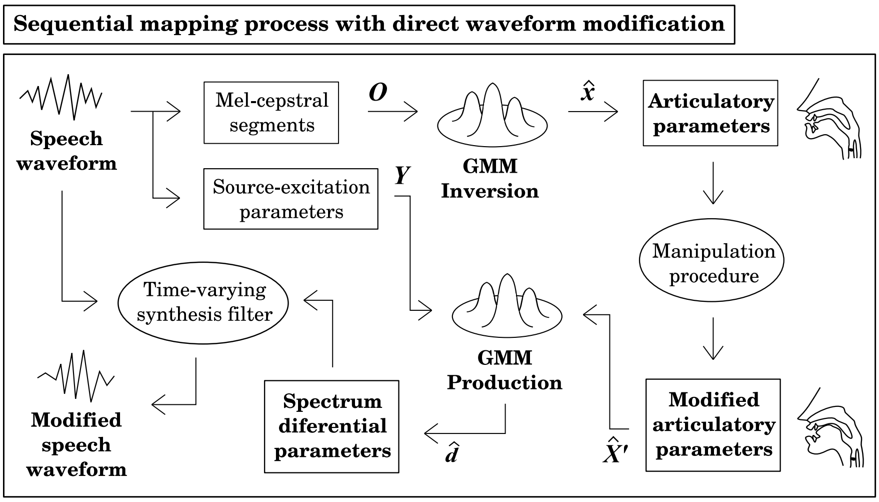

Sequential mapping process with direct waveform modification
Speech waveform
Mel-cepstral segments
Source-excitation parameters
O
Y
GMM Inversion
$\hat{x}$
Articulatory parameters
Manipulation procedure
Time-varying synthesis filter
GMM Production
Modified speech waveform
Spectrum diferential parameters
$\hat{d}$
$\hat{X}'$
Modified articulatory parameters

Figure 3.3: *Flow of the articulatory controllable speech modification system using direct waveform modification with spectrum differential parameters to generate a modified speech waveform.*

3.4 Speech Modification without Vocoder-Based Waveform Generation

3.4.1 Problem in Terms of Speech Quality

In Section 3.3, the articulatory controllable speech modification system was introduced, where an input speech signal can be conveniently modified through manipulation of the unobserved articulatory movements. In this system, to generate a modified speech waveform, a vocoder-based framework is employed, where the modified spectral envelope parameters and the source excitation parameters are used to generate the speech signal. However, it is well known that a vocoder-based procedure tends to degrade the quality of synthetic speech signals. Combined with its sensitivity to errors in speech parameterization, vocoder-based waveform generation would lead to a significant degradation in the quality of the modified speech waveform. In this work, to alleviate this problem, several implementations of a direct waveform modification procedure [96] capable of avoiding the use of a vocoder-based excitation generation scheme by using spectrum differential parameters to directly filter an input speech signal are presented. In this case, the spectrum differential parameters refer to the differences between modified and unmodified spectral envelope parameters. The flow of the sequential inversion and production mappings using the spectrum differential parameters is shown in Fig. 3.3.

By implementing a direct waveform modification procedure with spectrum differential parameters, the quality degradation caused by the use of a vocoder-based excitation generation process can be alleviated. However, in this framework, considering that the spectrum differential parameters are computed by using converted parameters, i.e., the converted spectral envelope parameters of modified speech, the quality of the modified speech waveform is still not optimized owing to the oversmoothed characteristics inherited from the trajectory-based conversion process [3]. One way to address the oversmoothing problem is by taking into account the global variance (GV) [3] and/or the modulation spectrum (MS) [100]. Nevertheless, the statistics of the GV or MS, which are obtained from the training data, do not exactly address the issue in new data. In this work, to exactly address the oversmoothing problem, two other implementations of a direct waveform modification method that can preserve the fine structure of the spectral envelope from the input speech waveform are presented.

3.4.2 Implementations of Direct Waveform Modification Method using Spectrum Differential Parameters

In a direct waveform modification method, an input speech signal is modified using the spectrum differential parameters by utilizing a time-varying synthesis filter, such as an MLSA filter [20]. To determine the best way of generating the spectrum differential parameters, three different methods are presented: a basic method (DiffBM), a refined method (DiffRM), and a refined method with differential GMM (DiffGMM).

Basic method (DiffBM)

In the basic method of calculating the spectrum differential parameters (DiffBM), extracted spectral envelope parameters of the input speech waveform and oversmoothed spectral envelope parameters of the modified speech waveform are employed. Let $\boldsymbol{c}$ be the time sequence of the spectral envelope parameters extracted from the input speech waveform and $\hat{\boldsymbol{c}}' = [\hat{\boldsymbol{c}}_1'^{\top}, \ldots, \hat{\boldsymbol{c}}_T'^{\top}]^{\top}$ be that of the oversmoothed spectral envelope parameters for the modified speech waveform. The time sequence of the DiffBM spectrum differential parameters $\boldsymbol{d}_{BM}$ is then given by

$$\boldsymbol{d}_{BM} = \hat{\boldsymbol{c}}' - \boldsymbol{c} = [[\hat{\boldsymbol{c}}_1' - \boldsymbol{c}_1]^{\top}, \ldots, [\hat{\boldsymbol{c}}_T' - \boldsymbol{c}_T]^{\top}]^{\top}. \tag{3.5}$$

A modified speech waveform is then generated by filtering the input speech waveform using the $\boldsymbol{d}_{BM}$ spectrum differential parameters. Therefore, the modified speech waveform can be characterized by a time sequence of DiffBM spectral envelope parameters $\boldsymbol{c}_{BM}$, which is given by

$$\begin{aligned} \boldsymbol{c}_{BM} &= \boldsymbol{c} + \boldsymbol{d}_{BM} \\ &= [[\boldsymbol{c}_1 + (\hat{\boldsymbol{c}}_1' - \boldsymbol{c}_1)]^{\top}, \ldots, [\boldsymbol{c}_T + (\hat{\boldsymbol{c}}_T' - \boldsymbol{c}_T)]^{\top}]^{\top} \\ &= [\hat{\boldsymbol{c}}_1'^{\top}, \ldots, \hat{\boldsymbol{c}}_T'^{\top}]^{\top}. \end{aligned} \tag{3.6}$$

Thus, the speech waveform modified by this basic method (DiffBM) is represented by a time sequence of the oversmoothed modified spectral envelope parameters $\hat{\boldsymbol{c}}'$. However, this sequence is completely different from that of the conventional vocoder-based system

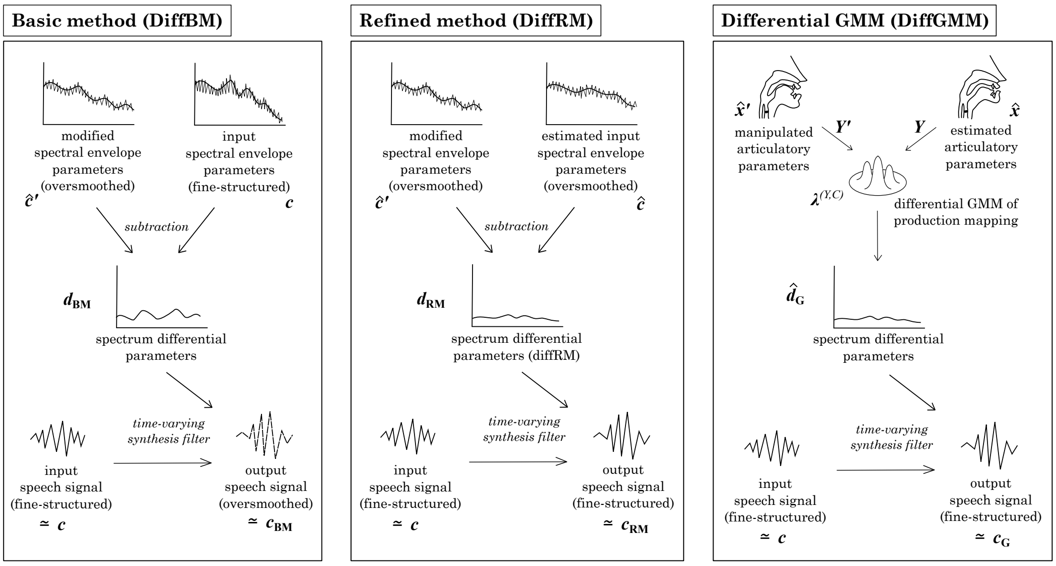

Basic method (DiffBM)
modified spectral envelope parameters (oversmoothed)
input spectral envelope parameters (fine-structured)
$\hat{\boldsymbol{c}}'$
$\boldsymbol{c}$
subtraction
$\boldsymbol{d}_{\mathrm{BM}}$
spectrum differential parameters
time-varying synthesis filter
input speech signal (fine-structured)
$\simeq \boldsymbol{c}$
output speech signal (oversmoothed)
$\simeq \boldsymbol{c}_{\mathrm{BM}}$
Refined method (DiffRM)
modified spectral envelope parameters (oversmoothed)
estimated input spectral envelope parameters (oversmoothed)
$\hat{\boldsymbol{c}}'$
$\hat{\boldsymbol{c}}$
subtraction
$\boldsymbol{d}_{\mathrm{RM}}$
spectrum differential parameters (diffRM)
time-varying synthesis filter
input speech signal (fine-structured)
$\simeq \boldsymbol{c}$
output speech signal (fine-structured)
$\simeq \boldsymbol{c}_{\mathrm{RM}}$
Differential GMM (DiffGMM)
$\hat{\boldsymbol{x}}'$
manipulated articulatory parameters
$\boldsymbol{Y}'$
$\boldsymbol{Y}$
$\hat{\boldsymbol{x}}$
estimated articulatory parameters
$\boldsymbol{\lambda}^{(Y,C)}$
differential GMM of production mapping
$\hat{\boldsymbol{d}}_{\mathrm{G}}$
spectrum differential parameters
time-varying synthesis filter
input speech signal (fine-structured)
$\simeq \boldsymbol{c}$
output speech signal (fine-structured)
$\simeq \boldsymbol{c}_{\mathrm{G}}$

Figure 3.4: *Three different flows for the implementation of a direct waveform modification procedure in the articulatory controllable speech modification system according to the calculation scheme for the spectrum differential parameters.*

in terms of the excitation signal. This is because the direct filtering procedure of the input speech waveform avoids the use of vocoder-based excitation generation. The DiffBM scheme is shown on the left panel in Fig. 3.4.

Refined method to alleviate oversmoothing (DiffRM)

In the refined method for calculating the spectrum differential parameters (DiffRM), the oversmoothing problem, which still appears in the basic method DiffBM, is alleviated by preserving the fine structure of the input speech waveform by employing oversmoothed spectral envelope parameters of both modified speech and unmodified speech waveforms. Let $\hat{\boldsymbol{c}}'$ be the time sequence of the oversmoothed spectral envelope parameters of the modified speech waveform and $\hat{\boldsymbol{c}} = [\hat{\boldsymbol{c}}_1^\top, \ldots, \hat{\boldsymbol{c}}_T^\top]^\top$ be that of the unmodified speech waveform. The time sequence of the DiffRM spectrum differential parameters $\boldsymbol{d}_{RM}$ is given by

$$\boldsymbol{d}_{RM} = \hat{\boldsymbol{c}}' - \hat{\boldsymbol{c}} = [[\hat{\boldsymbol{c}}_1' - \hat{\boldsymbol{c}}_1]^\top, \ldots, [\hat{\boldsymbol{c}}_T' - \hat{\boldsymbol{c}}_T]^\top]^\top, \tag{3.7}$$

where $\hat{\boldsymbol{c}}$ is given in Eq. (2.7).

Similarly to in the basic method, the modified speech waveform is generated by filtering the input speech waveform using the $\boldsymbol{d}_{RM}$ spectrum differential parameters. Thus, this modified speech waveform can be characterized by a time sequence of DiffRM spectral envelope parameters $\boldsymbol{c}_{RM}$, which is given by

$$\begin{aligned} \boldsymbol{c}_{RM} &= \boldsymbol{c} + \boldsymbol{d}_{RM} \\ &= [[\boldsymbol{c}_1 + (\hat{\boldsymbol{c}}_1' - \hat{\boldsymbol{c}}_1)]^\top, \ldots, [\boldsymbol{c}_T + (\hat{\boldsymbol{c}}_T' - \hat{\boldsymbol{c}}_T)]^\top]^\top \\ &= [[\hat{\boldsymbol{c}}_1' + \boldsymbol{\epsilon}_1]^\top, \ldots, [\hat{\boldsymbol{c}}_T' + \boldsymbol{\epsilon}_T]^\top]^\top, \end{aligned} \tag{3.8}$$

where the refining factors are $\boldsymbol{\epsilon}_t = \boldsymbol{c}_1 - \hat{\boldsymbol{c}}_1$ at frame t. Hence, the modified speech waveform of the refined method (DiffRM) is represented not only by a time sequence of the oversmoothed modified spectral envelope parameters $\hat{\boldsymbol{c}}'$ but also by the residuals given in a time sequence of the refining factors $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_T^\top]^\top$ to preserve the fine structure of the spectral envelope. Consequently, the oversmoothed characteristic of the modified speech waveform is alleviated. The DiffRM scheme is shown in the middle panel in Fig. 3.4.

Refined method with differential GMM (DiffGMM)

Finally, a method that works in a similar way to in the refined method DiffRM but in a more sophisticated manner by utilizing a differential GMM (DiffGMM) is presented. In this method, rather than generating the spectral envelope parameters twice, as in the DiffRM method, they are generated only once using the differential

GMM of the production mapping. Let $\hat{\boldsymbol{Y}}' = [\hat{\boldsymbol{Y}}_1'^\top, \ldots, \hat{\boldsymbol{Y}}_T'^\top]^\top$ be the time sequence of the source excitation parameters and the modified articulatory parameters, and $\hat{\boldsymbol{Y}} = [\hat{\boldsymbol{Y}}_1^\top, \ldots, \hat{\boldsymbol{Y}}_T^\top]^\top$ be that of the unmodified articulatory parameters. At frame t, their corresponding feature vectors are respectively given by $\hat{\boldsymbol{Y}}_t' = [\hat{\boldsymbol{x}}_t'^\top, \boldsymbol{s}_t^\top, \Delta\hat{\boldsymbol{x}}_t'^\top, \Delta\boldsymbol{s}_t^\top]^\top$ and $\hat{\boldsymbol{Y}}_t = [\hat{\boldsymbol{x}}_t^\top, \boldsymbol{s}_t^\top, \Delta\hat{\boldsymbol{x}}_t^\top, \Delta\boldsymbol{s}_t^\top]^\top$. Then, the time sequence of DiffGMM spectrum differential parameters $\hat{\boldsymbol{d}}_G$ is estimated as follows:

$$\begin{aligned}\hat{\boldsymbol{d}}_G = \operatorname*{argmax}_{\boldsymbol{d}_G} P\left(\boldsymbol{D}_G | \boldsymbol{Y}', \boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}\right), \\ \text{subject to } \boldsymbol{D}_G = \boldsymbol{C}' - \boldsymbol{C} \text{ and } \boldsymbol{D}_G = \boldsymbol{W}^{(c)}\boldsymbol{d}_G. \end{aligned} \tag{3.9}$$

Then, a modified speech waveform of the DiffGMM method is generated by filtering the input speech waveform using the time sequence of the estimated spectrum differential parameters $\hat{\boldsymbol{d}}_G = [\hat{\boldsymbol{d}}_{G_1}^\top, \ldots, \hat{\boldsymbol{d}}_{G_T}^\top]^\top$ as follows:

$$\boldsymbol{c}_G = \boldsymbol{c} + \hat{\boldsymbol{d}}_G. \tag{3.10}$$

Therefore, the corresponding modified speech waveform is characterized by the time sequence of spectral envelope parameters $\boldsymbol{c}_G$, where the oversmoothed structure has been alleviated by preserving the fine structure of the input speech waveform because $\hat{\boldsymbol{d}}_{G_t} = \hat{\boldsymbol{c}}_t' - \hat{\boldsymbol{c}}_t$. However, the procedure is different from that of the refined method DiffRM because the parameters are only generated once using the differential GMM of the production mapping. Furthermore, it would also be straightforward to apply additional techniques, such as GV [101] or MS modeling [100]. The DiffGMM scheme is shown on the right panel in Fig. 3.4.

3.5 Experimental Evaluation

3.5.1 Experimental Conditions

The parallel acoustic-articulatory data provided in MOCHA, described in Section 3.2 was used. As the spectral envelope parameters, the first through 24^{th} mel-cepstral coefficients converted from the spectral envelope were used, which were extracted frame-by-frame by STRAIGHT analysis [13]. As the source excitation parameters, log-scaled F_0 values also including an unvoiced/voiced binary decision feature and log-scaled power values extracted from the STRAIGHT spectrum were used. The fixed-point analysis [102] in STRAIGHT was used to extract the F_0 values. As the articulatory parameters, 14-dimensional EMA data, elaborated in Section 3.2, were used, which were converted into Z-scores. The speech data were sampled at 16 kHz. The frame shift was set to 5 ms. The contextual frame length in Eq. (2.1) was set to ±10 frames.

Both objective and subjective evaluations to assess the performance of the articulatory controllable speech modification system were performed. In the objective evaluation, first, the accuracy of the inversion mapping, described in Section 2.2.3, was measured by comparing the estimated articulatory parameters with the measured values. Then, the accuracy of the production mapping, described in Section 2.2.3, was measured by comparing the estimated spectral envelope parameters, converted from the measured articulatory parameters, with the extracted spectral envelope parameters. Finally, the accuracy of the sequential procedure of inversion and production mappings, described in Section 3.3.1, was measured by comparing the estimated spectral envelope parameters, converted from the estimated articulatory parameters, with the extracted spectral envelope parameters. On the other hand, in the subjective evaluation, both the quality of the generated speech sounds and the controllability of the system were evaluated. In the first subjective evaluation of the speech quality, the performance of the methods for manipulating articulatory parameters, described in Section 3.3.2, is compared in terms of the naturalness of the modified speech sounds. Then, in the second subjective evaluation, the performance of the implementations of the direct waveform modification method, described in Section 3.4, which avoids the use of a vocoder-based procedure, is compared by examining the quality of modified speech waveforms under several speaking conditions. Finally, the controllability of the system was evaluated by a categorical perception evaluation in which several vowel sounds were modified by manipulating the articulatory positions.

3.5.2 Objective Evaluation

Accuracy of inversion mapping

To measure the accuracy of the acoustic-to-articulatory inversion mapping described in Section 2.2.3, first, the root-mean-square (RMS) error of the estimated articulatory parameters relative to the measured values was calculated as follows:

$$\mathrm{RMSE}(d) = \sqrt{\frac{\sum_{t=1}^{T}(a_t^{(o)}(d) - a_t^{(e)}(d))^2}{T}}, \tag{3.11}$$

where RMSE(d) is the RMS error for the dth dimension of the articulatory parameters. The measured and estimated dth dimension articulatory parameters are respectively denoted as $a_t^{(o)}(d)$ and $a_t^{(e)}(d)$ at frame t. The lowest errors of 1.42 mm and 1.41 mm were achieved by using 128 mixture components for both male and female speakers, respectively. This result is consistent with the related work in [16].

Secondly, the correlation coefficient was measured, which was also calculated between the estimated and measured articulatory parameters as follows:

$$r(d)=\frac{\sum_{t=1}^{T}(a_t^{(o)}(d)-\hat{a}_t^{(o)}(d))(a_t^{(e)}(d)-\hat{a}_t^{(e)}(d))}{\sqrt{\sum_{t=1}^{T}(a_t^{(o)}(d)-\hat{a}_t^{(o)}(d))^2\sum_{t=1}^{T}(a_t^{(e)}(d)-\hat{a}_t^{(e)}(d))^2}}, \tag{3.12}$$

where $r(d)$ is the correlation coefficient for the dth dimension of the articulatory parameters. The mean values of the measured and estimated dth dimension articulatory parameters are respectively denoted as $\hat{a}_t^{(o)}(d)$ and $\hat{a}_t^{(e)}(d)$ at frame t. The highest correlation coefficients of 0.79 and 0.80 were yielded by using 128 mixture components for both male and female speakers, respectively. This result is also consistent with [16].

Accuracy of production mapping

To measure the accuracy of the articulatory-to-acoustic production mapping, described in Section 2.2.3, the mel-cepstral distortion between the estimated mel-cepstral parameters and the extracted values was calculated as follows:

$$\text{Mel-CD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(c^{(o)}(d)-c^{(e)}(d))^2}, \tag{3.13}$$

where $c^{(o)}(d)$ and $c^{(e)}(d)$ denote the dth dimension of the extracted and estimated mel-cepstral parameters, respectively. The final result was averaged over all samples of training data and over all 24 dimensions of mel-cepstral parameters. The lowest mel-cepstral distortion values of 4.70 dB and 4.94 dB were achieved by using 64 mixture components for both male and female speakers, respectively, which are also comparable results to those in [16].

Accuracy of sequential procedure of inversion and production mappings

To assess the effectiveness of the sequential inversion and production mappings, described in Section 3.3.1, the mel-cepstral distortion between estimated mel-cepstral parameters and extracted mel-cepstral parameters was also measured. The estimated mel-cepstral parameters were converted from the estimated articulatory parameters with the sequential inversion-production mapping. Furthermore, the mel-cepstral distortion results yielded by using the GMM of the production mapping trained with the estimated articulatory training data instead of the measured articulatory training data was also measured.

The lowest distortion values of 4.38 dB and 4.65 dB were achieved by using 128 mixture components for both male and female speakers, respectively. This improvement was achieved because the estimated articulatory parameters are the most likely ones to be converted from the input mel-cepstral parameter sequence. Therefore, by estimating the input mel-cepstral parameters, which is performed in the sequential inversion-production mapping, one can evaluate the appropriateness of this mapping procedure. On the other hand, by using the estimated articulatory training data to train the GMM production used in the sequential mapping, the lowest values of 3.99 dB and 4.20 dB were achieved by using 64 mixture components for both male and female speakers, respectively. In the following experiments for subjective evaluation, the GMMs for the sequential inversion and production mappings with 128 mixture components were used.

3.5.3 Subjective Evaluation of Speech Quality

Comparison of articulatory manipulation methods

In the first subjective evaluation of the speech quality, the performance of the methods for manipulating articulatory parameters in Section 3.3.2 is compared. To do this, the scaling factors of the tongue tip movements on the y-axis were modified using five scaling values from 1.0-fold to 5.0-fold (hyperarticulation can be achieved by exagger-

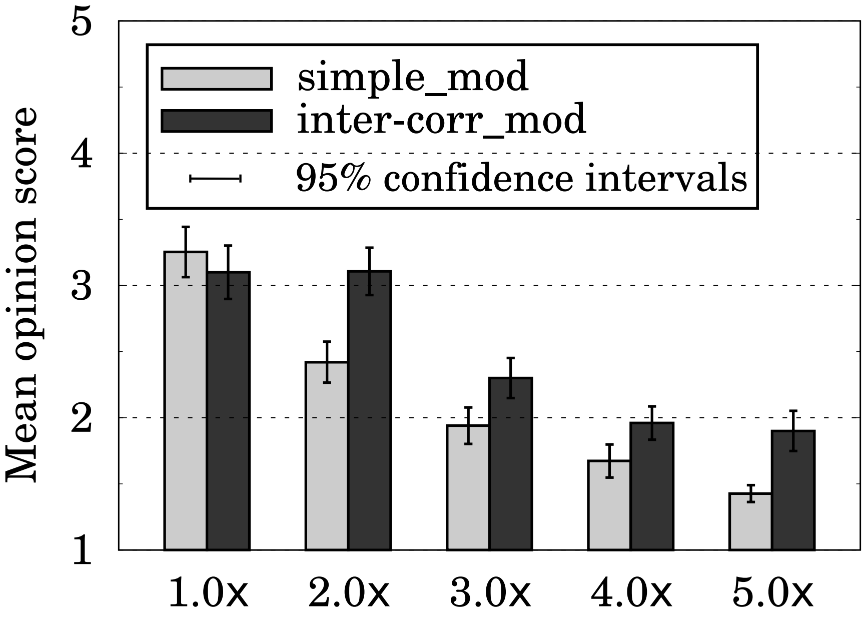


Figure 3.5: *Mean opinion score (MOS) results of male speaker (msak0) for the evaluation of modified speech quality by scaling the tongue tip movements in y-coordinate with two different methods of articulatory parameter manipulation.*

ating the articulatory motions with value more than 1.0, while hypoarticulation by dimishing their range with value less than 1.0). Listeners were asked to evaluate the quality of the modified speech sounds in a mean opinion score (MOS) evaluation using a range of scores from 1.0 to 5.0, where 5.0 was the highest. The number of listeners was 10. The number of distinct utterances per listener was 15, which were randomly taken from the 110 evaluation data.

The MOS results for the different methods of manipulating the articulatory parameters are shown in Figs. 3.5 and 3.6 for the male and female speaker, respectively. The results show that the method considering the intercorrelation of articulatory parameters, described in Section 3.3.2, gives higher scores than the simple linear transformation method for both of the male and female speaker data. It can also be observed that

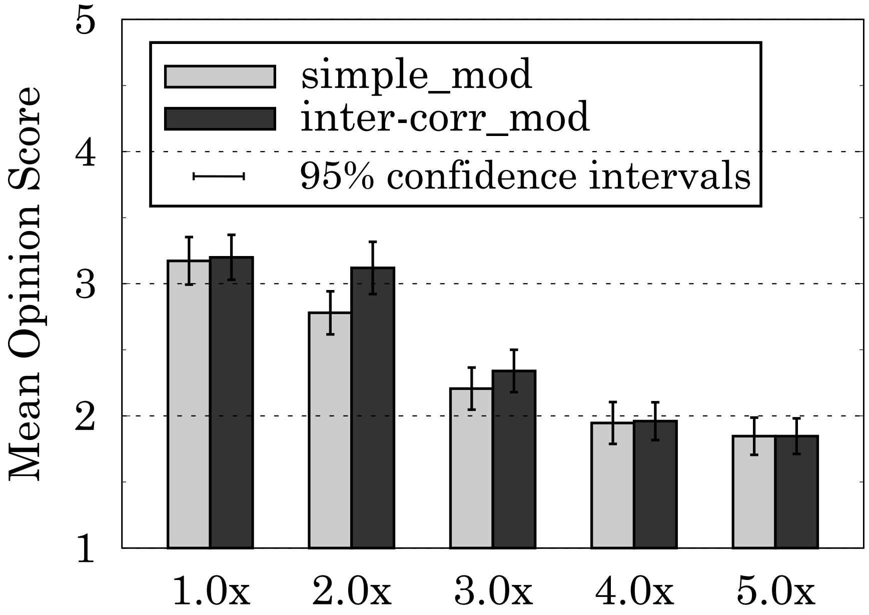
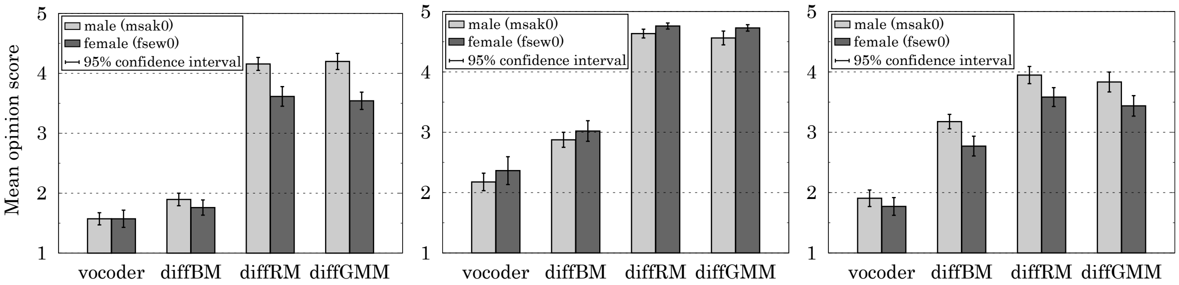


Figure 3.6: *Mean opinion score (MOS) results of female speaker (fsew0) for the evaluation of modified speech quality by scaling the tongue tip movements in y-coordinate with two different methods of articulatory parameter manipulation.*

this method still preserves the quality of the modified speech up to a scaling value of 2.0-fold, which implies that higher values would lead to possible abnormalities caused by the physical constraints within the vocal tract being exceeded.

Comparison of spectrum differential calculation methods

In the second subjective evaluation of speech quality, the performance of the implementations of the direct waveform modification method in Section 3.4, which avoids the use of a vocoder-based speech generation framework to alleviate the quality degradation of synthetic speech, is compared. To do this, three speaking conditions were emulated by scaling the trajectories of articulatory movements: normal articulation, hypoarticulation, and hyperarticulation. Naturally, prosodic elements, such as the



(a) *Hypoarticulation speaking condition (*0.5*-fold scaling)* (b) *Normal articulation speaking condition (*1.0*-fold scaling)* (c) *Hyperarticulation speaking condition (*2.0*-fold scaling)*

Figure 3.7: *MOS results for the evaluation of the modified speech quality under three speaking conditions using the vocoder-based framework and three implementations of the direct waveform modification method.*

speaking rate and glottal stops, are included in the characterization of speaking conditions [103]. However, because the spectral characteristic is the characteristic most closely related to the phonetic quality, it can be relatively easily modified by manipulating the articulatory movements. To accomplish this, 1.0-fold scaling to emulate the normal articulation condition, 0.5-fold scaling for hypoarticulation, and 2.0-fold scaling for hyperarticulation were used. A MOS evaluation was conducted to assess the quality of modified speech sounds, with a range of scores from 1.0 to 5.0. Four different speech generation procedures were compared: the vocoder-based method, the basic method of direct waveform modification (DiffBM), the refined method (DiffRM), and the refined method with a differential GMM (DiffGMM). The number of listeners was 12. The number of distinct utterances was 8.

The MOS results of the male and female speakers are shown in Fig. 3.7. These results demonstrate that the implementations of the direct waveform modification method, particularly the refined method (DiffRM) and the refined method with the differential GMM (DiffGMM), significantly improve the quality of the modified speech over all

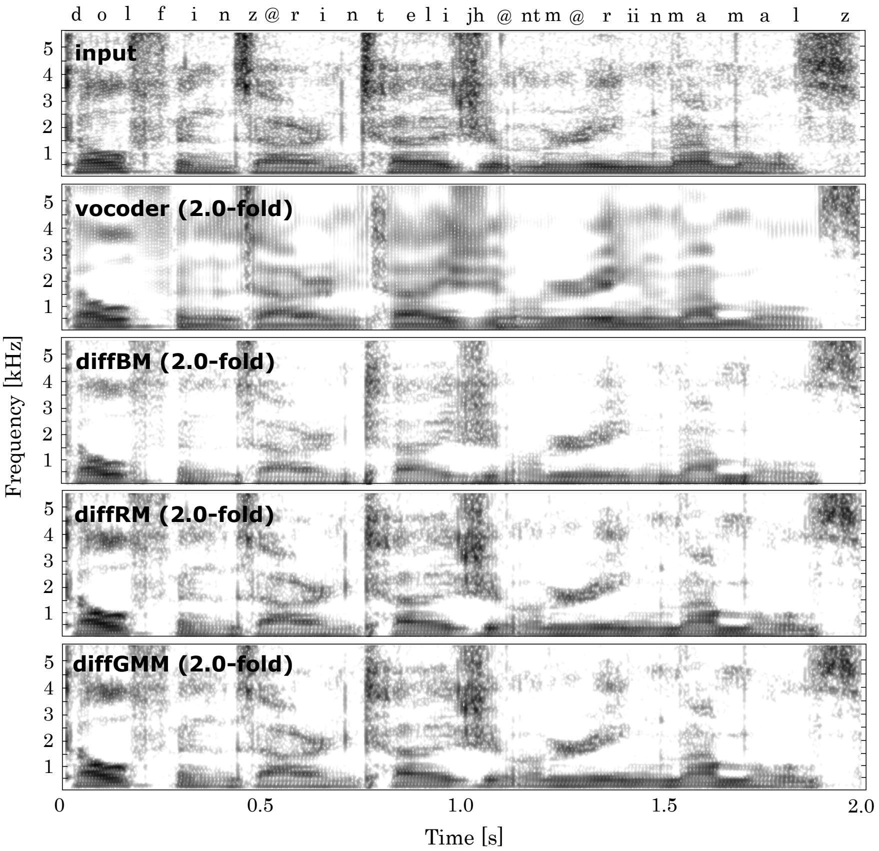


Figure 3.8: *Spectrograms of modified speech waveform under hyperarticulation (*2.0*-fold) condition using vocoder-based procedure (second top) and the three direct waveform modification schemes for the utterance "Dolphins are intelligent marine mammals" from the male speaker, while the spectrogram of input speech waveform is shown at the top.*

three speaking conditions for both male and female speakers. On the other hand, the basic method (DiffBM) yields only a small improvement from the conventional vocoder-based method compared with the DiffRM and DiffGMM. This is because, even though the vocoder-based excitation generation procedure is avoided, the overall structure of the generated speech waveform still inherits the oversmoothed characteristics, which

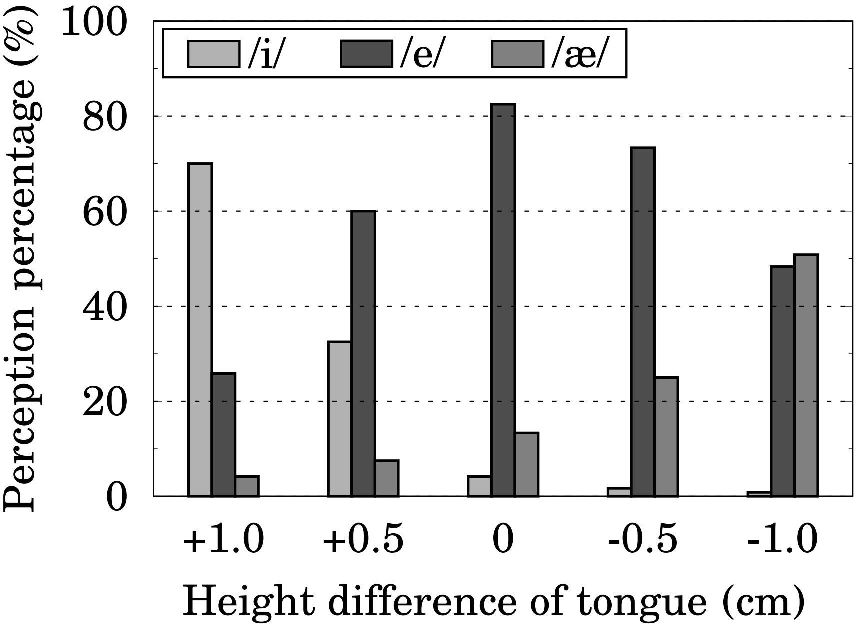


Figure 3.9: *Categorical perception results for evaluation of controllability of the system by modification of vowel sounds for male speaker.*

are alleviated in the DiffRM and DiffGMM methods. Higher scores yielded from the male speaker may have been caused by the higher accuracy in the estimation of speech spectrum for the male speaker than the female speaker. Spectrograms of a sample utterance, i.e., “Dolphins are intelligent marine mammals”, from the male speaker generated using all four speech generation procedures in the hyperarticulation speaking condition (2.0-fold scaling) are shown in Fig. 3.8. It can be observed that at higher frequency bands, a strong periodic structure is generated by the vocoder-based method, while the DiffBM method is capable of preserving the more natural aperiodic structure, which is further refined by using either DiffRM or DiffGMM.

3.5.4 Subjective Evaluation of Controllability

In the final subjective evaluation, the controllability of the articulatory controllable speech modification system by controlling several vowel sounds through the manipulation of articulatory positions was assessed. Specifically, three front vowels in English were modified, i.e., /ɪ/, /ɛ/, and /æ/. The prominent difference between these vowels in terms of articulation is in the height of the tongue during their pronunciation. For vowel /ɪ/, the tongue is located at the highest position among the three vowels, that for vowel /æ/ is located at the lowest position, and that for vowel /ɛ/ is located at an intermediate position. To simulate these conditions, the tongue position was set at the middle frame of vowel /ɛ/ at five different positions relative to the original position: +1.0 cm, +0.5 cm, 0 cm, −0.5 cm, and −1.0 cm. A more positive value means that the position is higher. The modified middle-frame position of the tongue height was interpolated to the middle-frame configurations of its surrounding left and right phonemes by using cubic-spline interpolation to ensure a smooth trajectory for the modified articulatory positions. In total, 10 distinct words containing the vowel /ɛ/ excerpted from the evaluation data were chosen. The evaluation involved a categorical perception procedure. Each of the modified speech samples of the chosen words was presented to the listeners, along with a label showing its written word including the modified vowel, which was written with a question mark. Each of the listeners was asked to guess the missing vowel, either /ɪ/, /ɛ/, or /æ/. The total number of listeners was 10 none of which were native English speakers. The refined direct waveform modification method with the differential GMM (DiffGMM) was used to generate the speech sounds. Frames corresponding to the target vowels, i.e., /ɪ/ and /æ/, were removed when training the GMMs.

Results of the categorical perception evaluation for both male and female speakers

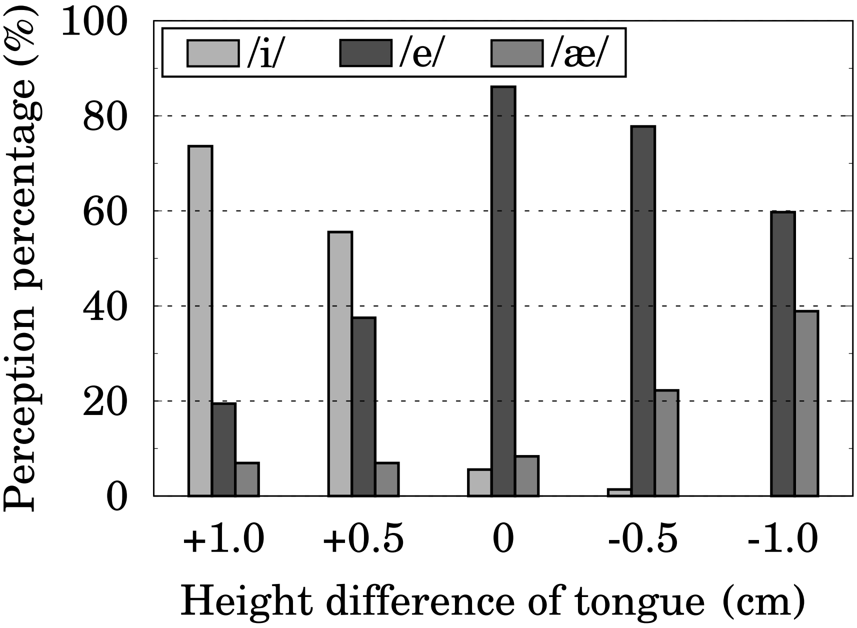


Figure 3.10: *Categorical perception results for evaluation of controllability of the system by modification of vowel sounds for female speaker.*

are respectively shown in Fig. 3.9 and Fig. 3.10. It can be observed that the differences in the articulatory position configuration indeed lead to a change in the perception of the vowel sounds, as has also been observed in [17]. This is shown by the high gradient of the perception for vowel /ɪ/ as the position of the tongue becomes higher and the moderate gradient for vowel /æ/ as the position becomes lower. This difference is most likely caused by the relatively similar spectral characteristics of vowels /ɪ/ and /æ/, considering that none of the listeners were native English speakers. Samples of linear predictive coding (LPC) spectrums containing the first three formants of the modified vowel for the word “stems” from the male speaker are shown in Fig. 3.11. These spectrums suggest the consistency of the formant characteristics of these three vowels, where vowel /ɪ/ has the lowest F1 and highest F2, vowel /æ/ has the highest F1 and lowest F2, and vowel /ɛ/ has intermediate values for both formants. Moreover, comparison of global variance (GV) of the mel-cepstral parameters, over all utterances containing the modified vowels, generated with the vocoder-based and the all three spectrum differential system, is given in Fig. 3.12. The graph shows that the GVs computed from the diffRM and diffGMM methods are very close to those of the original waveform, in which high-quality modified speech sounds can be confirmed. Note that because of the removal of the frames corresponding to the target vowels in the training procedure, this implies that alterations in articulatory configurations that are not yet known/learned can still lead to the appropriate production of intended speech sounds. Applications in which articulatory parameters are deliberately modified for effect, e.g. for language learning or speech therapy, rely on the underlying waveform generation to be well behaved when pushed into configurations outside of that found in the training data. This experiment can be considered as an important test of the robustness of the approach.

3.6 Discussion

To make the system more viable, it should not rely on only a limited amount of speaker characteristics. However, it is to be understood that the development of articulatory data is not a straightforward procedure. Therefore, an approach that could take advantage of the available speech and articulatory data to adapt with arbitrary speaker characteristics is to be considered. Such approach would alleviate the need in collecting articulatory data of a desired new speaker. One effective way of doing it is to use the vocal tract length normalization (VTLN) technique [104]. In this method, to compensate the difference in vocal tract lengths, a frequency warping function is employed in warping the frequency spectrum before computing the cepstral parameters. Hence, the acoustic space of available speakers can be adapted into the acoustic

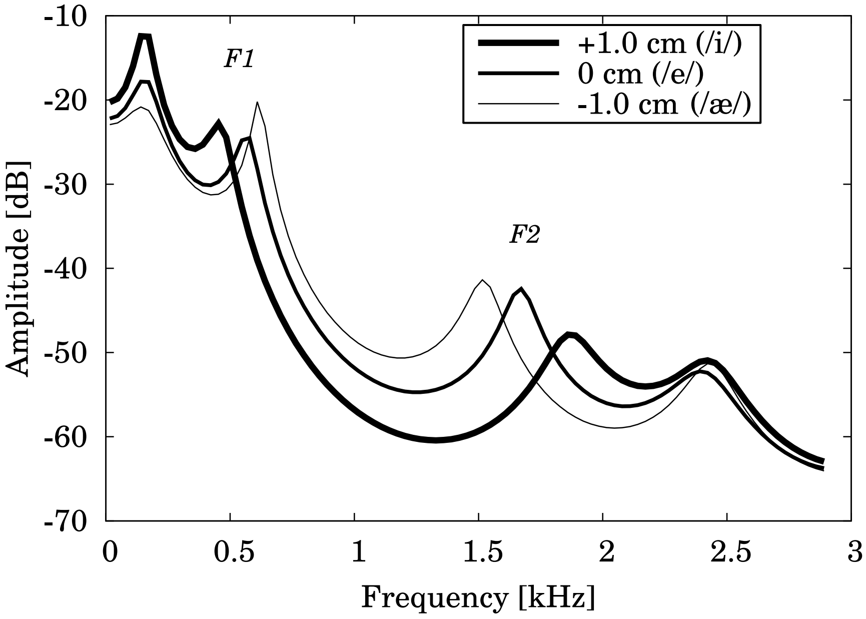


Figure 3.11: *LPC spectrums containing the first three formants from the speech section of the modified vowel for the word "stems" from the male speaker with three different relative heights of the tongue.*

space of the desired speaker. By using the trained mapping models with the adapted acoustic space, given an unseen speech data of the target speaker, its corresponding articulatory movements can be estimated. Note that their movements would imitate the articulatory configurations of the trained speakers, as it has also been implemented with a similar idea in [105]. Then, in the production mapping, given the articulatory parameters, the estimated acoustic spectrum should be warped into the spectrum of the desired speaker. An alternative way to do speaker-independent mapping is by employing the idea of eigenvoice-based voice conversion [106]. In this framework, the desired speaker characteristics can be controlled with an optimum weight set that influence the eigenvoice parameters. Even with only a limited amount of adaptation speech data from the new speaker, the optimum parameters can be easily obtained. In

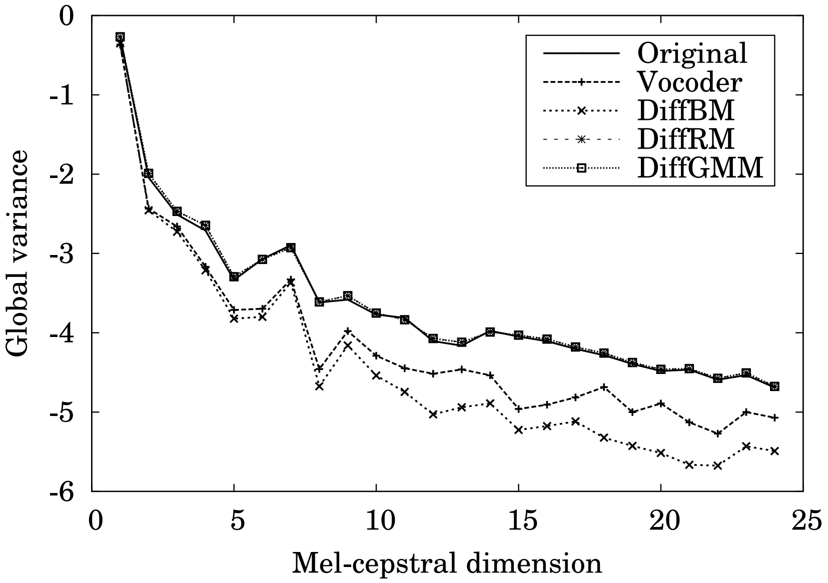


Figure 3.12: *Global variance of each mel-cepstral dimensions from all utterances containing modified vowel /ɪ/ onto /ɛ/ by leveraging the tongue height by* 1 *cm using the vocoder-based system and the three proposed spectrum differential systems. Global variances of the original waveform are also presented.*

the implementation, the idea would be similar to that of the VTLN one, but instead of using the frequency warping procedure, the voice conversion procedure would be used.

3.7 Conclusion

In this work, a successful way of exploiting articulatory parameters through an articulatory controllable speech modification system with a sequential procedure of inversion and production mappings has been presented. GMM-based statistical feature mapping technique is employed for each of the acoustic-to-articulatory inversion mapping and the articulatory-to-acoustic production mapping. An input speech signal is modified through manipulation of the unobserved articulatory movements with the use of sequential inversion and production mappings. In controlling the movements of the articulators, a method that considers the intercorrelation of articulatory parameters is deployed after a simple linear transformation is performed. Furthermore, to alleviate the quality degradation of modified speech sounds, several implementations of a direct waveform modification method that avoids the use of a vocoder-based excitation generation procedure are applied. The experimental results demonstrate the following. 1) The sequential inversion and production mappings yield higher accuracy in estimating spectral envelope parameters, i.e., of 4.38 dB and 4.65 dB mel-cepstral distortions for male and female speakers, compared with 4.70 dB and 4.94 dB for a conventional production mapping that uses the measured articulatory parameters, respectively. 2) The method for manipulating articulatory parameters by considering their intercorrelation generates more natural results than a simple linear transformation method, i.e., an MOS score of 3.1 compared with a score of 2.1 for a twofold scaling value. 3) The implementations of the direct waveform modification method significantly improve the quality of the modified speech by avoiding the use of a vocoder-based excitation generation process and overcoming the oversmoothing problem by yielding MOS scores of above 3.5, while for the vocoder-based process, the MOS score was usually below 2.0. 4) The controllability of the system is ensured by its capability of allowing the modification of vowel sounds through a certain manipulation of articulatory configurations, giving averages of about 45% for correct perceptions of vowel /æ/, 85% for vowel /ɛ/, and 70% for vowel /ɪ/. In the future, corresponding speaker-independent system [104–106] for employment in speech applications with interactive operation and also for voice conversion by vocal-tract shape modeling are worth be investigated.

4 Statistical Voice Conversion with Neural Network Spectral Mapping and WaveNet Vocoder

4.1 Introduction

Every human being has their own speech characteristics. The capability of handling the speaker characteristics within a speech signal has great potential to be employed in real-world applications. Indeed, this so-called voice conversion (VC) framework has been used in several works, such as, singing voice conversion [58, 96], body-conducted speech conversion [10], speech signal recovery [5, 6], and speech modification [18]. The growing interest in VC development motivated many researchers around the world to conceive the 1$^{\text{st}}$ Voice Conversion Challenge in 2016 [107]. Following this, the voice conversion framework elaborated within this chaper was participated in the 2$^{\text{nd}}$ Challenge, i.e., the Voice Conversion Challenge (VCC) 2018 [108].

In the development of a VC system, three aspects need to be considered: the conversion of spectral parameters, the conversion of prosodic parameters, and waveform generation. In the spectral parameter conversion, many techniques based on statistical methods have been proposed, such as, codebook-based conversion [1], Gaussian mixture model (GMM)-based mapping [2], and a neural-network based system [62, 109]. On the other hand, in the handling of prosodic parameters, such as fundamental frequency

(F0), several methods have been commonly used including a simple mean/variance linear transformation, a contour-based transformation [110], GMM-based mapping [111], and neural network [112]. For waveform generation, approaches include the source-filter vocoder system [3], the latest direct waveform modification technique [96], and the use of state-of-the-art WaveNet modeling [23–25].

In this work, neural network architectures for spectral modeling as well as a WaveNet-based vocoder for waveform modeling and generation are utilized. A neural network design is adopted for spectral parameter conversion, where a structure combining a deep neural network (DNN) and a deep mixture density network (DMDN) [113] is used to form a cascaded DMDN (CascDMDN). In a conventional DNN or DMDN, given a sequence of source spectral parameters, the target sequence is estimated using a single Gaussian distribution in a DNN or using a mixture of Gaussian distributions in a DMDN. In CascDMDN, a sequence of estimated source spectral parameters is first inferred within its first set of hidden layers, which is then fed into the second set to estimate the target sequence. For the conversion of prosodic parameters, a linear transformation of framewise F0 values of the source speaker into those of the target on the basis of their mean and variance statistics is used.

In the waveform-processing module, the state-of-the-art WaveNet-based vocoder [23–25] framework to directly model the waveform is used. In WaveNet [14], each waveform sample is conditioned using previous samples and possible auxiliary features within a stack of dilated convolutional layers. The structure of the dilated convolutions makes it possible to exponentially increase the receptive field of waveform samples efficiently. In addition, in this framework, the auxiliary features include the voiced/unvoiced (U/V) decision, continuous F0 values, mel-cepstrum parameters, and aperiodicity features. To obtain the set of refined speech parameters, a postprocessing method based on direct waveform modification [96] in several analysis-synthesis flows is employed. Then, a model selection procedure is performed to select the best waveform generation flow in an utterance-wise manner. In the evaluations carried out at the VCC 2018, the voice conversion system presented in this chapter achieved second place with an average mean opinion score (MOS) of 3.44 for speech quality and 85% accuracy for speaker similarity.

The rest of the chapter is organized as follows. Spectral parameter conversion models are elaborated in Section 4.2. The waveform-processing module is described in Section 4.3. Experimental results are presented in Section 4.4. Finally, the conclusion is given in Section 4.5.

4.2 Spectral parameter conversion models

In this section, the deep learning structures used to perform spectral parameter conversion are elaborated. Their graphical model representations are illustrated in Fig. 4.1. Moreover, the overall process described in this section is illustrated in the upper diagram of Fig. 4.2.

4.2.1 Conversion model with deep mixture density network (DMDN)

This system uses a DMDN [113] in the spectral parameter conversion by inferring a mixture of pdfs of the target spectral feature vector. Given an input source feature vector $\boldsymbol{X}_t$ at frame t, the conditional pdf of the target spectral feature vector $\boldsymbol{Y}_t$ is then defined as follows:

$$P_m(\boldsymbol{Y}_t|\boldsymbol{X}_t,\boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_{m,t} P(\boldsymbol{Y}_t|\boldsymbol{\mu}_{m,t},\boldsymbol{\Sigma}_{m,t}), \tag{4.1}$$

where the time-varying target mean vector and diagonal covariance matrix are respectively denoted as $\boldsymbol{\mu}_{m,t}$ and $\boldsymbol{\Sigma}_{m,t}$ for the mth mixture component. The weight of the mth mixture component is denoted as $\alpha_{m,t}$. The total number of mixture components is M. These time-varying mixture parameters are taken from the network output $f_\lambda(\boldsymbol{X}_t) = [f_\lambda^{(\alpha_1)}(\boldsymbol{X}_t), f_\lambda^{(\mu_1)}(\boldsymbol{X}_t)^\top, f_\lambda^{(\Sigma_1)}(\boldsymbol{X}_t)^\top, \ldots, f_\lambda^{(\alpha_M)}(\boldsymbol{X}_t),$
$f_\lambda^{(\mu_M)}(\boldsymbol{X}_t)^\top, f_\lambda^{(\Sigma_M)}(\boldsymbol{X}_t)^\top]^\top$ as

$$\alpha_{m,t} = \frac{f_\lambda^{(\alpha_m)}(\boldsymbol{X}_t)}{\sum_{n=1}^{M} f_\lambda^{(\alpha_n)}(\boldsymbol{X}_t)} \tag{4.2}$$

$$\boldsymbol{\mu}_{m,t} = f_\lambda^{(\mu_m)}(\boldsymbol{X}_t) \tag{4.3}$$

$$\boldsymbol{\Sigma}_{m,t} = \mathrm{diag}\,[\exp\,(f_\lambda^{(\Sigma_m)}(\boldsymbol{X}_t))^{\circ 2}], \tag{4.4}$$

where $\circ$ denotes a Hadamard elementwise product. The DMDN spectral conversion model is represented by the middle graph in Fig. 4.1.

In the training phase, a set of updated network parameters $\hat{\boldsymbol{\lambda}}$ is estimated by backpropagating the negative log likelihood derived from the conditional pdf given in Eq. (4.1) in a similar manner to the DNN in Eq. (2.9). On the other hand, in the conversion phase using the DMDN, given a source spectral feature vector sequence $\boldsymbol{X}$, the trajectory of the target spectral parameters $\hat{\boldsymbol{y}}$ is estimated by also using the MLPG [65] procedure as follows:

$$\hat{\boldsymbol{y}} = (\boldsymbol{W}^\top \overline{\boldsymbol{\Sigma}}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{\Sigma}}^{-1} \overline{\boldsymbol{\mu}}. \tag{4.5}$$

The sequence of the target mean vectors and that of the diagonal covariance matrices

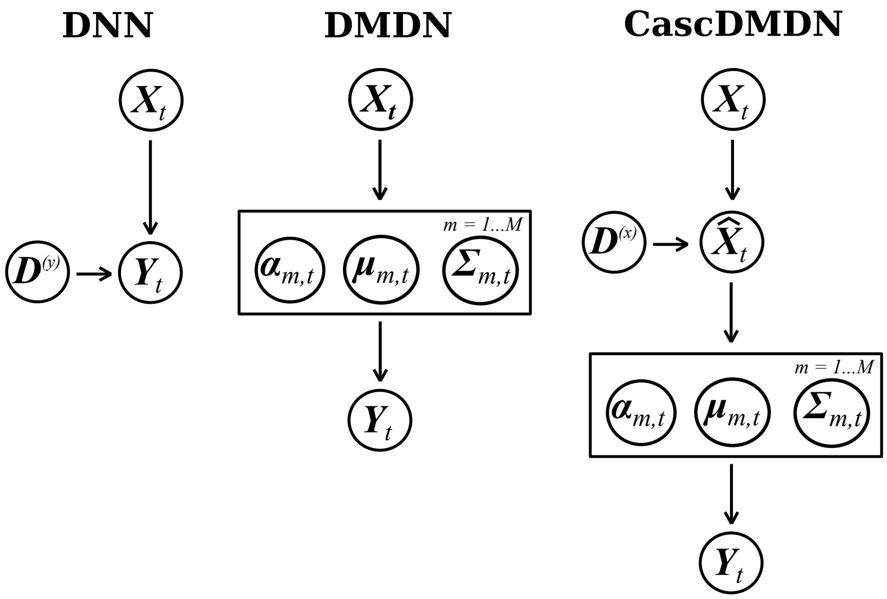


Figure 4.1: *Graphical representations of spectral conversion models using DNN, DMDN, and CascDMDN.*

in the above equation are respectively given by

$$\overline{\boldsymbol{\mu}} = [\boldsymbol{\mu}_{\hat{m}_1,1}^\top, \boldsymbol{\mu}_{\hat{m}_2,2}^\top, \ldots, \boldsymbol{\mu}_{\hat{m}_T,T}^\top]^\top \tag{4.6}$$

$$\overline{\boldsymbol{\Sigma}} = \operatorname{diag}[\boldsymbol{\Sigma}_{\hat{m}_1,1}, \boldsymbol{\Sigma}_{\hat{m}_2,2}, \ldots, \boldsymbol{\Sigma}_{\hat{m}_T,T}], \tag{4.7}$$

where the suboptimum mixture component sequence $\hat{\boldsymbol{m}} = \{\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_T\}$ is determined as follows:

$$\hat{\boldsymbol{m}} = \underset{\boldsymbol{m}}{\operatorname{argmax}} \prod_{t=1}^{T} \alpha_{m_1,t}. \tag{4.8}$$

4.2.2 Conversion model with cascaded DMDN (CascDMDN)

To develop a more flexible spectral parameter model, a cascading structure of the DNN and DMDN called the cascaded DMDN (CascDMDN) is employed. In CascDMDN, two sets of hidden layers are used, where the first set is used to estimate the pdf of the source spectral parameters and the second one is used to estimate a mixture of pdfs of the target parameters. Therefore, the conditional pdf of the target spectral feature vector is defined as follows:

$$P(\boldsymbol{Y}_t|\boldsymbol{X}_t,\boldsymbol{\lambda}) \simeq P_s(\hat{\boldsymbol{X}}_t|\boldsymbol{X}_t,\boldsymbol{\lambda}_1)P_m(\boldsymbol{Y}_t|\hat{\boldsymbol{X}}_t,\boldsymbol{\lambda}_2), \tag{4.9}$$

where the parameters of the first set are denoted as $\boldsymbol{\lambda}_1$ and those of the second one are denoted as $\boldsymbol{\lambda}_2$. The set of network parameters of CascDMDN is denoted as $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2\}$. In the above likelihood function, the first conditional pdf is similar to that of the DNN in Eq. (2.8), while the second one refers to the mixture output layer of the DMDN as in Eq. (4.1). The representation of CascDMDN is given by the right graphical model in Fig. 4.1.

In the training phase of CascDMDN, a set of updated network parameters $\hat{\boldsymbol{\lambda}}$ is estimated by backpropagating the following loss:

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} - \prod_{t=1}^{T} P_s(\boldsymbol{X}_t|\boldsymbol{X}_t,\boldsymbol{\lambda}_1)P(\boldsymbol{Y}_t|f_{\lambda_1}(\boldsymbol{X}_t),\boldsymbol{\lambda}_2). \tag{4.10}$$

Note that not only does the estimation of the source spectral feature vector in the first set give flexibility in the parameter inference, such as, for computing spectral differences between the estimated target and source spectral parameters as shown in the upper part of Fig. 4.2, but it also provides an additional regularization term in model training.

Then, in the conversion phase, given a source spectral feature vector sequence $\boldsymbol{X}$, the trajectory of the target spectral parameters $\boldsymbol{y}$ is estimated in a similar manner to the MLPG of the DMDN in Eq. (4.5), where the mixture output layer is denoted as $f_{\lambda_2}(f_{\lambda_1}(\boldsymbol{X}_1))$. Following the structure of the network, the trajectory of the source spectral parameters $\hat{\boldsymbol{x}} = [\hat{\boldsymbol{x}}_1^\top,\hat{\boldsymbol{x}}_2^\top,\dots,\hat{\boldsymbol{x}}_T^\top]^\top$ can be estimated as in the MLPG of the DNN in Eq. (2.11). In addition, the global variance (GV) [3] postfilter is applied to

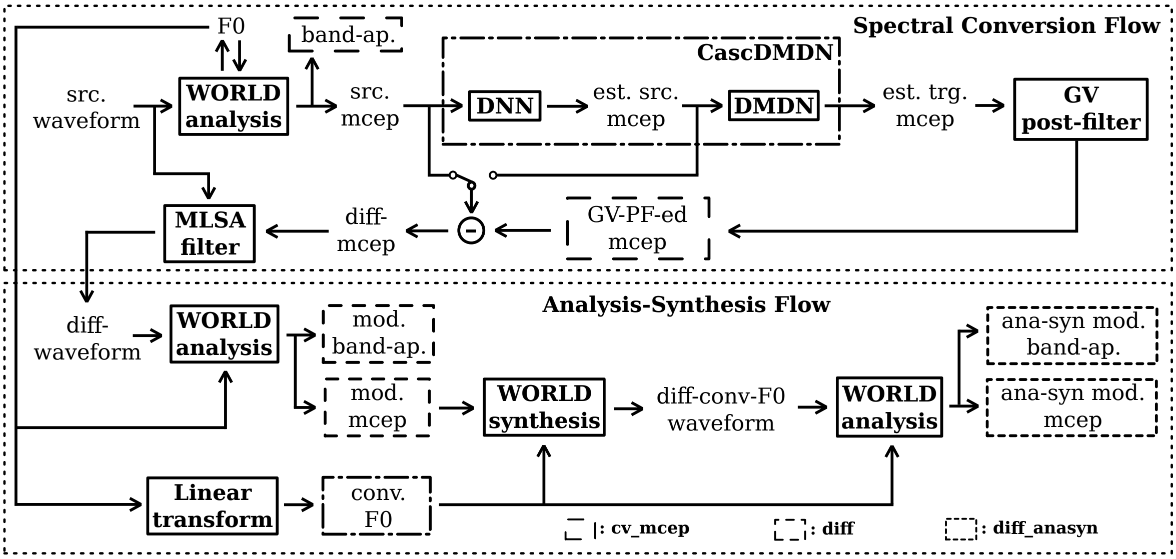


Figure 4.2: *Diagram of the spectral conversion flow (top) and the analysis-synthesis flow (bottom) to generate three different speech parameters, i.e., "cv_mcep", "diff", and "diff_anasyn", to be fed into the waveform-processing module.*

the converted spectral sequence to alleviate oversmoothed structures.

4.3 Waveform-processing module

A WaveNet-based vocoder [23–25] to model the waveform of the target speaker and generate the converted waveform using estimated speech features is used in this system. Several flows are used in producing the estimated spectral features, where the direct waveform modification [96] method is employed. In addition, a selection procedure is performed to obtain the best waveform generation flow in an utterance-wise manner.

4.3.1 Analysis-synthesis with direct waveform modification

It is well known that vocoder-based waveform generation usually causes quality degradation in the generated speech owing to the difficulty of modeling source excitation signals. To avoid this issue, the direct waveform modification (DiffVC) method [96] has been proposed to directly filter an input waveform according to spectral differences between the target waveform and the input waveform. However, because the excitation features are not converted, it is difficult to convert speaker characteristics with a large difference in prosody characteristics, such as in a cross-gender conversion. Here, an analysis-synthesis method to obtain refined spectral parameters that is based on the DiffVC method while making it possible to perform F0 conversion within the analysis-synthesis flow is described, as shown by the bottom flow in Fig. 4.2.

The analysis-synthesis procedure produces three different types of speech features to be fed into the WaveNet vocoder. The first one is called the "cv_mcep" set, which consists of the GV-postfiltered estimated target spectral parameters ("GV-PF-ed mcep") and the input band-aperiodicity features. Following this, the input waveform is then directly filtered ("diff-waveform") according to the spectral differences between GV-PF-ed mcep and the input spectral parameters. Then, by analyzing diff-waveform according to the original F0 values, the second feature set, called the "diff" set, which consists of the modified spectral and band-aperiodicity features, is obtained. Next, after performing the conversion of the F0 values, the F0-modified diff-waveform ("diff-conv-F0 waveform") is synthesized with a vocoder by using diff parameter set. Finally, the third set of parameters, called "diff_anasyn", is obtained by analyzing diff-conv-F0 waveform according to the converted F0 values. Note that the estimated target spectral parameters are generated in accordance with Section 4.2, while the converted F0 values are used by all three types of speech parameter set. Furthermore, the use of the estimated input spectral parameters to compute the spectral differences ("diff-mcep") was also investigated, as shown in the upper part of Fig. 4.2, although it did not yield significant improvements.

4.4 Experiments and results

4.4.1 Experimental conditions

The speech database for the HUB task of the VCC 2018 consisted of four source speakers and four target speakers, which had two female and two male speakers for the source and another two female and two male speakers for the target. In the training set, each speaker uttered the same set of 81 English sentences, whereas the evaluation set consisted only of the four source speakers uttering another set of 35 sentences. The speech signal sampling rate was 22,050 Hz. The WORLD [22, 114] package was used in speech analysis. From a speech signal, 35-dimensional mel-cepstrum parameters including the 0th power coefficient, F0 values, and 513-dimensional aperiodicity features, which were coded into two-band aperiodicity parameters, were used. The frame shift was set to 5 ms.

Following the spectral parameter conversion module described in Section 4.2, the DNN used four hidden layers. On the other hand, the DMDN used a total of three hidden layers and 16 mixture components. CascDMDN, which is a combination of these two structures, used one hidden layer for estimating source spectral parameters and four hidden layers with 16 mixture components for estimating target spectral parameters. ReLU activation function was used for the hidden units. For every model, the learning rate was set to 0.0006, the weights were initialized with the Xavier [115] method, the initial biases were set to zero, the Adam [116] optimization was employed, and an utterance-size batch was used.

This system used the WaveNet-based vocoder described in Section 2.4.2 for waveform modeling and generation. The hyperparameters of the WaveNet vocoder are as follows: the learning rate was set to 0.001 with a decay factor of 0.5 per 50,000 iteration steps, 20,000 batch-size samples were used with a total of 200,000 iteration steps, the number of residual blocks was 30, the dilation sequence was $1, 2, 4, \ldots, 512$ with three repetitions, the number of channels for residual blocks and dilated causal convolution was 512, the number of channels for skip connection was 256, and the Adam [116] algorithm was used for optimization. To train the WaveNet model, a speaker-independent (SI) network was first trained by using all the data of eight speakers in the HUB task plus the data of four speakers from the SPOKE task, i.e., with another 81 different sets of utterances, and the data of two speakers from the ARCTIC database, i.e., "rms" and "slt", with each having 1132 utterances. The SI-WaveNet model was then fine-tuned by updating only the output layers using the data of each of the four target speakers, which resulted in four WaveNet models.

In the waveform generation phase, the three different auxiliary features described in Section 4.3.1 were considered, i.e., cv_mcep, diff, and diff_anasyn, as shown in Fig. 4.2. The list of priorities was made heuristically, with the diff_anasyn set at the top followed by the diff set. As described in Section 2.4.3, to avoid collapsed segments in the WaveNet-generated waveforms, this system used a flow selection procedure to rule out waveforms with low quality.

The results of using mel-cepstral distortion to evaluate the spectral conversion module are given in the objective evaluation results. An internal subjective evaluation was conducted to assess the performance of the system with the provided baseline system, i.e., "sprocket" [117], where the results are given in the internal subjective evaluation section. Finally, the last three sections describe the official results of the subjective evaluation in VCC 2018.

4.4.2 Objective evaluation

To compare the several deep learning models presented in Section 4.2, an evaluation of mel-cepstral distortion was performed for the DNN, DMDN, and CascDMDN for the spectral parameter estimation. In this objective evaluation, the first 10 utterances from the training dataset were excluded while training the models. Then, they were used to compute the mel-cepstral distortion between the extracted target mel-cepstrum parameters and the estimated values as follows:

$$\text{Mel-CD[dB]}_t = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{34}(y_t(d) - \hat{y}_t(d))^2}, \tag{4.11}$$

where $y_t(d)$ and $\hat{y}_t(d)$ denote the dth dimension of the extracted mel-cepstrum parameters and that of the estimated values at frame t, respectively.

The trends of the mel-cepstral distortion averaged over all 16 speaker pairs during 70 training epochs are shown in Fig. 4.3. It can be observed that CascDMDN is capable of providing much more stable distortion than the conventional DNN and slightly more stable distortion than the conventional DMDN. The flexibility of the CascDMDN structure in providing access to estimated source spectral parameters makes a good choice for the spectral conversion model in this system. The overfitting condition observed in this objective evaluation served as a reference in training the final model using all training data.

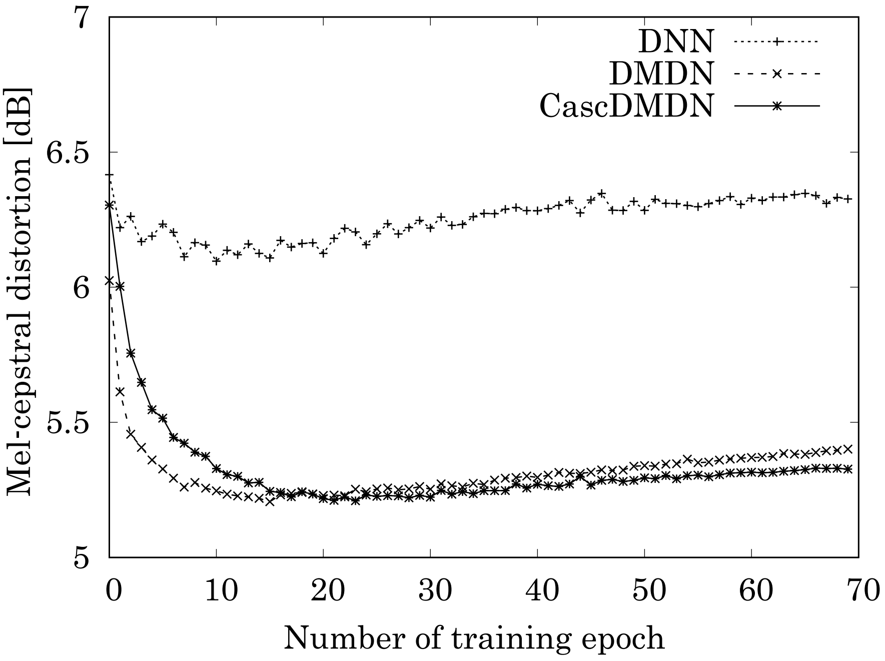


Figure 4.3: *Plot of mel-cepstral distortions for DNN, DMDN, and CascDMDN, measured with the first* 10 *utterances excluded from the training dataset.*

4.4.3 Internal subjective evaluation

In the internal subjective evaluation, two preference tests (naturalness and speaker similarity) were conducted to compare the performance of this system with that of the baseline system, i.e., sprocket [117]. All 16 speaker pair models for the four source and four target speakers were used in the evaluation. The total number of available evaluation utterances was 35. The total number of listeners was eight and none of them were native English speakers. In the naturalness test, two audio samples, one each for this system and the baseline system, of the same utterance were presented to a listener in a random order. Then, the listener was asked to select the audio preference according to naturalness. Meanwhile, in the preference test, in addition to two generated audio samples, two original audio samples of the corresponding target

Table 4.1: *Result of naturalness preference test in the internal subjective evaluation of this system (NU) and the baseline (sprocket) for same-gender and cross-gender conversions.*

Naturalness	Same-gender	Cross-gender
Baseline	68% ± 7%	42% ± 7%
NU	32% ± 7%	58% ± 7%

speaker randomly taken from the training dataset were presented. The listener was then asked to select their preference based on the similarity to the target speaker characteristic. From the 35 evaluation utterances, three were randomly taken for each speaker pair in each preference test, resulting in a total of 92 audio samples for each listener.

The results of the internal subjective evaluation are summarized in Tables 4.1 and 4.2. It can be observed that the baseline system achieves a significantly higher preference score in terms of naturalness for the same-gender conversions, with a score of 68%, compared with 32% for this system. However, this system yields a higher naturalness preference score for the cross-gender conversions, with a score of 58%, compared with 42% for the baseline. On the other hand, in the preference test for speaker similarity, this system achieves higher preference scores for both same- and cross-gender conversions, with scores of 57% and 55%, compared with 43% and 45% for the baseline, respectively. This result is reasonable because the baseline, i.e., sprocket, uses vocoder-free waveform generation for same-gender conversions and vocoder-based generation for the cross-gender conversions. This implies that the use of a WaveNet-based vocoder can improve the generated waveform quality compared with that obtained using the conventional vocoder and gives much higher accuracy than both the conventional vocoder

Table 4.2: *Result of speaker identity preference test in the internal subjective evaluation of this system (NU) and the baseline (sprocket) for same-gender and cross-gender conversions.*

Spk. Identity	Same-gender	Cross-gender
Baseline	43% ± 7%	45% ± 8%
NU	57% ± 7%	55% ± 8%

and the vocoder-free system, i.e., direct waveform modification.

4.4.4 Official subjective evaluation

In VCC 2018, to compare the performance of the submitted systems, an official subjective evaluation was conducted, which consists of a mean opinion score (MOS) test on the speech quality and a speaker similarity test. In the MOS test, each listener was given stimuli of audio samples and asked to evaluate the naturalness of the speech sounds using a five-point scale (1: Completely unnatural; 2: Mostly unnatural; 3: Equally natural and unnatural; 4: Mostly natural; 5: Completely natural). In the speaker similarity test, each listener was given a pair of audio samples as stimuli and asked to judge whether they were produced by the same speaker. Their confidence in the decision was given on a four-point scale (1: Same, absolutely sure; 2: Same, not sure; 3: Different, not sure; 4: Different, absolutely sure). The total number of listeners was 106 (49 female, 57 male).

The results of the official objective evaluation are summarized in Fig. 4.4. The results show the average MOS in terms of speech quality, plotted on the x-axis, for every submitted system, including the baseline (sprocket), as well as the original source

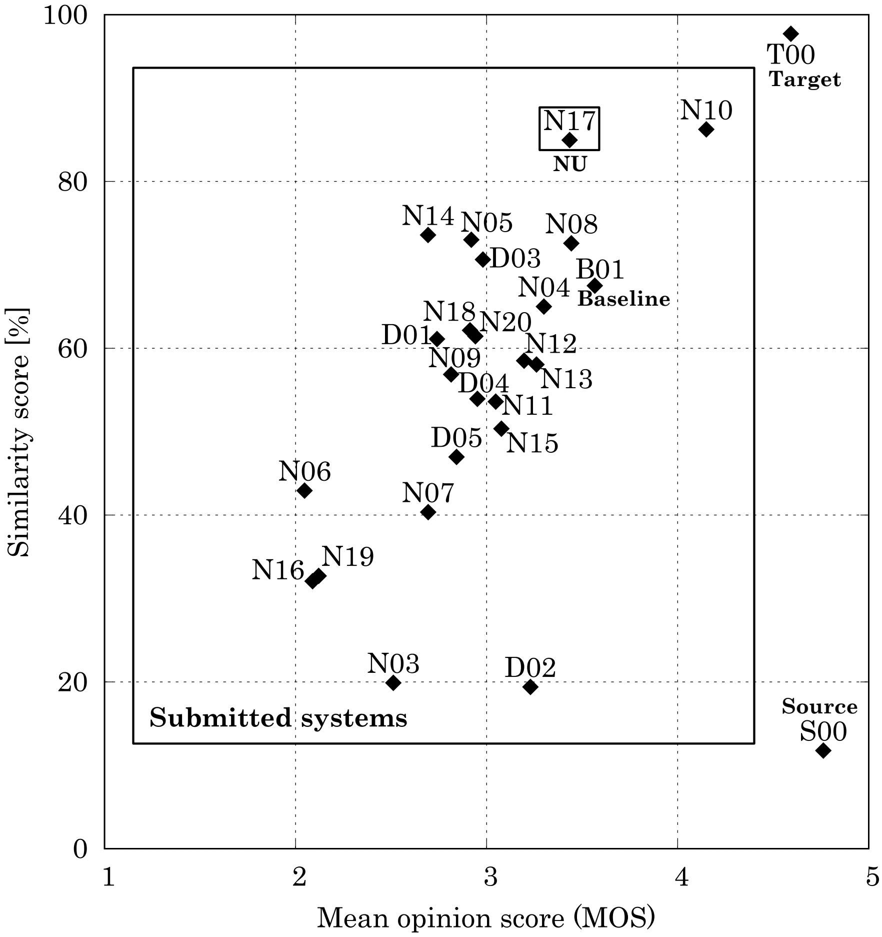


Figure 4.4: *Scatter plot of mean opinion score (MOS) for speech quality and speaker similarity score for the submitted systems along with baseline (sprocket) [B01], source [S00], and target [T00] speech. This VC system (NU) is N17.*

and target speakers. The similarity scores (in %), computed by adding up the two confidence scores in each binary similarity decision, are plotted on the y-axis. This VC system (NU), denoted as N17, achieves an MOS of 3.44 for speech quality, compared with 3.57 for the baseline, 4.15 for the top system, i.e., N10, and also 3.44 for the closest system, i.e., N08. On the other hand, for the speaker similarity test, this

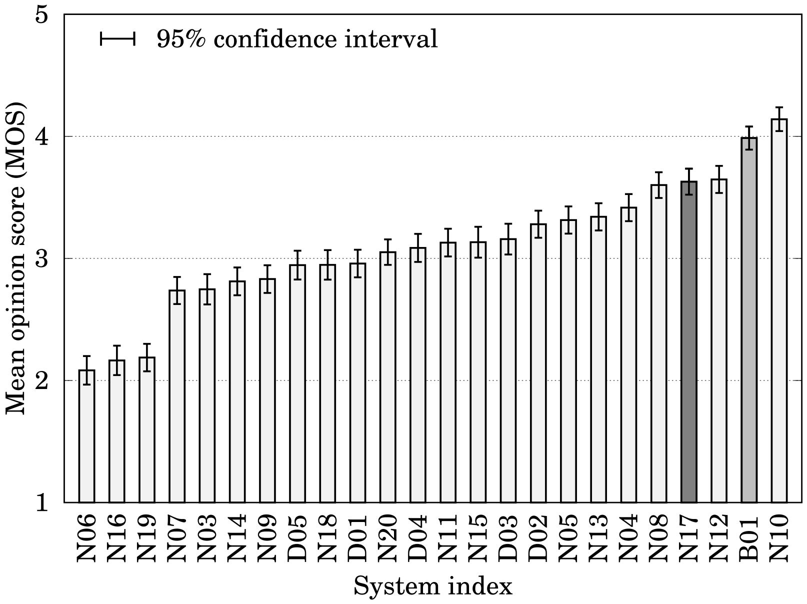


Figure 4.5: *MOS results for speech quality for same-gender conversions, i.e., female-to-female (F-F) and male-to-male (M-M) conversions. This VC system (NU) is N*17.

system achieves a similarity score of 85%, outperforming the baseline (68%) and all other systems except system N10, which has slightly higher accuracy (86%). Overall, this system was placed as the runner-up behind the top system, N10. Details of the official subjective evaluation results are given in the following sections.

4.4.5 Detailed results for speech quality

The detailed official results of the MOS for speech quality for all systems including the baseline, are given in Figs. 4.5 and 4.7. The results for same-gender conversions, which consist of female-to-female (F-F) and male-to-male (M-M) conversions, are shown in the Fig. 4.5, whereas, those for the cross-gender conversions, i.e., female-tomale (F-M) and male-to-female (M-F), are shown in Fig. 4.7.

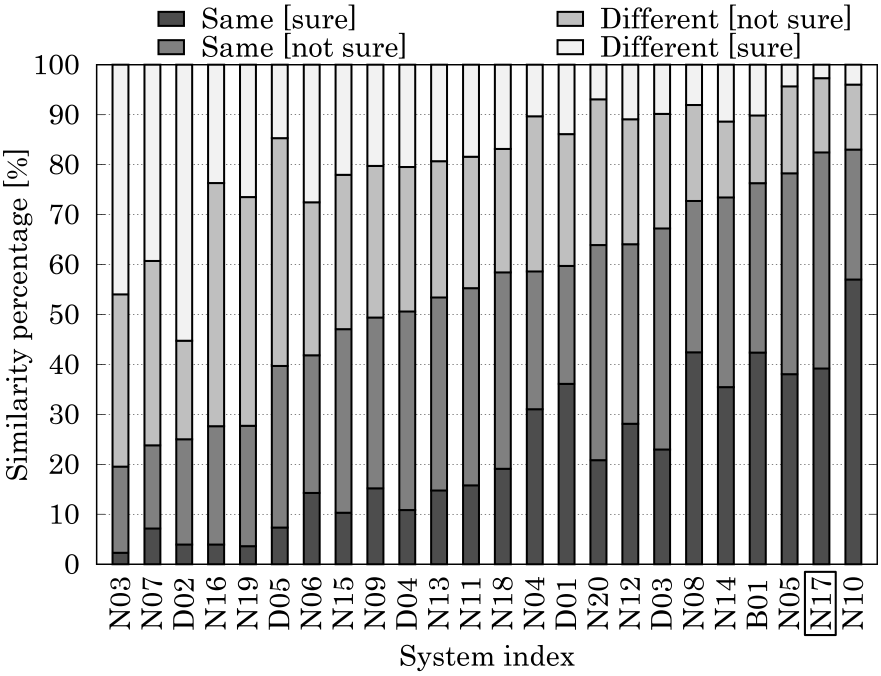


Figure 4.6: *Similarity percentage results for same-gender conversions, i.e., F-F and M-M conversions with two confidence levels in the binary decision. This VC system (NU) is N*17.

This VC system (NU), denoted as N17, achieves an MOS of 3.24 for the cross-gender conversions and 3.63 for the same-gender conversions, which place the system in the fourth and the third places, respectively. The MOSs for each gender conversion are 3.89 for F-F, 3.38 for M-M, and 3.24 for both F-M and M-F. Compared with the baseline, it is expected that this system will perform better in cross-gender conversions because sprocket uses the vocoder-based method in these conversions. The MOSs for the baseline system are 4.10, 3.88, 3.31, and 3.00 for the above conversions, respectively. However, this system is outperformed by the top system, i.e., N10, which achieves an MOS of over 4.10 for every gender-type conversion. Overall, compared with the other submitted systems, this system yields a good performance with an average MOS of

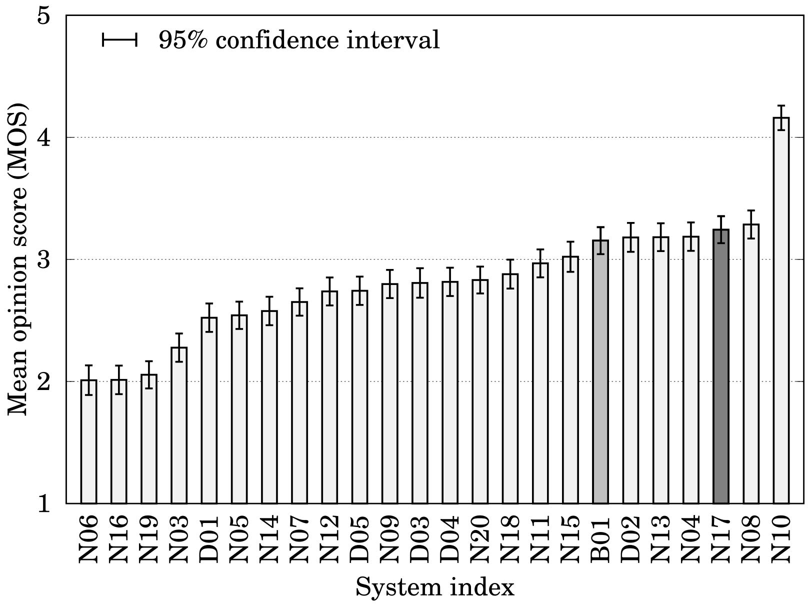


Figure 4.7: *MOS results for speech quality for cross-gender conversions, i.e., female-to-male (F-M) and male-to-female (M-F) conversions. This VC system (NU) is* $N17$.

3.44 over all speaker pairs, the same as the system N08, slightly below those of the baseline (3.57) and far behind system N10 (4.15).

4.4.6 Detailed results for speaker similarity

The official results for the speaker similarity evaluation are shown in Figs. 4.6 and 4.8. The results for same-gender conversions, i.e., F-F and M-M, are shown in Fig. 4.6, whereas the result for cross-gender conversions, i.e., F-M and M-F, are shown in Fig. 4.8. These figures show the percentage of speaker similarity decisions, i.e., “same” or “different”, each with two confidence levels, i.e., “sure” and “not sure”. To measure the final similarity score, the percentage scores of “same” (“sure”) and “same” (“not sure”) are added together.

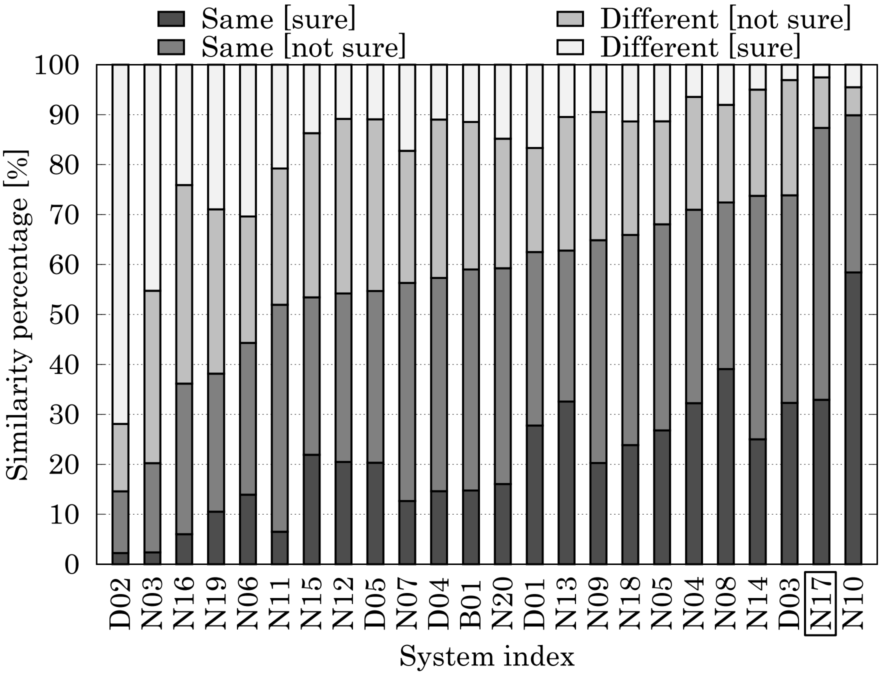


Figure 4.8: *Similarity percentage results for cross-gender conversions, i.e., F-M and M-F conversions with two confidence levels in the binary decision. This VC system (NU) is N*17.

This VC system (N17) has a total similarity score of 82% ("same" decisions), i.e., 18% "different" decisions for the same-gender conversions, and a similarity score of 87% for the cross-gender conversions. The details for each gender conversion are as follows: similarity scores of 82% for F-F conversion, 83% for M-M conversion, 93% for F-M conversion, and 76% for M-F conversion. In this speaker similarity evaluation, this system outperforms the baseline, as shown in both Figs. 4.6 and 4.8, where the baseline has the following similarity scores in the same order: 84%, 53%, 59%, and 60%. Compared with the top system, i.e., N10, this system yields similar results: where our system has a slightly better scores in F-F conversion (81% for N10) and in F-M conversion (91% for N10), a lower score in M-F conversion (85% for N10), and a much lower score in M-M conversion (94% for N10). Overall, this system yields a very good performance with an average similarity score of 85% over all speaker pairs, outperforming the baseline (68%) and all of the systems (the closest are N14 with 74%, and both N05 and N08 with 73%) except for N10, which has a slightly higher score of 86%.

4.5 Conclusion

In this work, a voice conversion (VC) system based on neural network spectral mapping and waveform modeling developed for the HUB task of the Voice Conversion Challenge 2018 has been presented. This VC system adopts a deep learning architecture to develop a spectral parameter conversion model by combining a deep neural network (DNN) and deep mixture density network (DMDN) to form a cascaded DMDN (CascDMDN). In the waveform modeling and generation, this system employs a WaveNet-based vocoder. The auxiliary features fed into the WaveNet system are chosen from several analysis-synthesis flows using a model selection procedure in an utterance-wise manner. The results of the challenge put this VC system in the second place with an average mean opinion score (MOS) of 3.44 for speech quality and a similarity score of 85% for speaker identity.

5 Voice Conversion with Cyclic Recurrent Neural Network and Finely Tuned WaveNet Vocoder

5.1 Introduction

As has been described in previous chapter, voice conversion (VC) [1–3,56] is a framework for transforming the voice characteristics of a source speaker into that of a particular target speaker. In order to develop a VC system, generally, a speech signal of the source speaker is decomposed into several components that can be transformed, in a more convenient way, into that of the target speaker, such as spectral and prosodic features. Then, the converted speech waveform is synthesized from the transformed speech features, such as by using vocoder-based waveform generation [3]. Observing that a lot of speech related works have been employing VC concept [4–8,18,58,59,118], as well as in other closely related tasks, e.g., speaking style conversion and in dialect conversion, it is worthwhile to improve the quality of VC, especially by an examination of the waveform generation.

In a conventional vocoder [13, 22, 119–121], assumptions on the speech production [11, 12] procedure are used for generating the speech signal. Recently, an alternative data-driven approach for speech waveform generation, by using neural network, has been becoming prominent [14, 66, 68]. The latter framework, which is usually caled neural vocoder, has shown significant improvements for generating synthetic speech signal with similar quality as that of the natural speech [14, 122], such as the WaveNet vocoder. This significant progress can be achieved thanks to the capability of neural vocoder in learning from the given speech data, instead of using predefined assumptions on speech production as in conventional vocoder. In this work, the autoregressive (AR) neural vocoder is used, particularly WaveNet vocoder, which is conditioned on spectral and prosodic features [23, 24] to generate speech sample-by-sample. Whereas in the training phase, the ground truth of the previous waveform samples are given along with the conditioning features.

As can be observed, owing to its data-driven capability, the use of neural vocoder has a potential to improve the quality of converted speech in a VC framework. Similarly, as in the conventional vocoder, to generate converted speech, the transformed speech features are fed in to the neural vocoder for generating the speech waveform. Though, thanks to its nature as a statistical model which may compensate the mismatches of speech features, it is possible that the quality and the accuracy of converted speech will be improved than that with conventional vocoder. In fact, indeed, the use of neural vocoder in VC systems has recently been proven to be successful by the top performers [15, 123] in the Voice Conversion Challenge (VCC) 2018 [108]. Note that compared to [123], which uses intermediate phonetic-based features for WaveNet vocoder, in this work, a parallel VC framework without any text/linguistic information is used because not all of the time a robust automatic speech recognition can be obtained.

In [15], as also within the previous Chapter 4, although it can achieve the 2nd rank in the VCC 2018, the mismatches between speech features have not been addressed directly within the WaveNet vocoder. These mismatches occur between the spectral features estimated from the spectral mapping model and the spectral features extracted from the speech signal. Because WaveNet is a data-driven neural vocoder, it is possible to use oversmoothed features (estimated spectral features) for fine-tuning a model that is pretrained with extracted spectral features. Though, in VC, this procedure is not straightforward to be done due to the difference of temporal alignment between the speech signal of the source speaker and that of the target speaker. In other words, further errors would be introduced, due to the difference of temporal structure, if the converted spectral features of the source speaker are used together with the speech waveform of the target speaker for fine-tuning the WaveNet vocoder.

In [35], the problem can be overcome by using a concatenated mapping of target-to-source spectral model and source-to-target spectral model, with recurrent neural network (RNN)-based architecture, which are trained separately (ConcatRNN). This way, an appraisal of oversmoothed target spectral features with the same temporal structure of the target speech signal can be obtained from the concatenated flow. However, due to the gap of connection between the two separately trained networks, the reliability of the oversmoothed target features is not guaranteed which can hamper the performance of the WaveNet fine-tuning.

In this work, a cyclic spectral mapping framework with RNN modules (CycleRNN) is presented, which can optimize both of the conversion flow, i.e., source-to-target, and the cyclic flow, i.e., to generate an appraisal of oversmoothed target spectral. The concept of both the proposed CycleRNN spectral mapping model and the conventional ConcatRNN closely resembles that of the back-translation method for style transfer in natural language processing [124,125]. Specifically, the proposed CycleRNN framework is trained by using two losses, namely the cyclic loss, i.e., for oversmoothed target spectral, and the conversion loss, i.e., for converted source-to-target spectral, which is computed with time-warping alignment procedure, such as dynamic-time-warping

(DTW), due to differences in temporal structure. Due to the difference in performing loss computations, where time-warping function is used for conversion loss, a weighting value for the cyclic loss is needed to balance the contribution of the loss in this multi-task-like learning framework. Note that, different than a cyclic network based on generative adversarial network (CycleGAN) [27], the proposed CycleRNN framework does not use a discriminator, but uses a time-warping function between source and target speaker for direct optimization of the converted spectral features, whereas the cyclic flow can be directly optimized because of the unchanging temporal structure.

After training the CycleRNN model, to perform WaveNet fine-tuning, the over-smoothed target spectral features generated by using cyclic flow are used together with the speech waveform of the target speaker. On the other hand, for the conversion phase, the conversion flow is used to simply generate the converted source-to-target spectral features, which are then fed into the fine-tuned WaveNet to generate the converted speech waveform. The experimental results demonstrate the effectiveness of the proposed framework achieving a mean opinion score of 3.50 for speech quality and a speaker similarity score of 78.33% for conversion accuracy by using 1e-6 as the weight of the cyclic loss.

This chapter is organized as follows. In Section 5.2, the RNN-based spectral model used in this work is described. In Section 5.3, the WaveNet fine-tuning procedure and the proposed CycleRNN spectral model are elaborated. In Section 5.4, the results of experimental evaluations are given, which are followed by a discussion in Section 5.5. Finally, the system is summarized in Section 5.6.

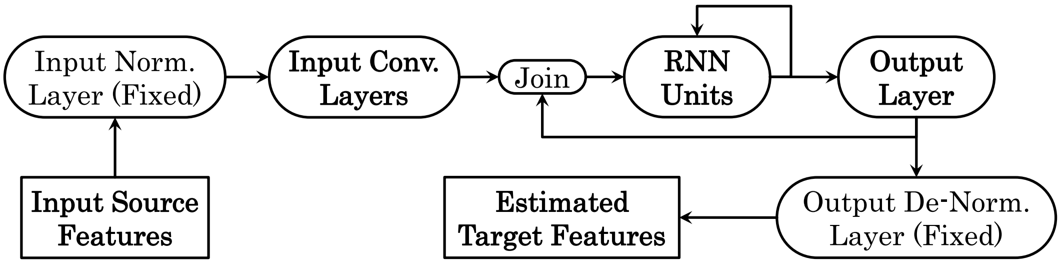


Figure 5.1: *Spectral conversion network with recurrent neural network (RNN) hidden units and autoregressive (AR) output.*

5.2 RNN-based Spectral Mapping Model

Let $\boldsymbol{x}_t = [x_t(1), x_t(2), \ldots, x_t(d), \ldots, x_t(D)]^\top$ and $\boldsymbol{y}_t = [y_t(1), y_t(2), \ldots, y_t(d), \ldots, y_t(D)]^\top$ be the D-dimensional spectral feature vectors of the source speaker and target speaker, respectively, at frame t. Their feature vector sequences are respectively denoted as $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_t^\top, \ldots, \boldsymbol{x}_T^\top]^\top$ and $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \ldots, \boldsymbol{y}_t^\top, \ldots, \boldsymbol{y}_T^\top]^\top$. Note that in this work, for a feedforward RNN function $f_{\boldsymbol{\lambda}}(\cdot)$, the output will either be the spectral feature vector of the target speaker $\boldsymbol{y}_t$ or be that of the source speaker $\boldsymbol{x}_t$. On the other hand, it is possible to augment the input feature vector with additional information, such as F0 and aperiodicity features. However, to simplify the notation, either $\boldsymbol{x}_t$ or $\boldsymbol{y}_t$ will be used to indicate an input feature vector of a source speaker or a target speaker, respectively, regardless of the content.

Given an input feature vector of the source speaker $\boldsymbol{x}_t$ at frame t, the estimated spectral feature vector of the target speaker $f_{\boldsymbol{\lambda}}(\boldsymbol{x}_t) = \hat{\boldsymbol{y}}_t$ is determined through using gated recurrent unit (GRU) [126] blocks with an AR flow as follows:

$$\boldsymbol{i}_t = [\tilde{\boldsymbol{x}}_t^\top, \hat{\boldsymbol{y}}_{t-1}^\top]^\top \tag{5.1}$$

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_{ir}\boldsymbol{i}_t + \boldsymbol{b}_{ir} + \boldsymbol{W}_{hr}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{hr}) \tag{5.2}$$

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_{iz}\boldsymbol{i}_t + \boldsymbol{b}_{iz} + \boldsymbol{W}_{hz}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{hz}) \tag{5.3}$$

$$\boldsymbol{n}_t = \tanh(\boldsymbol{W}_{in}\boldsymbol{i}_t + \boldsymbol{b}_{in} + \boldsymbol{r}_t \odot (\boldsymbol{W}_{hn}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{hn})) \tag{5.4}$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) \odot \boldsymbol{n}_t + \boldsymbol{z}_t \odot \boldsymbol{h}_{t-1} \tag{5.5}$$

$$\hat{\boldsymbol{y}}_t = \boldsymbol{W}_{yh}\boldsymbol{h}_t + \boldsymbol{b}_{yh}, \tag{5.6}$$

where the weights and biases are respectively denoted as $\boldsymbol{W}$ and $\boldsymbol{b}$, $\odot$ denotes the elementwise product, and $\tilde{\boldsymbol{x}}_t$ is the input feature vector processed with a convolutional input layer as illustrated in Fig. 5.1. The reset, the update, and the new gates are respectively denoted as $\boldsymbol{r}_t$, $\boldsymbol{z}_t$, and $\boldsymbol{n}_t$. The hidden state for the current frame t is denoted as $\boldsymbol{h}_t$. For a GRU with more than one layer, the input feature vector of the succeeding layer is the hidden state of the previous layer. At frame $t = 0$, the AR feature vector $\hat{\boldsymbol{y}}_{t-1}$ is initialized with the zero vector $\boldsymbol{0}$. The set of model parameters is denoted as $\boldsymbol{\lambda}$.

In the training phase, the optimized model parameters $\hat{\boldsymbol{\lambda}}$ are estimated as follows:

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} \frac{10\sqrt{2}}{\ln 10} \sum_{d=1}^{D} |\hat{y}_t(d) - y_t(d)|, \tag{5.7}$$

where $f_{\boldsymbol{\lambda}}(\boldsymbol{x}_t) = [\hat{y}_t(1), \hat{y}_t(2), \dots, \hat{y}_t(d), \dots, \hat{y}_t(D)]^\top$ and $|\cdot|$ denotes the absolute function. Note that, as illustrated in Fig. 5.1, owing to the use of normalization and de-normalization layers, the loss function can be defined, as in Eq. (5.7), to be within the spectral domain, i.e., in this case, it is the L1 loss in the mel-cepstrum domain. In the conversion phase, to generate a sequence of converted target spectral feature vectors $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \hat{\boldsymbol{y}}_2^\top, \dots, \hat{\boldsymbol{y}}_t^\top, \dots, \hat{\boldsymbol{y}}_T^\top]^\top$, feed the RNN is simply fed with a sequence of input feature vectors of the source speaker $\boldsymbol{x}$, i.e., $f_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \hat{\boldsymbol{y}}$. Finally, the convolutional input layers illustrated in Fig. 5.1 are designed to dynamically create segmental input feature vectors through the use of convolutional weights, which can capture several preceding and succeeding contextual input frames.

5.3 Fine-Tuning of the WaveNet Vocoder

5.3.1 WaveNet Fine-Tuning with Oversmoothed Features to Overcome Quality Degradation Problem in VC

In a statistical VC framework, as illustrated in the top diagram of Fig. 5.2, there are mismatches between the converted spectral features of the source speaker and the natural spectral features of the target speaker. These mismatches degrade the quality of the converted speech waveform generated using the WaveNet vocoder because it is developed with the natural spectral features. In a TTS system [122], the use of the predicted mel-spectrogram, i.e., oversmoothed spectral features, in the development of the WaveNet model, has increased the quality of statistical TTS to that of natural speech. Though, in VC, aligning the spectral features in the time domain with those of the target waveform would further introduce artifacts and phonetical mismatches, depending on the difference in the voice characteristics/speaking style between the two speakers. In [123], this problem is overcome through the use of phonetic-based features, i.e., a phonetic posteriorgram (PPG), which can be estimated independently for each speaker. In this system, the WaveNet model was developed using the predicted PPG features of each speaker. In this work, however, this problem is tackled without the use of any text/linguistic features because a robust automatic speaker recognition (ASR) system cannot always be built as in [123]. However, as has been stated, in VC, such a

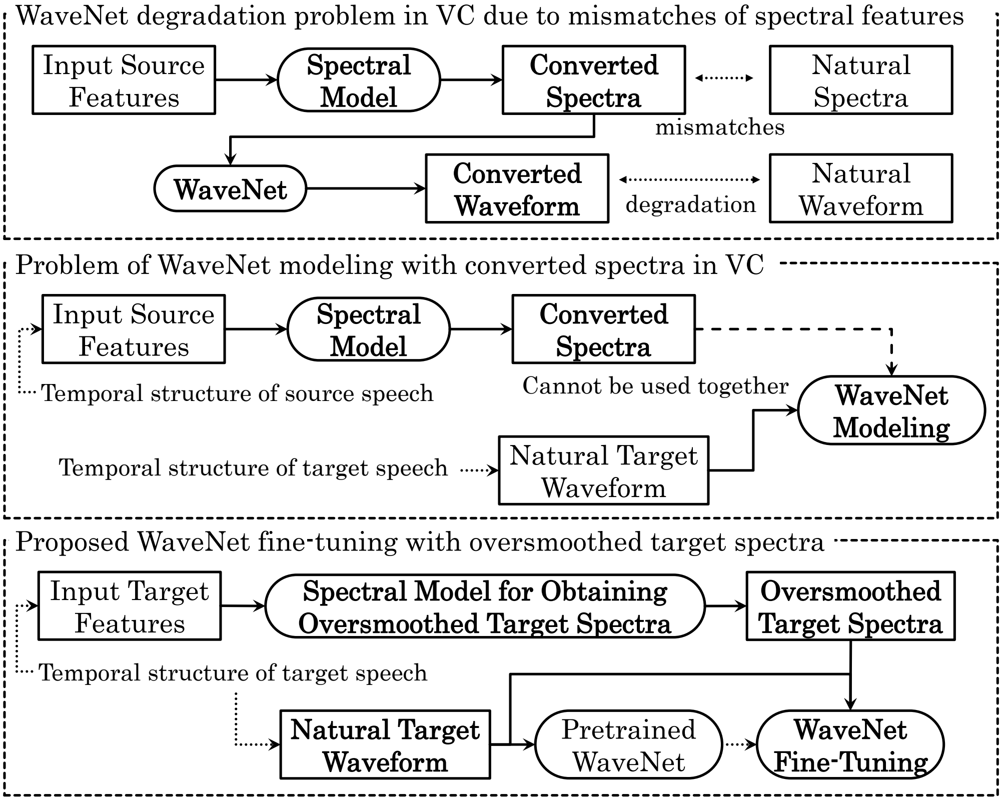

WaveNet degradation problem in VC due to mismatches of spectral features
Input Source Features
Spectral Model
Converted Spectra
Natural Spectra
mismatches
WaveNet
Converted Waveform
degradation
Natural Waveform
Problem of WaveNet modeling with converted spectra in VC
Input Source Features
Spectral Model
Converted Spectra
Temporal structure of source speech
Cannot be used together
WaveNet Modeling
Temporal structure of target speech
Natural Target Waveform
Proposed WaveNet fine-tuning with oversmoothed target spectra
Input Target Features
Spectral Model for Obtaining Oversmoothed Target Spectra
Oversmoothed Target Spectra
Temporal structure of target speech
Natural Target Waveform
Pretrained WaveNet
WaveNet Fine-Tuning

Figure 5.2: *Degradation problem of WaveNet vocoder in VC due to mismatches of speech features (top). Difficulties in using oversmoothed features to develop WaveNet in a VC framework (middle). Proposed WaveNet fine-tuning with oversmoothed target features to alleviate the degradation problem (bottom).*

procedure is not straightforward to implement owing to the difference in the temporal sequence alignment between a source speaker and a target speaker as illustrated in the middle diagram of Fig. 5.2.

Before the conventional spectral mapping method used to enable WaveNet fine-tuning in VC without any linguistic features is described, a simple modification of the likelihood function in the WaveNet fine-tuning procedure will be explained. Following the WaveNet likelihood function in Eq. (2.31), let us redefine the notation of the auxiliary parameters in frame t as $\boldsymbol{h}_t = [\boldsymbol{g}_t^\top, \boldsymbol{y}_t^\top]^\top$, where $\boldsymbol{g}_t$ contains both excitation features, such as F0 values and voice/unvoiced decisions, and aperiodicity features. First, before performing fine-tuning, a pretrained WaveNet model is developed by using the natural spectral feature vectors of all available speakers. In the conventional VC framework, this multispeaker WaveNet model is fine-tuned with the natural spectral features of the target speaker $\boldsymbol{y}_t$. In the proposed method, to improve the WaveNet model in a VC framework, instead of using the natural spectral features, the WaveNet is finely tuned by using the predicted target spectral feature vectors $\hat{\boldsymbol{y}}_t^{(pred)}$, which should resemble the spectral features of the target speaker with oversmoothed characteristics. Therefore, in the WaveNet fine-tuning, the optimized parameter set of the WaveNet model $\hat{\boldsymbol{\theta}}$ is estimated as follows:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{t=1}^{T} P(s_t | \boldsymbol{s}_{t-p}, \boldsymbol{g}_t, \hat{\boldsymbol{y}}_t^{(pred)}, \boldsymbol{\theta}). \tag{5.8}$$

The proposed WaveNet fine-tuning procedure with oversmoothed features is illustrated in the bottom diagram of Fig. 5.2. The conventional procedure to obtain the predicted target spectral features is described in the next subsection.

5.3.2 Concatenated Mapping with RNN-based Spectral Models to Obtain Oversmoothed Features

To obtain oversmoothed target spectral features with the same temporal structure as the target speech waveform, an RNN-based concatenated mappings between the target and source speakers [35] can be employed. Let the feedforward RNN function of the source-to-target mapping (STmap) be $f_{\boldsymbol{\lambda}_{\mathrm{ST}}}(\cdot)$ and that of the target-to-source mapping (TSmap) be $f_{\boldsymbol{\lambda}_{\mathrm{TS}}}(\cdot)$. Given a sequence of input spectral feature vectors of the source speaker $\boldsymbol{x}$, the sequence of converted spectral feature vectors corresponding to

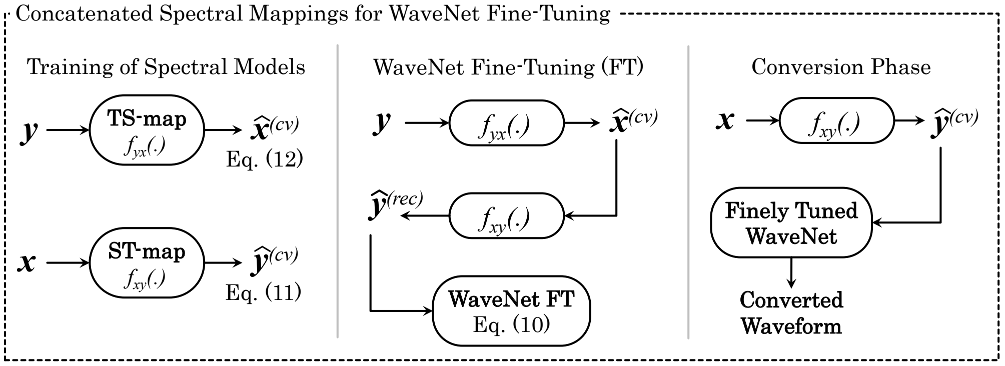


Figure 5.3: *Flow of the concatenated spectral mappings to enable the WaveNet fine-tuning (FT) procedure in a VC framework.*

the target speaker is given by $f_{\boldsymbol{\lambda}_{\mathrm{ST}}}(\boldsymbol{x}) = \hat{\boldsymbol{y}}^{(conv)}$, for which the optimized set of STmap parameters $\hat{\boldsymbol{\lambda}}_{\mathrm{ST}}$ is estimated as

$$\hat{\boldsymbol{\lambda}}_{\mathrm{ST}} = \operatorname*{argmin}_{\boldsymbol{\lambda}_{\mathrm{ST}}} |\overline{\hat{\boldsymbol{y}}}^{(conv)} - \boldsymbol{y}|, \tag{5.9}$$

where the above distance function follows the L1 loss in the mel-cepstrum domain, as in Eq. (5.7), and $\overline{\hat{\boldsymbol{y}}}^{(conv)}$ denotes the time-aligned sequence of the converted spectral feature vectors. Conversely, given a sequence of input spectral feature vectors of the target speaker $\boldsymbol{y}$, the sequence of converted spectral feature vectors corresponding to the source speaker is given by $f_{\boldsymbol{\lambda}_{\mathrm{TS}}}(\boldsymbol{y}) = \hat{\boldsymbol{x}}^{(conv)}$, for which the optimized set of TSmap parameters $\hat{\boldsymbol{\lambda}}_{\mathrm{TS}}$ is estimated as

$$\hat{\boldsymbol{\lambda}}_{\mathrm{TS}} = \operatorname*{argmin}_{\boldsymbol{\lambda}_{\mathrm{TS}}} |\overline{\hat{\boldsymbol{x}}}^{(conv)} - \boldsymbol{x}|, \tag{5.10}$$

where the above distance function similarly follows the L1 loss in the mel-cepstrum domain given in Eq. (5.7) and $\overline{\hat{\boldsymbol{x}}}^{(conv)}$ denotes the time-aligned sequence of the converted spectral feature vectors.

To enable the WaveNet fine-tuning procedure, as illustrated in Fig. 5.3, the above two mapping functions are simply concatenated so that $f_{\boldsymbol{\lambda}_{\mathrm{ST}}}(f_{\boldsymbol{\lambda}_{\mathrm{TS}}}(\boldsymbol{y})) = \hat{\boldsymbol{y}}^{(pred)}$. Consequently, $\hat{\boldsymbol{y}}^{(pred)}$ inherits the oversmoothed characteristics of the estimated spectral trajectory with the same temporal structure as the input natural target spectral $\boldsymbol{y}$, and accordingly as the waveform of the target speech. However, it can be observed that the sequence of self-predicted target spectra $\hat{\boldsymbol{y}}^{(pred)}$, which is used in Eq. (5.8), is not optimized in Eq. (5.9) or in Eq. (5.10). In other words, there is still inconsistency between the oversmoothed features used in the WaveNet fine-tuning and the concatenated RNN-based spectral mappings. This inconsistency is addressed through the proposed CycleRNN-based spectral modeling described in the next subsection.

5.3.3 Proposed CycleRNN Spectral Mapping Model to Improve WaveNet Fine-Tuning

In this work, to improve the WaveNet fine-tuning procedure, a CycleRNN spectral mapping model is proposed, which is optimized using both the cyclic loss, i.e., of the self-predicted target spectra, and the conversion loss, i.e., of the source-to-target conversion. Specifically, the CycleRNN model consists of two mapping modules, namely the feedforward RNNs $f_{\boldsymbol{\lambda}_{\mathrm{c}_1}}(\cdot)$ and $f_{\boldsymbol{\lambda}_{\mathrm{c}_2}}(\cdot)$. In the conversion flow, i.e., to perform source-to-target conversion, given a sequence of input spectral feature vectors of the source speaker $\boldsymbol{x}$, the converted sequence is given by $f_{\boldsymbol{\lambda}_{\mathrm{c}_2}}(\boldsymbol{x}) = \hat{\boldsymbol{y}}^{(conv)}$. On the other hand, in the cyclic flow, given a sequence of input target spectral feature vectors $\boldsymbol{y}$, its self-predicted sequence is given by $f_{\boldsymbol{\lambda}_{\mathrm{c}_2}}(f_{\boldsymbol{\lambda}_{\mathrm{c}_1}}(\boldsymbol{y})) = \hat{\boldsymbol{y}}^{(pred)}$. Hence, the set of optimized parameters of the CycleRNN model $\hat{\boldsymbol{\lambda}}_{\mathrm{c}} = \{\hat{\boldsymbol{\lambda}}_{\mathrm{c}_1}, \hat{\boldsymbol{\lambda}}_{\mathrm{c}_2}\}$ is estimated as follows:

$$\hat{\boldsymbol{\lambda}}_{\mathrm{c}} = \underset{\hat{\boldsymbol{\lambda}}_{\mathrm{c}}}{\operatorname{argmin}}\, |\overline{\hat{\boldsymbol{y}}}^{(conv)} - \boldsymbol{y}| + \alpha|\hat{\boldsymbol{y}}^{(pred)} - \boldsymbol{y}|, \tag{5.11}$$

where the above distance functions also follow the L1 loss function in the mel-cepstral domain as in Eq. (5.7), $\overline{\hat{\boldsymbol{y}}}^{(conv)}$ denotes the time-aligned sequence of the converted

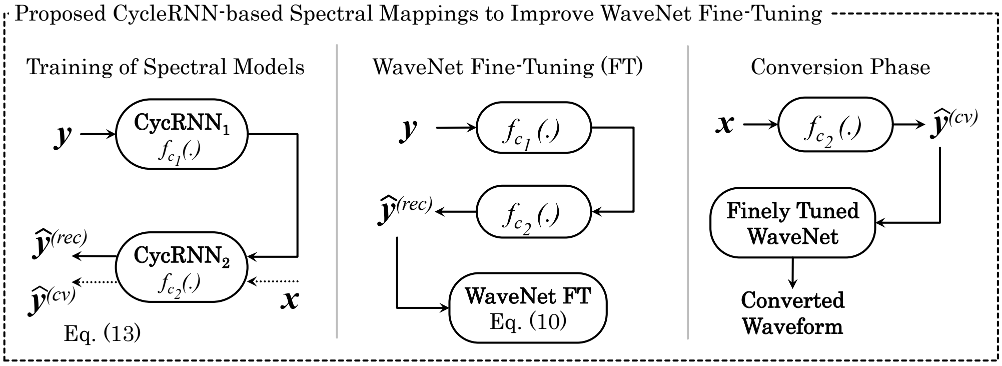


Figure 5.4: *Proposed CycleRNN-based spectral mapping model, to improve the WaveNet fine-tuning (FT) procedure, by incorporating not only the conversion, but also the cyclic losses in the spectral model development.*

spectral feature vectors, and α denotes the weight of the cyclic loss.

The training procedure of the CycleRNN spectral mapping model is illustrated in Fig. 5.4. It can be observed, as given in Eq. (5.11), that the parameters of the CycleRNN model are optimized according to both the estimation of the converted spectral feature vectors $\hat{\boldsymbol{y}}^{(conv)}$ and that of the self-predicted target spectral feature vectors $\hat{\boldsymbol{y}}^{(pred)}$. Compared with the simple concatenated mappings in Section 5.3.2, the proposed CycleRNN model should be more beneficial to the WaveNet fine-tuning procedure described in Section 5.3.1, which employs the self-predicted target spectral features in the parameter optimization as in Eq. (5.8). Further, it is also very important to consider the weight of the cyclic loss α because the conversion loss and cyclic loss are in different domains. The conversion loss utilizes a time-warping function, for example, obtained through DTW, to align the time sequence, whereas the cyclic loss is a direct framewise comparison that makes it easier to be more accurate, resulting in the need for a weighting value. This feature is thoroughly investigated in the following experimental evaluation.

5.4 Experimental Evaluation

5.4.1 Experimental Conditions

WORLD [22, 114] was used to parameterize the speech waveform signal. Framewise fundamental frequency (F0) values were extracted using Harvest [127] in WORLD. The spectral envelope of the speech spectrum was computed frame by frame using CheapTrick [128,129] in WORLD then parameterized into the zeroth through 34th mel-cepstrum coefficients. As the aperiodicity features, two-dimensional coding parameters were extracted from the computed aperiodicity values of D4C [130] in WORLD. 1024 points were used for fast Fourier transform analysis. The frame shift was set to 5 ms.

In the WaveNet modeling, a multispeaker WaveNet vocoder [24, 131] was trained using the speech data of 12 speakers in the VCC 2018 [108] dataset and two speakers in the CMU Arctic dataset ("bdl" and "slt"). The total number of utterances per speaker in the VCC 2018 dataset was 81, whereas in the CMU Arctic dataset, it was 1132. As the training set, the final 71 utterances in VCC 2018 and the first 992 utterances in CMU Arctic were used, giving a total of $2,834$ short audio files (with an average duration of 3.5 s). On the other hand, as the validation set, the first 10 utterances in VCC 2018 and the final 140 utterances in CMU Arctic were used, giving a total of 420 short audio files. The sampling rate of speech signal from VCC 2018 and CMU Arctic datasets was 22050 Hz.

The hyperparameters of the WaveNet model are as follows. The length of one dilation sequence, i.e., a sequence of residual blocks with causal dilated convolutions, was 11 $(1, 2, 4, \ldots, 1024)$. The number of repeats of the dilation sequence was four, giving a total of 8190 samples in the receptive field. The numbers of channels for the residual blocks and skip connections were 128 and 256, respectively. Two convolution layers with a kernel size of 3 and dilation sizes of 1 and 3 were used to capture the context of ±4 frames of auxiliary speech parameters. A trainable upsampling layer was used after the input convolutions to match the time resolution of the auxiliary parameters with that of the waveform samples. Dropout [132] layers with 0.5 probability were used after the upsampling layer and after each repeat of the dilation sequence for the residual connections. The speech auxiliary features consisted of voiced/unvoiced (UV) binary decisions, continuous F0 values, 35-dimensional mel-cepstrum parameters, and two-dimensional coded aperiodicity parameters. A batch sequence length of 8800 waveform samples was used. The model parameters were initialized with the Glorot [115] method. The Adam algorithm [116] was used to optimize the parameters with a learning rate of 1e-4. To reduce the errors in the higher-frequency region, a noise shaping [133] method was used.

To train the RNN-based spectral mapping model, the speech data of only four speakers in the VCC 2018 dataset was used, i.e., “SF1”, “SM1”, “TF1”, and “TM1”, where “S”, “T”, “F”, and “M” denote source, target, female, and male, respectively. Similarly to in the development of the WaveNet model, the final 71 utterances were used in the training set and the first 10 utterances were used for the validation set. In the subjective evaluation (listening test), another 35 utterances provided in the VCC 2018 dataset were used for the evaluation set. For the proposed cyclic RNN architecture (CycleRNN) described in Section 5.3.3, the total number of trained models was four, i.e., the total number of combinations of target source–target speaker pairs. On the other hand, for the concatenated RNN mappings (ConcatRNN) described in Section 5.3.2, the total number of trained models was eight because the source-to-target and target-to-source mapping models were trained separately for each speaker pair.

The hyperparameters of the RNN-based spectral models were as follows. The numbers of hidden layers and hidden units for the GRU [126] were 1 and 1024, respectively. A similar structure of the convolution layers to that in the WaveNet model was used to capture the context of ±4 frames of input features. Dropout layers with 0.5 probability were used after the input convolution layers and for the output of the GRU, i.e., before the output projection layer. The input features consist of not only the 35-dimensional mel-cepstrum parameters but also the V/UV binary decisions, the log of continuous F0 values, and two-dimensional aperiodicity coding parameters. The output features consist of only the 35-dimensional mel-cepstrum parameters, i.e., the network estimates only spectral features. A batch sequence length of 80 speech frames was used for the CycleRNN models, whereas the utterance batch size was used for the ConcatRNNs. The network parameters were initialized with the Glorot method and optimized using the Adam algorithm with a learning rate of 0.0001. To compute the conversion loss, i.e., of the source-to-target or target-to-source conversion, for the spectral modeling, time-warping functions were used, which were computed using the dynamic-time-warping algorithm [134] corresponding to only the speech frames, i.e., non-silent frames, of the speech sequences.

To use the multispeaker WaveNet model to generate the converted speech waveform, two types of fine-tuning procedure were performed. The first one was by using the natural (extracted) speech parameters of the corresponding target speaker, which is basically the conventional fine-tuning [24]. The second one was by using the target speech parameters that consisted of oversmoothed mel-cepstrum features obtained through using either the ConcatRNN [35] or the CycleRNN [36]. Hence, in the proposed fine-tuning approach, there was a total of eight fine-tuned WaveNet models, i.e., for all four source-target speaker pairs and for both the ConcatRNN and CycleRNN. Both objective and subjective evaluations were conducted to assess the performance of

Table 5.1: *Mel-cepstral distortion (MCD) [dB] of reconstructed target spectral features in training set (Trc) and converted source-to-target spectral features in validation set (Vcv). Reconstructed target spectral features were used for the WaveNet fine-tuning, which were estimated either using concatenated target-to-source and source-to-target mappings (CatRNN) or using the proposed CycleRNN models with three different weights for the cyclic loss, i.e., 1 (CycRNN0), 0.001 (CycRNN3), and 0.000001 (CycRNN6).*

MCD [dB]	**CatRNN**		**CycRNN0**		**CycRNN3**		**CycRNN6**	
	Trc	**Vcv**	**Trc**	**Vcv**	**Trc**	**Vcv**	**Trc**	**Vcv**
SF1-TF1	5.25	6.41	2.67	6.43	3.90	6.40	4.04	**6.39**
SM1-TM1	4.63	5.42	2.23	5.41	3.37	5.41	3.50	**5.40**
SF1-TM1	4.93	**5.93**	2.43	5.98	3.54	**5.93**	3.67	**5.93**
SM1-TF1	5.31	**6.40**	2.53	6.43	3.92	6.42	4.07	6.41

the spectral mapping models and the WaveNet fine-tuning approach with the converted speech waveform samples.

5.4.2 Objective Evaluation

In the objective evaluation, the performance of the spectral mapping models was assessed by computing the mel-cepstral distortion (MCD) [3] and log-GV distance (GV). To compute the MCD, the following formula was used:

$$\mathrm{MCD[dB]} = \frac{1}{T}\sum_{t=1}^{T}\frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{34}(\hat{y}_t(d) - y_t(d))^2}, \tag{5.12}$$

where $\hat{y}_t(d)$ and $y_t(d)$ respectively denote the dth dimension of the converted mel-cepstrum and that of the natural mel-cepstrum at frame t. To compute the LGD, the

Table 5.2: *Log-GV distortion (LGD) of reconstructed target spectral features in training set (Trc) and converted source-to-target spectral features in validation set (Vcv). Reconstructed target spectral features were used for the WaveNet fine-tuning, which were estimated either using concatenated target-to-source and source-to-target mappings (CatRNN) or using the proposed CycleRNN models with three different weights for the cyclic loss, i.e., 1 (CycRNN0), 0.001 (CycRNN3), and 0.000001 (CycRNN6).*

LGD	**CatRNN**		**CycRNN0**		**CycRNN3**		**CycRNN6**	
	Trc	**Vcv**	**Trc**	**Vcv**	**Trc**	**Vcv**	**Trc**	**Vcv**
SF1-TF1	2.05	1.76	0.70	**1.55**	1.40	1.68	1.45	1.69
SM1-TM1	1.71	1.37	0.51	**1.09**	1.06	1.34	1.10	1.35
SF1-TM1	1.82	1.46	0.56	**1.16**	1.20	1.43	1.24	1.45
SM1-TF1	2.04	1.69	0.62	**1.38**	1.41	1.73	1.45	1.74

following formula was used:

$$\mathrm{LGD} = \frac{1}{D}\sum_{d=1}^{35}\sqrt{(\log\mathrm{GV}(\hat{y}(d)) - \log\mathrm{GV}(y(d)))^2}, \tag{5.13}$$

where $\mathrm{GV}(\hat{y}(d))$ and $\mathrm{GV}(y(d))$ respectively denote the global variance of the dth dimension of the converted mel-cepstrum and that of the natural mel-cepstrum. The GV [3] was computed as follows:

$$\mathrm{GV}(y(d)) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\sum_{t=1}^{T_n}(y_t(d) - \overline{y}_n(d))^2, \tag{5.14}$$

where $\overline{y}_n(d)$ denotes the mean value of the dth mel-cepstrum parameter of the nth utterance and the total number of utterances is denoted as N.

The MCD and LGD of both the reconstructed target spectral features in the training set (Trc) and the converted source-to-target spectral features in the validation set (Vcv) were computed. The reasons for computing these values were twofold: to measure

Table 5.3: *Absolute measurement difference between mel-cepstral distortion (MCD) [dB] (|MCD(Vcv) − MCD(Trc)|) of converted source-to-target spectral features in validation data (Vcv) and reconstructed target spectral features in training data (Trc). The corresponding error values are given in Table 5.1.*

|MCD(Vcv)− MCD(Trc)|	**CatRNN**	**CycRNN0**	**CycRNN3**	**CycRNN6**
SF1-TF1	**1.16**	3.76	2.50	2.35
SM1-TM1	**0.79**	3.17	2.04	1.90
SF1-TM1	**1.00**	3.55	2.39	2.26
SM1-TF1	**1.10**	3.90	2.50	2.35

the model accuracy in converting source spectral features into target spectral features (Vcv) and to monitor the difference between the converted spectral features and the reconstructed target spectral features (Trc), where the latter has the same temporal alignment as the target speech because it was used for WaveNet fine-tuning. In other words, it has to be ensured that the WaveNet model is not fed with overly accurate or inaccurate features in the fine-tuning phase while still maintaining the conversion accuracy, which is important in the conversion phase. The models that were compared were the conventional RNN-based spectral mapping with a concatenated flow to obtain oversmoothed target spectral features (CatRNN) and the proposed cyclic RNN-based spectral model with three different cyclic weights: 1 (CycRNN1), 1e-3 (CycRNN3), and 1e-6 (CycRNN6).

The obtained values of MCD and LGD are respectively given in Tables 5.1 and 5.2. On the one hand, these results show that there are no large differences in the converted validation set (Vcv) among the four spectral models, although, CycRNN6 gives the lowest MCD, whereas CycRNN0 gives the lowest LGD. On the other hand, there are

Table 5.4: *Absolute measurement difference between log-GV distortion (LGD) ($|LGD(Vcv) - LGD(Trc)|$) of converted source-to-target spectral features in validation data (Vcv) and reconstructed target spectral features in training data (Trc). The corresponding error values are given in Table 5.2.*

|LGD(Vcv)− LGD(Trc)|	**CatRNN**	**CycRNN0**	**CycRNN3**	**CycRNN6**
SF1-TF1	0.29	0.85	0.27	**0.24**
SM1-TM1	0.34	0.58	0.28	**0.25**
SF1-TM1	0.36	0.60	0.23	**0.20**
SM1-TF1	0.35	0.77	0.32	**0.29**

noticeable differences in both the MCD and LGD values for the reconstructed target in the training set (Trc) among the four spectral models. CycRNN0 gives the highest accuracy (lowest MCD) and the highest trajectory variance (least oversmoothed/lowest LGD) amoung the four models. As the cyclic weight is decreased to 1e-3 (CycRNN3) and 1e-6 (CycRNN6), lower accuracy and increased oversmoothing can be seen for the Trc features. The lowest accuracy and the most oversmoothed Trc were obtained with CatRNN, which is reasonable because Trc is not directly optimized in the CatRNN spectral model. Although it seems that CycRNN0, i.e., with a cyclic weighting value of 1, gives the best values, what is actually required is a balanced performance between Vcv and Trc. In other words, reasonably small differences are needed between the MCD and LGD values for these two features.

Therefore, to emphasize the reason for carrying out the objective evaluation, the absolute differences between the MCD values of Vcv and Trc and between the LGD values of Vcv and Trc were computed. Although these values are already implied within Tables 5.1 and 5.2, to make them directly visible, the absolute differences are

Table 5.5: *Mean Opinion Score (MOS) for naturalness (*1 *to* 5 *scale) using conventional DiffGV waveform generation, conventional WaveNet (WN), and proposed WaveNet fine-tuning (WNFT). WN-NoGV denotes the use of plain converted spectral features. WN-DiffGV denotes the use of postprocessed spectral features with the DiffGV-based method. WNFT models were fine-tuned with either concatenated RNN (Cat) or the proposed CycleRNN (Cyc). The weight of the cyclic loss in CycleRNN was* 1 *(0),* 0.001 *(3), or* 0.000001 *(6). NoGV spectral features were used for WNFT models. S-Gender and X-Gender denote same-gender and cross-gender conversions, respectively.* ± *denotes the* 95% *confidence interval. [·] denotes a system with a statistically significantly lower score than the highest score in each conversion category.*

MOS of Naturalness	**DiffGV**	**WN-NoGV**	**WN-DiffGV**	**WNFT-Cat**	**WNFT-Cyc0**	**WNFT-Cyc3**	**WNFT-Cyc6**
All Pairs	[2.53±0.20]	[2.34±0.13]	[2.78±0.15]	[3.23±0.17]	[2.96±0.17]	[3.29±0.15]	**3.50±0.14**
S-Gender	3.39±0.23	[2.56±0.19]	[2.96±0.24]	3.51±0.24	[3.33±0.23]	3.56±0.20	**3.69±0.19**
SF1-TF1	3.40±0.32	[2.20±0.22]	3.23±0.37	3.25±0.35	[2.75±0.27]	3.43±0.34	**3.58±0.28**
SM1-TM1	[3.38±0.33]	[2.93±0.28]	[2.70±0.29]	3.78±0.32	**3.90±0.27**	3.70±0.22	3.80±0.27
X-Gender	[1.68±0.21]	[2.13±0.18]	[2.60±0.19]	[2.94±0.24]	[2.60±0.24]	[3.01±0.21]	**3.31±0.21**
SF1-TM1	[1.35±0.21]	[2.43±0.25]	[2.43±0.29]	3.13±0.36	2.93±0.37	2.95±0.31	**3.30±0.29**
SM1-TF1	[2.00±0.34]	[1.83±0.22]	[2.78±0.26]	[2.75±0.31]	[2.28±0.28]	3.08±0.29	**3.33±0.31**

respectively given in Tables 5.3 and 5.4. The results show that CatRNN gives the lowest absolute difference between the MCD of Vcv and Trc. This means that considering the conversion performance (Vcv accuracy) is not particularly different among the four spectral models, CatRNN might provide the oversmoothed target (Trc) that is closest to the converted features for the WaveNet fine-tuning. However, this condition may have the disadvantage of overly poorly reconstructed features, resulting in an overestimation in the WaveNet fine-tuning procedure. If this assumption holds, it means that the absolute MCD difference between Vcv and Trc should not be too small if Vcv is not

Table 5.6: *Mean Opinion Score (MOS) of naturalness (*1 *to* 5 *scale) for natural waveform samples of source and target speakers. In the evaluation, the original waveforms were mixed with the converted waveforms, where the MOS values for the converted waveforms are given in Table 5.5.*

MOS of Naturalness	**SF1**	**SM1**	**TF1**	**TM1**
Original	4.83±0.14	4.93±0.09	4.90±0.10	4.83±0.12

particularly accurate (which is true in this case). In other words, CycRNN3 and CycRNN6 might give a better reconstructed target (Trc) owing to their intermediate absolute difference values. On the other hand, the absolute differences between the LGD values of Vcv and Trc show that CycRNN6 gives the smallest difference between the converted and reconstructed spectra in the oversmoothness measurement. The latter result might allow CycRNN6 to provide a balanced performance in spectral modeling that considers both the WaveNet fine-tuning phase and the conversion phase. Meanwhile, it can be clearly seen that CycRNN0 gives the largest absolute difference for both MCD and LGD among the four models. This analysis of the results of the objective evaluation supports the results of the perceptual evaluation given in the next subsection.

5.4.3 Subjective Evaluation

In the subjective evaluation, a mean opinion score (MOS) test was conducted to evaluate the naturalness of the converted speech waveforms and a speaker similarity test to evaluate the accuracy of the converted speech with respect to the natural tar-

Table 5.7: *Results of speaker similarity score (%) aggregated from “same_sure” and “same_not sure” decisions using conventional DiffGV waveform generation, conventional WaveNet (WN), and proposed WaveNet fine-tuning (WNFT). WN-NoGV denotes the use of plain converted spectral features. WN-DiffGV denotes the use of post-processed spectral features with the DiffGV-based method. WNFT models were fine-tuned with either concatenated RNN (Cat) or the proposed CycleRNN (Cyc). The weight of the cyclic loss in CycleRNN was* 1 *(0),* 0.001 *(3), or* 0.000001 *(6). NoGV spectral features were used for WNFT models. S-Gender and X-Gender denote same-gender and cross-gender conversions, respectively. [·] denotes a system with a statistically significantly lower score than the highest score in each conversion category.*

Speaker Similarity Scores (%)	**DiffGV**	**WN-NoGV**	**WN-DiffGV**	**WNFT-Cat**	**WNFT-Cyc0**	**WNFT-Cyc3**	**WNFT-Cyc6**
All Pairs	[47.50]	[62.50]	[58.33]	66.67	[62.50]	[68.33]	**78.33**
S-Gender	[58.33]	[70.00]	[65.00]	73.33	[66.67]	[70.00]	**81.67**
SF1-TF1	60.00	[80.00]	80.00	70.00	[73.33]	66.67	**86.67**
SM1-TM1	[56.67]	[60.00]	[50.00]	**76.67**	[60.00]	73.33	**76.67**
X-Gender	[36.67]	[55.00]	[51.67]	60.00	[58.33]	66.67	**75.00**
SF1-TM1	[23.33]	[36.67]	[43.33]	60.00	56.67	60.00	**73.33**
SM1-TF1	[50.00]	73.33	60.00	60.00	60.00	73.33	**76.67**

get speech. Seven different systems were employed to generate the converted speech waveforms: direct waveform modification using the spectrum differential [4] and GV [3] postfilter (DiffGV), which was similar to the baseline of VCC 2018 [117]; WaveNet-based generation with plain converted spectra (WN-NoGV); WaveNet-based generation with postprocessed converted spectral features using the DiffGV-based method (WN-DiffGV) [15], which was the VC system described in the previous Chapter 4 for VCC 2018; WaveNet fine-tuned (WNFT) with oversmoothed target spectral features estimated using concatenated RNN mappings (WNFT-Cat) [15]; and the WNFT-based

Table 5.8: *Speaker similarity scores (%) for original waveform samples with respect to each of the target speakers TF1 and TM1. In the evaluation, these pairs were mixed with the pairs using converted audios, where the similarity scores for the converted audios are given in Table 5.7.*

Speaker Similarity Scores (%)	**TF1**	**TM1**
SF1	3.33	0.00
SM1	0.00	0.00
TF1	90.00	-
TM1	-	86.67

model with the CycleRNN framework [36] using weights of cyclic loss of 1e0 (WNFT-Cyc0), 1e-3 (WNFT-Cyc3), and 1e-6 (WNFT-Cyc6). The converted speech waveforms were generated from the conversion of four speaker pairs: source female to target female (SF1-TF1), source male to target male (SM1-TM1), source female to target male (SF1-TM1), and source male to target female (SM1-TF1). This resulted in a total of 28 different combinations of systems and speaker pairs. Note that in the DiffGV-based waveform generation, for cross-gender conversions, i.e., SF1-TM1 and SM1-TF1, the vocoder-based excitation was used, whereas waveform generation for same-gender conversions was vocoder-free. Also, for WN-NoGV and WN-DiffGV, the multispeaker WaveNet model was fine-tuned with the natural spectral parameters of the corresponding target speaker.

In the MOS test, each listener was given one audio stimuli at a time then was asked to judge its naturalness based on a Likert five-scale score, i.e., 1: completely unnatural, 2: mostly unnatural, 3: equally natural and unnatural, 4: mostly natural, and 5: completely natural. On the other hand, in the speaker similarity test, each listener was given a pair of audio stimuli then was asked to judge whether they were produced by the same speaker. The similarity decision was based on two main scores, i.e., "same" or "different", with two confidence measures, i.e., "sure" or "not sure". The number of distinct test utterances per combination of systems and speaker pair in the MOS test was four, whereas in the similarity test it was three, which were randomly selected for each listener from a testing set consisting of 35 utterances. The number of listeners was 10. Natural speech waveforms were also included in both the MOS and similarity tests with the same number of distinct utterances per speaker in each corresponding test. This configuration gave a total of 128 audio samples to be evaluated in the MOS test and 102 audio samples to be evaluated in the similarity test for each listener. To summarize the results of the MOS test, these values were computed: the average value in each conversion pair/category; the 95% confidence interval of the sample average; and the p-value using the Mann–Whitney U test [135] with $\alpha < 0.05$ (two-tailed) to infer the statistical significance of the best system in each conversion pair/category. To summarize the similarity test, the total similarity score was computed by summing the "same-sure" and the "same-not sure" decisions. Similarly, the statistical inference of the best system in each conversion pair/category was computed using the Mann–Whitney U test with $\alpha < 0.05$ (two-tailed).

The results of the MOS test are given in Tables 5.5 and 5.6 for the converted and natural speech waveforms, respectively. The results show that the WaveNet model fine-tuned with oversmoothed features generated through the proposed CycleRNN spectral model with a weight of cyclic loss of 1e-6 (WNFT-Cyc6) yields a statistically significantly higher score than the other conversion categories. From the conversion pairs/categories, it can be observed that the proposed WNFT-Cyc6 system always yields better and more consistent performances, usually having a significant difference, than the other systems, especially compared with the conventional WaveNet model fine-tuned with natural features (WN-NoGV and WN-DiffGV) and DiffGV systems. The WaveNet fine-tuned with oversmoothed features from concatenated spectral mappings (WNFT-Cat) also gives reasonable naturalness performance, as can be predicted from the objective measurements, and more consistent results than WNFT-Cyc0, with similar performance to WNFT-Cyc3 but usually lower performance than WNFT-Cyc6. Note that for DiffGV-based waveform generation, because of the avoidance of vocoder-based excitation in the same-gender conversions, much better naturalness performance than that for cross-gender conversions can be observed. The proposed WNFT-Cyc6 also shows strong cross-gender conversion performance, especially for male-to-female conversion, compared to the other systems. In short, the results show significantly improved naturalness of the converted speech waveforms for the proposed WaveNet fine-tuning using the CycleRNN-based spectral mapping with balanced spectral mapping accuracy, i.e., with a cyclic weight of 1e-6 in this work.

The results of the similarity tests for the converted and natural speech waveforms are given in Tables 5.7 and 5.8, respectively. Similarly, to the naturalness performance obtained from the MOS test, the proposed WNFT-Cyc6 system gives higher performance for all conversion categories, mostly with statistical significance, than the other systems. This particularly applies especially when it is compared with the conventional WaveNet fine-tuning with natural target features (WN-NoGV and WN-DiffGV) as well as the DiffGV-based waveform generation. Note that owing to the excitation of the original waveform, DiffGV gives the worst similarity performance, even though it gives quite high naturalness for same-gender conversions. These speaker similarity test results also have a similar tendency when comparing the results for different weights of cyclic loss, i.e., 1e0 (WNFT-Cyc0), 1e-3 (WNFT-Cyc3), and 1e-6 (WNFT-

Cyc6), where WNFT-Cyc6 yields the best and most consistent performance, followed by WNFT-Cyc3 then WNFT-Cyc0. The importance of tuning the weight of the cyclic loss is again emphasized by the similarity performance for the latter two weights, where the WaveNet model fine-tuned with oversmoothed features from concatenated mappings (WNFT-Cat) yields a better and more consistent performance than WNFT-Cyc0. The WNFT-Cat system indeed gives a reasonable performance compared with the proposed WNFT-Cyc6, even though its overall similarity scores are still inferior to those of the WNFT-Cyc6. Moreover, only the proposed WNFT-Cyc6 gives accuracy consistently higher than 70% for both of same-gender and cross-gender conversion from all speaker pairs, which allows for concluding the effectiveness of the proposed method in improving the speaker conversion accuracy of converted speech waveforms. All audio samples are available at `http://bit.ly/2RkLmXC`.

5.5 Discussion

A thorough objective and subjective experiments have been presented to assess the proposed CycleRNN spectral modeling in performing WaveNet fine-tuning in a VC framework. A correlation was observed between the objective measurements of the accuracy and variance of the spectral trajectory and the subjective perceptual results. The CycleRNN with a cyclic loss weight of 1e-6 (CycRNN6) gives the most balanced values of MCD and LGD as shown in Section 5.4.2. It also gives the best overall perceptual performance for both naturalness and speaker conversion accuracy as shown in Section 5.4.3. Both performances are followed by those of the concatenated RNN mappings (CatRNN), CycleRNN with a cyclic loss weight of 1e3 CycRNN3), and CycleRNN with a cyclic loss weight of 1e0 (CycRNN0). These results demonstrate the importance of monitoring objective measurements while developing a spectral mapping model in this framework because the same spectral features in WaveNet fine-tuning and in the conversion phase are not used. Further, it would be interesting to determine how well the proposed CycleRNN framework performs with the use of smaller cyclic loss weights, such as 1e-7 and 1e-8. These two values are at the lower limit in the experimental configuration of this work because with cyclic loss weight of 1e-9 and smaller, the cyclic loss does not converge to reasonable values.

Another factor that might be at least as important as the spectral modeling is the WaveNet fine-tuning phase. In this procedure, owing to the limited amount of training data of the target speaker in the experiments of this work, the overfitting condition has to be handled with care. This problem is overcome through the use of dropout connections with suitable locations for WaveNet modeling. Further, the use of convolutional layers to capture contextual frames of input features is also important to improve the generated waveform. The monitoring of WaveNet loss, i.e., binary cross-entropy, is crucial in determining the training duration. Thus, it is highly recommended that separate development and evaluation sets are used to monitor the fine-tuning procedure. Finally, it can be straightforwardly observed that the proposed technique within this work can be easily extended to various other neural waveform generators. Although, the use of more data can greatly alleviate the overfitting problem, the achievements of this work through the use of a limited amount of data will generally be beneficial for other researchers.

5.6 Conclusion

In this work, a novel parallel voice conversion (VC) framework based on the cyclic structure of a recurrent neural network (CycleRNN) and a finely tuned WaveNet vocoder has been presented. The CycleRNN spectral mapping model utilizes multiple losses (multitask learning model), which are the conversion (source-to-target) loss and cyclic (reconstructed target) loss. The CycleRNN architecture consists of two concatenated RNN modules where the reconstructed target spectral features are obtained by feeding the original target spectral to the first RNN then the second RNN. In contrast, in the conversion the input source features are fed to the second RNN module to obtain the converted source-to-target features. Different than the simple concatenated spectral mapping flow with the RNN (CatRNN), where two mapping modules are separately trained, i.e., target-to-source and source-to-target, in the proposed CycleRNN, the synchronization of the reconstructed target and the converted features owing to the use of the corresponding losses can be ensured. In experiments, it has been demonstrated that the proposed CycleRNN model with weight of the cyclic loss of 1e-6 (CycRNN6) gives statistically significantly better overall naturalness and conversion accuracy than the conventional direct waveform modification with global variance (DiffGV) waveform generation, the conventional WaveNet fine-tuning with natural target spectral features, CycleRNN with cyclic loss weights of 1e0 and 1e-3, and CatRNN. The experimental results also demonstrate the importance of both tuning the cyclic loss weight and monitoring with objective measurements while developing the statistical model, which in the experiments of this work were in agreement with the subjective perceptual results. In future work, the proposed concept to other neural waveform generators, such as WaveNet vocoder with shallow architecture [136] and nonparallel neural vocoder [137], and to non-parallel VC [29].

6 Non-Parallel Voice Conversion with Cyclic Variational Autoencoder

6.1 Introduction

In the previous two chapters, voice conversion (VC) frameworks have been presented, i.e., neural-network (NN)-based spectral and waveform modeling for VC in Chapter 4 and high-quality VC with fine-tuned WaveNet vocoder using CycleRNN spectral mapping in Chapter 5. Indeed, by looking back to other related works, within two decades, many speech applications have been realized by employing the VC framework, such as creation of speech database with various voice characteristics [57], singing voice conversion [4], recovery of impaired speech signal [5, 6], expressive speech synthesis [7, 8], body-conducted speech processing [9, 10], and articulatory controllable speech modification [18]. In this chapter, for improving the flexibility in the development of related applications, a VC technique that can be realized using easily available speech data, such as with non-parallel speech dataset, is presented.

Basically, in general, there are two main VC frameworks, non-parallel VC and parallel VC. In the non-parallel VC, it is not straightforward to measure the correspondence between source spectral features and the target spectral features, owing to the inavailability of optimization with paired utterances. On the other hand, in a parallel

VC [2,3], because of the availability of the paired utterances, their correspondence can be directly achieved by performing time-alignment, such as with dynamic-time-warping (DTW) algorithm. However, not all of the time a proper parallel dataset, i.e., where the source and the target speakers utter the same set of sentences, can be collected for the development of a VC system. Consequently, as the main focus in this work, a consideration for a reliable non-parallel VC using data-driven statistical modeling would be highly beneficial for real-life applications.

Indeed, the challenge in developing the non-parallel spectral conversion model has attracted many works within the recent years, such as: with the use of clustered spectral matching algorithms [138, 139]; with adaptation/alignment of speaker model parameters [140, 141]; with restricted Boltzmann machine [64]; with generative adversarial networks (GAN)-based methods [28, 142]; and with variational autoencoder (VAE)-based frameworks [38, 143–145]. In this work, the focus is on the use of VAE-based system, owing to its potential in employing latent space to represent common hidden aspects of speech signal, between different speakers, e.g., phonetical attributes. Further, its implementation can be flexibly realized through any network architectures, such as with convolutional or recurrent models.

In a VAE framework [37], a latent space, usually with a Gaussian prior, is used for encoding a set of input features. In a VAE-based VC [38], additional speaker-coding features are used, alongside the encoded latent features, to reconstruct the spectral features in the generation phase. Speaker-code associated with the source (original) speaker is used to estimate the reconstructed spectra, while speaker-code associated with a desired target speaker is used to estimate converted spectra. However, due to the non-parallel condition, the spectral model parameters are optimized with respect only to the reconstructed spectra. Hence, because of the only reliance in speaker-code capability to disentangle speaker identity, the performance of a conventional VAE-based VC is still insufficient.

In this work, to improve VAE-based VC, a cycle-consistent mapping flow [146] is proposed to be used, i.e., CycleVAE-based VC, that indirectly optimizes the conversion flow by recycling the converted spectral features. Specifically, in the CycleVAE, the converted features are fed-back into the system to generate corresponding cyclic reconstructed spectra that can be directly optimized. The cyclic flow can, then, be continued by feeding the cyclic reconstructed features back into the system. Therefore, the conversion flow, i.e., the estimation of converted spectra, is indirectly considered in the computation of both the reconstruction losses and the regularizations of latent space. In the experiments, it has been demonstrated that the proposed CycleVAE-based VC shows higher correlation degree of latent features, i.e., more similar latent attributes between different speakers (possibly within phonetical space), and higher accuracy of converted spectra. Perceptual evaluation also shows significant improvements in both quality and accuracy of converted speech, especially when the speaker identities are considerably distant, such as in cross-gender conversions.

6.2 Proposed CycleVAE-based VC

In this work, to improve the VAE-based VC, as illustrated in Fig. 6.1, CycleVAE is proposed, which is capable of recycling the converted spectra back into the system, so that the conversion flow is indirectly considered in the parameter optimization. A similar idea has also been proposed as a cycle-consistent flow in a self-supervised method for visual correspondence [146].

In the proposed CycleVAE-based VC, the parameter optimization is defined as fol-

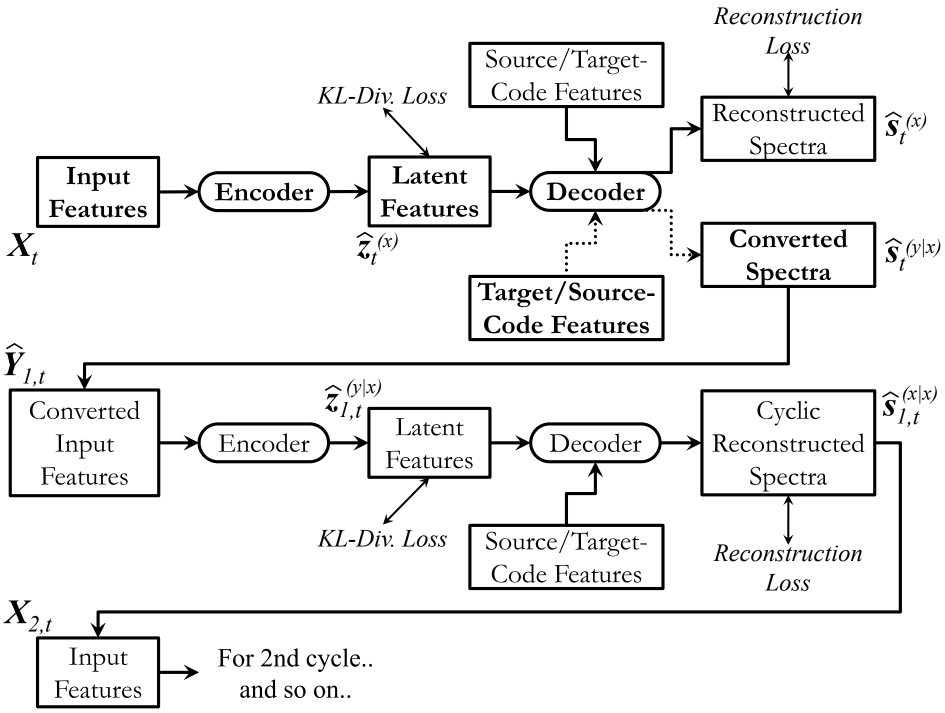


Figure 6.1: *Flow of the conventional VAE-based (upper-part) and the proposed CycleVAE-based (whole diagram) VC. Converted input features include converted excitation features, such as linearly transformed F_0 values. One full-cycle includes the estimation of both reconstructed and cyclic reconstructed spectra. Each of encoder and decoder networks are shared for all cycles.*

lows:

$$\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}\} = \underset{\boldsymbol{\theta},\boldsymbol{\phi}}{\operatorname{argmax}} \sum_{t=1}^{T} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{X}_t, \boldsymbol{c}^{(x)}, \boldsymbol{c}^{(y)}), \tag{6.1}$$

where

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{X}_t, \boldsymbol{c}^{(x)}, \boldsymbol{c}^{(y)}) = \sum_{n=1}^{N} & -D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{X}_{n,t})||p_{\boldsymbol{\theta}}(\boldsymbol{z}_t)) \\
& - D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\hat{\boldsymbol{Y}}_{n,t})||p_{\boldsymbol{\theta}}(\boldsymbol{z}_t)) \\
& + \mathbb{E}_{q_{\boldsymbol{\phi}(\boldsymbol{z}_t|\boldsymbol{X}_t)}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x)} = \boldsymbol{s}_t^{(x)}|\boldsymbol{z}_{n,t}, \boldsymbol{c}^{(x)})] \\
& + \mathbb{E}_{q_{\boldsymbol{\phi}(\boldsymbol{z}_t|\hat{\boldsymbol{Y}}_t)}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x|x)} = \boldsymbol{s}_t^{(x)}|\boldsymbol{z}_{n,t}, \boldsymbol{c}^{(x)})], \qquad (6.2)
\end{aligned}$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\hat{\boldsymbol{Y}}_{n,t}) = \mathcal{N}(\boldsymbol{z}_{n,t}; f_{\boldsymbol{\phi}}^{(\mu)}(\hat{\boldsymbol{Y}}_{n,t}), \mathrm{diag}(f_{\boldsymbol{\phi}}^{(\sigma)}(\hat{\boldsymbol{Y}}_{n,t})^2)), \qquad (6.3)$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x|x)}|\boldsymbol{z}_{n,t}, \boldsymbol{c}^{(x)}) \approx \mathcal{N}(\boldsymbol{s}_t^{(x)}; g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{n,t}^{(y|x)}, \boldsymbol{c}^{(x)}), \boldsymbol{I}), \qquad (6.4)$$

$$\hat{\boldsymbol{z}}_{n,t}^{(y|x)} = f_{\boldsymbol{\phi}}^{(\mu)}(\hat{\boldsymbol{Y}}_{n,t}) + f_{\boldsymbol{\phi}}^{(\sigma)}(\hat{\boldsymbol{Y}}_{n,t}) \odot \boldsymbol{\epsilon} \text{ s. t. } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \qquad (6.5)$$

where $\boldsymbol{s}_{n,t}^{(x)}$ and $\boldsymbol{s}_{n,t}^{(x|x)}$ are random variables, $\boldsymbol{s}_t^{(x)}$ is an observed value, and

$$\hat{\boldsymbol{Y}}_{n,t} = [\hat{\boldsymbol{e}}_t^{(y|x)^\top}, \hat{\boldsymbol{s}}_{n,t}^{(y|x)^\top}]^\top, \qquad (6.6)$$

$$\hat{\boldsymbol{s}}_{n,t}^{(y|x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{n,t}^{(x)}, \boldsymbol{c}^{(y)}), \qquad (6.7)$$

$$\hat{\boldsymbol{s}}_{n,t}^{(x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{n,t}^{(x)}, \boldsymbol{c}^{(x)}), \qquad (6.8)$$

$$\boldsymbol{X}_{n,t} = [\boldsymbol{e}_t^{(x)^\top}, \hat{\boldsymbol{s}}_{n-1,t}^{(x|x)^\top}]^\top, \qquad (6.9)$$

$$\hat{\boldsymbol{s}}_{n,t}^{(x|x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{n,t}^{(y|x)}, \boldsymbol{c}^{(x)}). \qquad (6.10)$$

The index of the n-th cycle is denoted as n. The total number of cycle is N. $\hat{\boldsymbol{Y}}_{n,t}$ denotes the converted input features at n-th cycle, $\hat{\boldsymbol{e}}_t^{(y|x)}$ denotes the converted source-to-target excitation features, e.g., linearly transformed F_0, $\hat{\boldsymbol{s}}_{n,t}^{(x|x)}$ denotes the cyclic reconstructed spectra at n-th cycle, and at $n = 1$, $\hat{\boldsymbol{s}}_{1,t}^{(y|x)} = \hat{\boldsymbol{s}}_t^{(y|x)}$, $\hat{\boldsymbol{s}}_{1,t}^{(x)} = \hat{\boldsymbol{s}}_t^{(x)}$, $\hat{\boldsymbol{z}}_{1,t}^{(x)} = \hat{\boldsymbol{z}}_t^{(x)}$ and $\boldsymbol{X}_{1,t} = \boldsymbol{X}_t$. Hence, in the proposed CycleVAE-based VC, the conversion flow is indirectly optimized through the consideration of the converted spectra $\hat{\boldsymbol{s}}_{n,t}^{(y|x)}$ in each n-th cycle.

6.3 Experimental Evaluation

6.3.1 Experimental conditions

A subset of the Voice Conversion Challenge (VCC) 2018 [108] dataset was used, which included four speakers, i.e., SF1, SM1, TF1, and TM1. The speaker notations are as follows: S denotes source speaker, T denotes target speaker, F denotes female speaker, and M denotes male speaker. The total number of utterances in the training and the testing sets were 81 and 35, respectively. The average length per one audio sample is about 3.5 seconds. To develop a non-parallel training dataset, the first 40 utterances were used for corresponding source speaker, while the last 41 were for the target speaker.

WORLD [22] package was used to perform speech analysis. As the spectral envelope parameters, the zeroth through 34th mel-cepstrum coeficients converted from the spectral envelope were used, which were extracted frame-by-frame. As the excitation features, log-scaled of continuous F_0 also including an unvoiced/voiced binary decision feature, and 2-dimensional aperiodicity coding coefficients were used. To perform excitation conversion, mean and variance transformation [3] was performed with respect to the log-scaled F_0 values. The sampling rate of the speech signal was 22,050 kHz. The number of FFT points was 1024. The frame shift was set to 5 ms.

To develop the spectral networks, a recurrent neural network (RNN)-based model was used, which was as follows: dilated convolutional layers were used, to capture the context of -4/+4 input frames, with a kernel size of 3 and 2 layers of 1 and 3 dilation, respectively; gated recurrent unit (GRU) [126] was used with 1024 hidden units and 1 hidden layer; a linear output layer was used; output frame was also fed-back into GRU. Fixed normalization and denormalization layers were used before convolutional

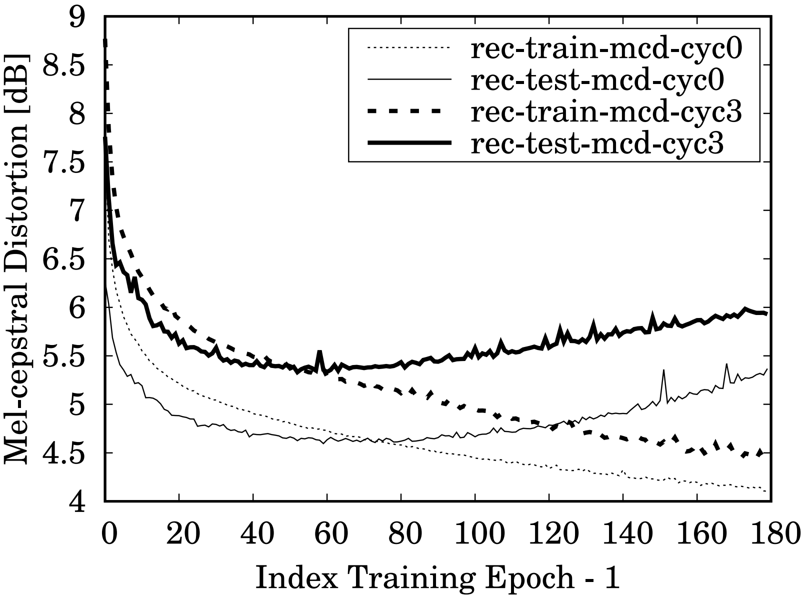


Figure 6.2: *Mel-cepstral distortion (mcd) of reconstructed (rec) spectra, estimated using the conventional VAE-based (cyc0) and the proposed CycleVAE-based (cyc3) VC, during 180 training epochs, for training (train) and testing (test) sets. mcds were computed with only the speech frames of the input speech.*

and after output layers, respectively, that were set with the statistics of training data. Dropout [132] layers were used with 0.5 probability after convolutional and GRU layers. Network parameters are initialized with Glorot [115] method, and optimized using Adam [116] with 0.0001 learning rate. A batch-frame size of 80 was used.

Four one-to-one spectral models were developed for each of the conventional VAE- and the proposed CycleVAE-based VC, with respect to the four corresponding speaker pairs, i.e., SF1-TF1, SF1-TM1, SM1-TF1, and SM1-TM1. To code the speaker identity, a binary decision value was used. Search of hyperparameters was conducted by varying the number of latent dimensions to 8, 16, 32, 50, and 64, and the number of cycles N, in Eq. (6.5), to 1, 2, 3, 4, and 5. The optimum number of latent dimensions for

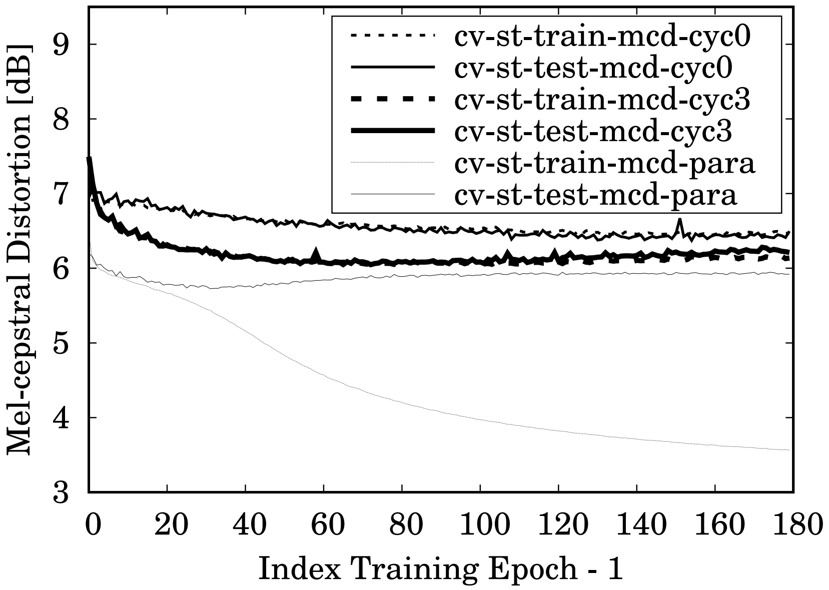


Figure 6.3: *Mel-cepstral distortion (mcd) of converted source-to-target (cv-st) spectra, estimated using the conventional VAE-based (cyc0) and the proposed CycleVAE-based (cyc3) VC, during 180 training epochs, for training (train) and testing (test) sets. mcds were computed, through DTW alignment, with only the speech frames of corresponding source and target speech.*

both VAE and CycleVAE was 16. The optimum number of cycles for CycleVAE was 3. Objective evaluation was performed to measure the accuracy of the reconstructed and the converted spectra, and the degree of latent features correlation. Another RNN-based parallel spectral conversion models were developed as the upper bound in measuring conversion accuracy. Subjective evaluation was performed to perceptually measure the quality and the accuracy of converted speech between conventional VAE and proposed CycleVAE [1].

[1] Implementation is being made available at `https://github.com/patrickltobing/cyclevae-vc`

Table 6.1: *Mel-cepstral distortion (MCD) of converted spectra (Cv) and GV-postfiltered [3] converted spectra (PF) with the conventional VAE, the proposed CycleVAE, and parallel spectral modeling as the lower bound, for each speaker-pair conversions. (S: source speaker; T: target speaker; F: female speaker; M: male speaker; Init. denotes the initial MCD values.)*

MCD [dB]	**Init.**	**VAE**		**CycleVAE**		**Parallel**	
		Cv	**PF**	**Cv**	**PF**	**Cv**	**PF**
SF1-TF1	8.18	6.41	6.95	**6.24**	**6.78**	5.92	6.42
SF1-TM1	8.73	6.49	7.03	**5.97**	**6.49**	5.60	6.03
SM1-TF1	9.06	6.83	7.42	**6.29**	**6.78**	6.00	6.43
SM1-TM1	7.68	5.74	6.15	**5.71**	**6.10**	5.36	5.72

6.3.2 Objective evaluation

Mel-cepstral distortion (MCD) [3] was used to measure the accuracy of both the reconstructed and the converted spectra. Their values are respectively charted, during 180 training epochs, in Figs. 6.2 and 6.3. It can be observed that the proposed CycleVAE-based VC yields higher accuracy of converted spectra and lower accuracy of reconstructed spectra compared to the conventional VAE. This trend is somewhat inline with [147], where reconstruction performance is not a proper measure for a better disentanglement of speaker identity (or for better conversion performance). Moreover, MCD values of converted spectra were also computed after applying global variance (GV)-postfilter [3], as given in Table 6.1. The result shows that the proposed CycleVAE is more suited to additional postfiltering phase compared to the conventional VAE, especially when the speaker identities are considerably distant.

To measure the condition of the latent features, the cosine similarities between the latent features of the source and of the target speaker within the same utterances

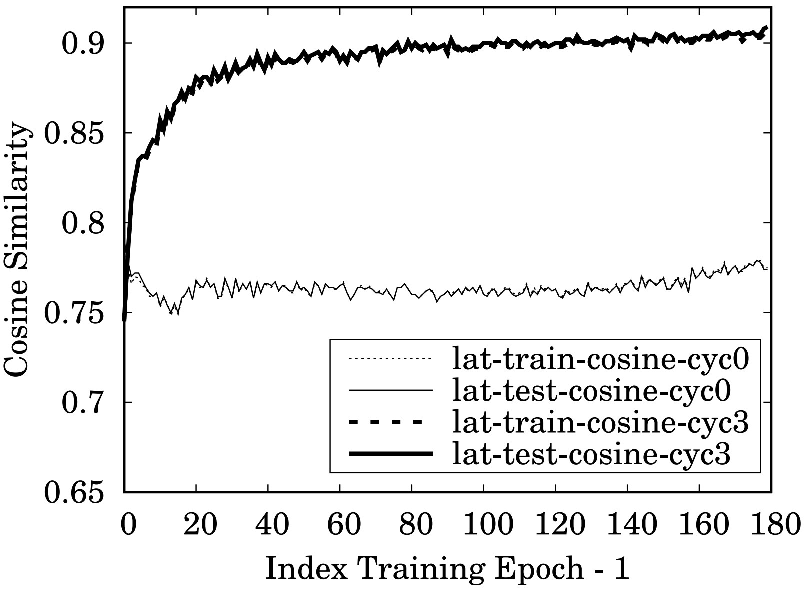


Figure 6.4: *Cosine similarity (cosine) between latent features of corresponding source and target speech, encoded with the conventional VAE-based (cyc0) and the proposed CycleVAE-based (cyc3) VC, during 180 training epochs, for training (train) and testing (test) sets. cosines were computed, through DTW alignment, with only the speech frames of source and target speech.*

were computed, which were charted during 180 training epochs, as in Fig. 6.4. It can be clearly seen that the proposed CycleVAE-based VC generates latent features with higher correlation degree compared to conventional VAE. As studied in [148], higher cosine similarities would be produced by latent attributes that represent either equal phonetic space or equal speaker identities. Hence, CycleVAE is more likely to give latent representations that are closer to phonetic domain due to different speaker identities.

Table 6.2: *Result of preference test on speech quality for all, same-gender (S-Gender), and cross-gender (X-Gender) conversion categories using the conventional VAE and the proposed CycleVAE-based VC. CI denotes the 95% confidence interval of the sample mean. p-values were computed using the two-tailed Mann–Whitney U-test with $\alpha < 0.05$. Bold indicates statistically significant better scores.*

Quality Preference	**VAE**	**CycleVAE**	**CI**	**p-value**
All	40.83%	**59.17%**	±6.27%	6.01e-05
S-Gender	52.50%	47.50%	±9.07%	4.40e-01
X-Gender	29.17%	**70.83%**	±8.25%	1.18e-10

6.3.3 Subjective evaluation

Perceptual evaluation was performed to compare the quality and the conversion accuracy of converted speech, between the conventional VAE- and the proposed CycleVAE-based VC, by conducting two forced-choice preference tests. In the quality preference test, each listener was presented with two audio stimuli at a time, and was asked to choose a prefered audio by considering both speech naturalness and intelligibility. In the similarity preference test, i.e., to measure the conversion accuracy, each listener was given two audio stimulis, and a reference audio with different utterance, then, was asked to choose a prefered audio that has the closer speaker characteristics to the reference speaker. The numbers of distinct utterances in quality and similarity tests were 6 and 5, respectively, which were randomly chosen from the testing set. Converted speech using parallel spectral models were also included. GV-postfiltered converted spectra was used. The number of listeners was 10.

Table 6.3: *Result of preference test on speaker similarity (Spk. Sim.) for all, same-gender (S-Gender), and cross-gender (X-Gender) conversion categories using the conventional VAE and the proposed CycleVAE-based VC. CI denotes the 95% confidence interval of the sample mean. p-values were computed using the two-tailed Mann–Whitney U-test with $\alpha < 0.05$. Bold indicates statistically significant better scores.*

Spk. Sim. Preference	**VAE**	**CycleVAE**	**CI**	**p-value**
All	39.00%	**61.00%**	±6.82%	1.11e-05
S-Gender	46.00%	54.00%	±9.94%	2.59e-01
X-Gender	32.00%	**68.00%**	±9.30%	3.81e-07

The results of quality and similarity preference tests are given in Tables 6.2 and 6.3, respectively. These results show that the proposed CycleVAE-based VC significantly improves the overall quality and accuracy of converted speech, especially for cross-gender (SF1-TM1, SM1-TF1) conversions, compared to conventional VAE. Their performances for same-gender conversions are statistically similar. This tendency is inline with the objective measurements shown in Table 6.1, where the conventional VAE-based VC suffers from degradation in cross-gender conversions and the CycleVAE significantly improves them. All audio samples and complete perceptual results can be accessed at `http://bit.ly/2Wg3oIt`.

6.4 Conclusions

A novel framework to improve conventional VAE, for a non-parallel VC, by using a cycle-consistent flow, i.e., the CycleVAE, has been presented. Specifically, the converted spectra, which is not directly optimized, is recycled back into the system, to generate cyclic reconstructed spectra that can be directly optimized. The cyclic flow can be continued by feeding the cyclic reconstructed features back into the system. The experimental results demonstrate that the proposed CycleVAE-based VC yields higher correlation degree of latent features and more accurate converted spectra, while significantly improves the quality and conversion accuracy of the converted speech. Future work includes development of many-to-many VC, and incorporates the use of discrete latent space [149], better prior [150], i-vector [151], additional classifier/discriminator network [145], duration modeling with recurrent Markov structure [152,153], and neural waveform generator [14, 136] to produce naturaly sounding converted speech [26, 36].

7 Conclusions

7.1 Summary of the Thesis

As an essential element in daily-life and communication, human voice (speech) is inevitably versatile and valuable. Automation of speech processing through the use of recent progress in machine learning techniques will definitely be a fine contribution for the society. As within this thesis, this may also include investigation on techniques for achieving high-quality and flexible voice conversion. Using a voice conversion, one can perform a transformation of the voice characteristics from a source speaker into that of a target speaker. Such framework will indeed be beneficial in various real-world applications including, but not limited to, entertainment, education, medical, and within the speech research area itself. To achieve high-quality and flexible voice conversion framework, within this thesis, specifically, statistical techniques for spectral mapping and waveform modeling have been presented.

In this thesis, two main approaches have been studied to achieve spectral mapping modeling. First is the use of articulatory information for developing a voice modification system with intuitive/flexible manipulation of the positions of speech organs. This has the potential to be developed into voice conversion by the modeling of the vocal-tract shape, though it is not straightforward to obtain articulatory data for statistical model development. On the other hand, the second method is to perform mapping of spectral envelope parameters between source and target speakers, which can be obtained in a more straightforward/flexible manner from a speech signal compared to the articulatory data. Further, ultimate flexibility in the development of voice conversion system has also been presented by means of nonparallel spectral mapping modeling, where source and target speakers do not have to utter a same set of sentences, i.e., arbitrary speech data can be used.

Finally, to achieve high-quality voice conversion, statistical waveform modeling based on neural network (neural vocoder), particularly WaveNet vocoder is investigated. The advantage of data-driven neural vocoder is the concept of data-driven approach in its development, which does not use any assumptions on speech production mechanism, unlike its conventional rule-based vocoder counterpart. Hence, the quality of the synthetic speech can be close to that of the natural speech, provided that the statistical model is well-trained/developed. In this thesis, this advantage of neural vocoder is exploited for the use of voice conversion, especially its capability of performing compensation of mismatches between the neural vocoder model and the spectral mapping model. Two main approaches have been investigated to address mismatches between spectral and waveform model, i.e., with a postprocessing method and with direct fine-tuning approach of the neural vocoder, on which they have proved to surpass the performance of the conventional voice conversion with vocoder-based generation to ultimately achieve high-quality converted speech output.

Chapter 2 has described the general overview of the thesis and its related works. These include: the acoustic-to-articulatory inversion mapping problem and articulatory-to-acoustic production mapping problem for the development of voice modification with articulatory mapping and manipulation; the works on voice conversion framework and its several issues related to nonparallel spectral mapping modeling, and speech quality degradation due to the oversmoothing and due to the use of vocoder-based excitation generation assumption; and lastly, recent works on the neural vocoder framework, such as WaveNet vocoder, which is capable of producing natural sounding synthetic speech, as well as its possible limitation when used in a voice conversion system.

In Chapter 3, the voice modification system with articulatory mapping and manipulation has been presented. This framework is realized through the use of a sequential mapping flow between Gaussian mixture model (GMM)-based inversion and production mappings, where intermediate representations of articulatory parameters can be intuitively manipulated. Articulatory manipulation is also performed while considering their interdimensional correlation, which can be derived from the statistical inversion mapping model. Additionaly, direct waveform modification with spectrum differential technique is also used, within the production mapping flow, to generate high-quality modified speech signal by avoiding the use vocoder-based excitation generation assumption. In the experiments, it has been demonstrated that the proposed system is capable of producing high-quality modified speech for varying articulation effort, such as hypo- and hyper-articulation, and is capable of producing modified vowel sounds through manipulating the positions of the tongue.

In Chapter 4, the voice conversion framework based on neural network (NN) architecture, for spectral mapping modeling and waveform modeling, has been elaborated. Specifically, the voice conversion system utilizes a structure of deep mixture density network for the spectral mapping solution, and employs the use of WaveNet vocoder to achieve better quality of converted speech waveform. However, there exists an issue of mismatches between the spectral and the waveform models, where the converted spectral features estimated from the statistical mapping model are not the same as the natural spectral features used in training the WaveNet. A postprocessing method based on the direct waveform modification technique is used to reduce these mismatches. In the experimental evaluation, the NN-based voice conversion system is capable of achieving higher quality and speaker-similarity of converted speech for cross-gender situations compared to conventional voice conversion with vocoder-based excitation generation assumption, while giving higher speaker-similarity in same-gender situations compared to the conventional voice conversion with direct waveform modification technique.

In Chapter 5, this thesis has introduced a novel framework to improve the performance of the voice conversion system with WaveNet vocoder. Specifically, to directly address the mismatches between the statistical spectral mapping model and the WaveNet vocoder, a pretrained WaveNet model (trained by using natural speech features) is fine-tuned with estimated (oversmoothed) spectral parameters. However, in a voice conversion, it is not straightforward to obtain these oversmoothed spectral parameters, owing to the differences in temporal structure between source and target speakers. The converted source-to-target spectral parameters cannot be used to fine-tune a WaveNet vocoder for the target speech waveform. To solve this problem, a cyclic structure of recurrent neural network (CycleRNN) is proposed for the spectral mapping solution, which is capable of estimating both of the converted source-to-target spectra and the oversmoothed target spectra, where the latter has the same temporal structure as the target speech through the use of the cyclic flow. The experimental results have demonstrated the effectiveness of the CycleRNN-based spectral mapping for fine-tuning a WaveNet vocoder in a voice conversion system by the significant displays of improvements in both of the speaker similarity and speech quality of the converted speech compared to the best previous voice conversion with posprocessing method for WaveNet vocoder.

Lastly, chapter 6 has given a novel solution for the non-parallel spectral modeling in voice conversion. In a completely non-parallel speech dataset, the source and the target speakers utter different sets of sentences. Hence, there are no pairings can be made for the optimization of the statistical model. Compared to parallel speech dataset, non-parallel set is easier to obtain and, therefore, has greater availability/flexibility for development of speech applications. Moreover, it is also not possible to obtain parallel speech data where the speakers are speaking in different languages. To solve this problem, this thesis proposes to use the variational autoencoder (VAE)-based framework for the voice conversion, where a regularized latent space is used as a shared space between different speakers. Time-invariant speaker code is used to determine the speaker characteristics and the VAE-based network is optimized based on the reconstructed spectral features and latent space regularization. However, due the non-consideration of converted spectral features in model optimization, the performance of VAE-based network is significantly degraded for the voice conversion. To improve, a cyclic structure of VAE (CycleVAE) has been proposed in this thesis, which is capable of recycling the converted spectral parameters back into the system to obtain corresponding cyclic reconstructed spectra that can be directly optimized. Hence, in the optimization, the condition of spectral conversion is also considered within the CycleVAE-based model. Evidently, the experimental results have demonstrated the effectiveness of the CycleVAE-based framework that significantly improves the performance of non-parallel voice conversion for the speaker similarity and the speech quality.

7.2 Future Work

Although comprehensive investigations and studies have been conducted in this thesis towards the development of high-quality and flexible voice conversion system, there is still a lot of works need to be done for continuous improvements.

Table 7.1: *Based on the contribution of each chapter (chap.) within this thesis for realization of high quality and flexible voice conversion of Table 1.1, future work may include a combination of Chapter 6, 5, and 3. This way, high quality output from fine-tuned neural vocoder can be achieved, along with flexible system development using nonparallel spectral modeling and flexible control of speech features through possible manipulation of latent space utilized within the nonparallel spectral model.*

Aspect	**Technique**	**Chap. 3**	**Chap. 4**	**Chap. 5**	**Chap. 6**	**Chaps. 6-(3)-5**
Flexibility	**Control**	○				○
	Parallel		○	○		
	Nonparallel				○	○
High quality	**Wav-mod**	○	○			
	Wav-gen	○	○	○		○
	FT wav-gen			○	○	○

Physical constraints for speech modification with articulatory control:

Naturally, speech production mechanism in human vocal tract is bounded by the physical constraints of the speech organs (articulators). Therefore, it is much more reasonable to adopt these constraints in the development of an articulatory controllable speech modification and the like. Further, a merger between statistical modeling and direct physical approach would be beneficial in general for both real-world applications and research studies. Such approach will also enable the physical vocal-tract shape modeling, where interpretative voice conversion may actually be realized through the conversion of vocal-tract shape between speakers.

Focus on improvements of the CycleVAE-based voice conversion:

The CycleVAE-based method for non-parallel voice conversion is a very versatile framework, which deserves continuous enhancements. First, investigations on the use of other latent space, such as with Laplacian distribution or possibly with a discrete space, e.g., using vector-quantized dictionary, are necessary. Secondly, examinations of its performance with various type of speech database, such as cross-language speech data or different type of speech, e.g., body-conducted speech, are worth to be conducted to see the extent of its versatileness and how can it be improved to accomodate the corresponding needs. Third, augmentation of speech dataset would also be straightforward for a CycleVAE-based voice conversion, such as by using waveform similarity based overlap-add (WSOLA) [154] technique of fundamental frequency (F0) transformation to obtain speech waveforms with varying F0 level corresponds to possible variation of speaker pair in a dataset. In this case, the WSOLA-based method can actually be used to also making it possible in using direct waveform modification [155] for cross-gender conversion in a straightforward manner. Fourth, the use of continuous space in speaker-coding features would be necessary to achieve speaker interpolation, where the speaker characteristics of a speech signal can be conveniently moved between different speaker space. Various improvements are also possible to be integrated in this resourceful CycleVAE framework, such as duration modeling, joint optimization with neural vocoder, and real-time processing, which are briefly mentioned within the next following contents. These points are also shown in Table 7.1, where the CycleVAE framework in Chapter 6 serves as a fundamental system for flexible nonparallel spectral modeling, then the fine-tuning approach of neural vocoder in Chapter 5 can be straightforwardly applied to achieve high-quality output, and finally flexibility in control can be achieved, as in Chapter 3, by further exploring the usage of latent space for a set of possible controllable/versatile features.

Duration modeling in a voice conversion framework:

To develop a voice conversion system, it would also be useful to handle the modeling of the duration of the speech, which might naturally varies between different speakers. In doing so, one may resort to a sequence-to-sequence-based method that inherently models the duration of the speech features. However, taking into account that the real-world applications of voice conversion in most of the time require streaming-like procedure, sequence-based models might face quite difficulties. Hence, a solution of duration modeling that can handle real-time low-latency processing, such as by handling segmental duration, should be thought about and investigated. One possible way is by incorporating the use of hidden semi-Markov structure in a recurrent neural network [152, 153] or with VAE-based non-linear switching dynamical systems (SNLDS) [156], which will definitely be investigated for future work. Further, by being able to automatically determine segmental durations and also able to extract speaker-independent features within the latent space for a voice conversion system, automatic speech recognition (ASR) and text-to-speech (TTS) systems can be easily realized by using the encoder network and the decoder network, respectively. As such is the versatile VC system.

Joint optimization of spectral modeling and neural vocoder:

Although, in this thesis, significant improvements of quality and accuracy have been shown by fine-tuning of neural vocoder with respect to the spectral modeling output, such procedure is still cumbersome due to the need of separate model trainings and fine-tuning. Especially, in the case of voice conversion, where there may exist a lot of possible speaker pairs. Therefore, to improve the efficiency in development, it is worthwhile to start resorting to a method that can jointly optimize both of the spectral (or speech features in general) modeling and the neural vocoder. One may apply the use of language-dependent features to develop a unified speech modeling, such as with phonetical posteriorgrams (PPG), though it may not be straightforward for cross-language voice conversion, and it also requires a reliable automatic speech recognition (ASR) system. Another possible way is to utilize the latent space that is shared between speakers as in the CycleVAE framework, for example, to be used as the conditioning features in a neural vocoder. In the latter case, the CycleVAE and the neural vocoder might be jointly developed. However, there may arise several problems in the implementations, owing to the instability when training from scratch, and so on. Hence, future investigations are worth to be conducted.

Real-time processing for real-world applications:

Finally, to deploy a voice conversion system for real-world applications, it is inevitable that real-time processing can be administered. This, specifically, relates to the low-latency streaming-like procedure, where the output of the system should be obtained within a certain delay range with respect to the input. In the case of having two separate statistical modelings, i.e., for spectral mapping model and for speech waveform model, considerations of both of the size of the models and the processing-time need to be taken into account, especially for neural vocoder that performs in the waveform-sample domain. Implementation of neural network architecture in central processing unit (CPU) instead of graphics processing unit (GPU) would also be necessary in the view of client-side deployments.

Acknowledgments

I would like to convey my deepest gratitude to my thesis advisor, Professor Tomoki Toda of Nagoya University, for everything that he has done to support my study and research.

I would like to express my sincere appreciation to Professor Kazuya Takeda of Nagoya University for his sound advices and support to broaden my research work.

I would also like to note a genuine thankfulness to Associate Professor Hirokazu Kameoka of NTT Corporation for providing the invaluable opportunities on the internship works at the NTT's Communication Science Laboratory.

I would especially like to express my humble gratefulness to NEC C&C Foundation for the opportunity to receive an indispensable support on my research work.

I would like to acknowledge my earnest thanks to the staffs of Toda Laboratory and Takeda Laboratory for their patience and kind assistances. I would also like to convey my thanks to laboratory colleagues for their support, especially Dr. Kazuhiro Kobayashi of TARVO, Inc., and Dr. Tomoki Hayashi of Human Dataware Lab. Co., Ltd., for their aids on the research works.

Finally, I would like to express my wholehearted recognition to my family for their encouragement and support throughout my study and research life.

References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. of Jpn. (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.

[3] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[4] K. Kobayashi, T. Toda, and S. Nakamura, "Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential," *Speech Commun.*, vol. 99, pp. 211–220, 2018.

[5] A. B. Kain, J.-P. Hosom, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Commun.*, vol. 49, no. 9, pp. 743–759, 2007.

[6] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on spectral subtraction and statis-

tical voice conversion," in *Proc. INTERSPEECH*, Lyon, France, Sep. 2013, pp. 3067–3071.

[7] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Commun.*, vol. 51, no. 3, pp. 268–283, 2009.

[8] O. Türk and M. Schröder, "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 965–973, 2010.

[9] A. Subramanya, Z. Zhang, Z. Liu, and A. Acero, "Multisensory processing for speech enhancement and magnitude-normalized spectra for speech modeling," *Speech Commun.*, vol. 50, no. 3, pp. 228–243, 2008.

[10] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 9, pp. 2505–2517, 2012.

[11] H. Dudley, "Remaking speech," *J. Acoust. Soc. Am.*, vol. 11, no. 2, pp. 169–177, 1939.

[12] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations.* Walter de Gruyter, 1970, vol. 2.

[13] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representation using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[14] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR arXiv preprint arXiv:1609.03499*, 2016.

[15] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "NU voice conversion system for the Voice Conversion Challenge 2018," in *Proc. Odyssey*, Les Sables d'Olonne, France, Jun. 2018, pp. 219–226.

[16] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.*, vol. 50, no. 3, pp. 215–227, 2008.

[17] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 21, no. 1, pp. 207–219, 2013.

[18] P. L. Tobing, K. Kobayashi, and T. Toda, "Articulatory controllable speech modification based on statistical inversion and production mappings," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2337–2350, 2017.

[19] A. A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 145–148.

[20] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electron. Commun. Jpn. (Part I: Commun.)*, vol. 66, no. 2, pp. 10–18, 1983.

[21] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proc. ICSLP*, Yokohama, Japan, Sep. 1994, pp. 1043–1046.

[22] 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, no. 7, pp. 1877–1884, 2016.

[23] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 1118–1122.

[24] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. IEEE ASRU*, Okinawa, Japan, Dec. 2017, pp. 712–718.

[25] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 1138–1142.

[26] P. L. Tobing, Y. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Voice conversion with CycleRNN-based spectral mapping and finely tuned WaveNet vocoder," *IEEE Access*, vol. 7, pp. 171 114–171 125, 2019.

[27] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. EUSIPCO*, Rome, Italy, Sep. 2018, pp. 2100–2104.

[28] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, “StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks,” in *Proc. SLT*, Athens, Greece, Dec. 2018, pp. 266–273.

[29] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, “Non-parallel voice conversoin with cyclic variational autoencoder,” in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 674–678.

[30] P. Mermelstein, “Articulatory model for the study of speech production,” *J. Acoust. Soc. Am.*, vol. 53, no. 4, pp. 1070–1082, 1973.

[31] B. S. Atal, J. J. Chang, M. V. Matthews, and J. W. Tukey, “Articulatory compensation: A study of ambiguities in the acoustic-articulatory mapping,” *J. Acoust. Soc. Am.*, vol. 60, no. S1, p. S77, 1976.

[32] H. Wakita, “Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 3, pp. 281–285, 1979.

[33] S. Maeda, “Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model,” in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. Netherlands: Springer, 1990, pp. 131–149.

[34] J. Schroeter and M. M. Sondhi, “Techniques for estimating vocal-tract shapes from the speech signal,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 133–150, 1994.

[35] P. L. Tobing, T. Hayashi, Y.-C. Wu, K. Kobayashi, and T. Toda, "An evaluation of deep spectral mappings and WaveNet vocoder for voice conversion," in *Proc. SLT*, Athens, Greece, Dec. 2018, pp. 297–303.

[36] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Voice conversion with cyclic recurrent neural network and fine-tuned WaveNet vocoder," in *Proc. ICASSP*, Brighton, UK, May 2019, pp. 6815–6819.

[37] D. P. Kingma and J. Ba, "Auto-encoding variational bayes," *CoRR arXiv preprint arXiv:1312.6114*, 2013.

[38] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. APSIPA*, Jeju, South Korea, Dec. 2016, pp. 1–6.

[39] P. Mermelstein, "Determination of the vocal tract-shape from measured formant frequencies," *J. Acoust. Soc. Am.*, vol. 41, no. 5, pp. 1283–1294, 1967.

[40] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, pp. 231–267.

[41] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *J. Acoust. Soc. Am.*, vol. 100, no. 3, pp. 1819–1834, 1996.

[42] S. Suzuki, T. Okadome, and M. Honda, "Determination of articulatory positions from speech acoustics by applying dynamic articulatory constraints," in *Proc. ICSLP*, Sydney, Australia, Dec. 1998, pp. 2251–2254.

[43] K. Richmond, S. King, and P. Taylor, “Modelling the uncertainty in recovering articulation from acoustics,” *Comput. Speech Lang.*, vol. 17, no. 2–3, pp. 153–172, 2003.

[44] K. Richmond, “Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion,” in *Proc. Int. Conf. Non-Linear Speech Process. (NOLISP)*, Paris, France, May 2007, pp. 263–272.

[45] S. Hiroya and M. Honda, “Estimation of articulatory movements from speech acoustics using an HMM-based speech production model,” *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 175–185, 2004.

[46] A. B. Youssef, G. B. Pierre Badin, and P. Heracleous, “Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models,” in *Proc. INTERSPEECH*, Brighton, United Kingdom, Sep. 2009, pp. 2255–2258.

[47] T. Toda, A. W. Black, and K. Tokuda, “Acoustic-to-articulatory inversion mapping with Gaussian mixture model,” in *Proc. INTERSPEECH*, Jeju, Korea, Oct. 2004, pp. 1129—-1132.

[48] P. L. Tobing, T. Toda, H. Kameoka, and S. Nakamura, “Acoustic-to-articulatory inversion mapping based on latent trajectory gaussian mixture model,” in *Proc. INTERSPEECH*, San Fransisco, CA, USA, Sep. 2016, pp. 953–957.

[49] P. L. Tobing, H. Kameoka, and T. Toda, “Deep acoustic-to-articulatory inversion mapping with latent trajectory modeling,” in *Proc. APSIPA*, Kuala Lumpur, Malaysia, Dec. 2017, pp. 1274–1277.

[50] B. S. Atal, J. J. Chang, M. V. Matthews, and J. W. Tukey, “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique,” *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1535–1555, 1978.

[51] T. Kaburagi and M. Honda, “Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database,” in *Proc. ICSLP*, Sydney, Australia, Dec. 1998, pp. 433–436.

[52] C. T. Kello and D. C. Plaut, “A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters,” *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2354–2364, 2004.

[53] S. Aryal and R. Gutierrez-Osuna, “Data driven articulatory synthesis with deep neural networks,” *Comput. Speech Lang.*, vol. 36, pp. 260–273, Mar. 2016.

[54] K. Nakamura, T. Toda, Y. Nankaku, and K. Tokuda, “On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum,” in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 93–96.

[55] T. Toda, A. W. Black, and K. Tokuda, “Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis,” in *5th ISCA Tutorial and Research Workshop on Speech Synthesis*, Pittsburgh, PA, USA, Jun. 2004, pp. 31–36.

[56] D. B. Childers, B. Yegnanarayana, and K. Wu, “Voice conversion: Factors responsible for quality,” in *Proc. ICASSP*, Tampa, FL, USA, Mar. 1985, pp. 748–751.

[57] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in *Proc. ICASSP*, Seattle, WA, USA, May 1998, pp. 285–288.

[58] F. Villavicencio and J. Bonada, “Applying voice conversion to concatenative singing-voice synthesis,” in *Proc. INTERSPEECH*, Makuhari, Japan, Sep. 2010, pp. 2162–2165.

[59] H. Doi, T. Toda, H. Saruwatari, and K. Shikano, “Alaryngeal speech enhancement based on one-to-many eigenvoice conversion,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 1, pp. 172–183, 2014.

[60] Y. Tajiri, K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “Non-audible murmur enhancement based on statistical conversion using air- and body-conductive microphones in noisy environments,” in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2769–2773.

[61] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, “Spectral mapping using artificial neural networks for voice conversion,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 954–964, 2010.

[62] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, “Voice conversion using deep neural networks with layer-wise generative training,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1859–1872, 2014.

[63] L. Sun, S. Kang, K. Li, and H. Meng, “Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,” in *Proc. ICASSP*, South Brisbane, Australia, Apr. 2015, pp. 4869–4873.

[64] T. Nakashika, T. Takiguchi, and Y. Minami, “Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2032–2045, 2016.

[65] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, Detroit, MI, USA, May 1995, pp. 660–663.

[66] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. C. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *CoRR arXiv preprint arXiv:1612.07837*, 2016.

[67] Y. Ai, H.-C. Wu, and Z.-H. Ling, "SampleRNN-based neural vocoder for statistical parametric speech synthesis," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5659–5663.

[68] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *CoRR arXiv preprint arXiv:1802.08435*, 2018.

[69] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Real-time neural text-to-speech with sequence-to-sequence acoustic model and WaveGlow or single Gaussian WaveRNN vocoders," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 1308–1312.

[70] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "FFTNet: A real-time speaker-dependent neural vocoder," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 2251–2255.

[71] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Investigations of real-time Gaussian FFTNet and parallel WaveNet neural vocoders with simple acoustic features," in *Proc. ICASSP*, Brighton, UK, May 2019, pp. 7020–7024.

[72] J.-M. Valin and J. Skoglund, "LPCNet: improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, Brighton, UK, May 2019, pp. 5891–5895.

[73] J.-M. Valin and J. Skoglund, "A Real-Time Wideband Neural Vocoder at 1.6 kb/s Using LPCNet," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 3406–3410.

[74] A. van den 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," *CoRR*, vol. abs/1711.10433, 2017.

[75] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," *CoRR*, vol. abs/1807.07281, 2018.

[76] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: a flow-based generative network for speech synthesis," in *Proc. ICASSP*, Brighton, UK, May 2019, pp. 3617–3621.

[77] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. ICASSP*, Brighton, UK, May 2019, pp. 5916–5920.

[78] Y.-C. Wu, K. Kobayashi, T. Hayashi, P. L. Tobing, and T. Toda, "Collapsed speech segment detection and suppression for wavenet vocoder," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 1988–1992.

[79] M. M. Sondhi and J. Schroeter, “A hybrid time-frequency domain articulatory speech synthesizer,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 35, no. 7, pp. 955–967, 1987.

[80] S. Parthasarathy, J. Schroeter, C. H. Coker, and M. M. Sondhi, “Articulatory analysis and synthesis of speech,” in *TENCON ’89. 4th IEEE Region 10 Int. Conf.*, Bombay, India, Nov. 1989, pp. 760–764.

[81] M. M. Sondhi, “Articulatory modeling: a possible role in concatenative text-to-speech synthesis,” in *IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, CA, USA, Sep. 2002, pp. 73–78.

[82] B. Bollepali, A. W. Black, and K. Prahallad, “Modeling a noisy-channel for voice conversion using articulatory features,” in *Proc. INTERSPEECH*, Portland, OR, USA, Sep. 2012, pp. 2202–2205.

[83] K. Kirchhoff, G. A. Fink, and G. Sagerer, “Combining acoustic and articulatory feature information for robust speech recognition,” *Speech Commun.*, vol. 37, no. 3, pp. 303–319, 2002.

[84] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, “Speech production knowledge in automatic speech recognition,” *J. Acoust. Soc. Am.*, vol. 121, no. 2, pp. 723–742, 2007.

[85] L. Revéret, G. Bailly, and P. Badin, “MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation,” in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 755–758.

[86] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka, “An audiovisual talking head for augmented speech generation: Models and animations based on a real

speaker ’ s articulatory data,” in *Proc. Articulated Motion and Deformable Objects (AMDO).* Mallorca, Spain: Berlin, Heidelberg: Springer, Jul. 2008, pp. 132–143.

[87] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, “Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract,” *Brain Lang.*, vol. 31, no. 1, pp. 26–35, 1987.

[88] A. A. Wrench and W. J. Hardcastle, “A multichannel articulatory database and its application for automatic speech recognition,” in *Proc. 5th Seminar of Speech Prod.*, Kloster Seeon, Bavaria, Germany, May 2000, pp. 305–308.

[89] K. Richmond, P. Hoole, and S. King, “Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus,” in *Proc. INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 1505–1508.

[90] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. C. Lammert, M. I. Proctor, V. Ramanarayanan, and Y. Zhu, “A multimodal real-time MRI articulatory corpus for speech research,” in *Proc. INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 837–840.

[91] B. J. Kröger, J. Gotto, S. Albert, and C. Neuschaefer-Rube, “A visual articulatory model and its application to therapy of speech disorders: a pilot study,” *Speech production and perception: Experimental analyses and models. ZAS Papers in Linguistics*, vol. 40, pp. 79–94, 2005.

[92] B. J. Kröger, V. Graf-Borttscheller, and A. Lowit, "Two- and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008, pp. 2639–2642.

[93] D. W. Massaro, "The psychology and technology of talking heads: Applications in language learning," in *Advances in Natural Multimodal Dialogue Systems*. Netherlands: Springer, 2005, vol. 30, pp. 183–214.

[94] B. J. Kröger, P. Birkholz, R. Hoffmann, and H. Meng, "Audiovisual tools for phonetic and articulatory visualization in computer-aided pronunciation training," in *Development of Multimodal Interfaces: Active Listening and Synchrony*. Berlin, Heidelberg: Springer, 2010, pp. 337–345.

[95] P. L. Tobing, T. Toda, G. Neubig, S. Sakti, S. Nakamura, and A. Purwarianti, "Articulatory controllable speech modification based on statistical feature mapping with Gaussian mixture models," in *Proc. INTERSPEECH*, Singapore, Sep. 2014, pp. 2298–2302.

[96] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Proc. INTERSPEECH*, Singapore, Sep. 2014, pp. 2514–2518.

[97] P. L. Tobing, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Articulatory controllable speech modification based on Gaussian mixture models with direct waveform modification using spectrum differential," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 3350–3354.

[98] A. Wrench. (1999) The MOCHA-TIMIT articulatory database. Queen Margaret University College. [Online]. Available: http://www.cstr.ed.ac.uk/artic/mocha.html

[99] B. Youssef, A. Badin, and G. Bailly, "Can tongue be recovered from face? The answer of data-driven statistical models," in *Proc. INTERSPEECH*, Makuhari, Japan, Sep. 2010, pp. 2002–2005.

[100] 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 Lang. Process.*, vol. 24, no. 4, pp. 755–767, 2016.

[101] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion based on direct waveform modification with global variance," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2754–2758.

[102] H. Kawahara, H. Katayose, A. de Cheveigné, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. EUROSPEECH*, Budapest, Hungary, Sep. 1999, pp. 2781–2784.

[103] B. Picart, T. Drugman, and T. Dutoit, "Analysis and synthesis of hypo- and hyperarticulated speech," in *7th ISCA Tutorial and Research Workshop on Speech Synthesis*, Kyoto, Japan, Sep. 2010, pp. 270–275.

[104] G. Sivaraman, V. Mitra, H. Nam, M. K. Tiede, and C. Y. Espy-Wilson, "Vocal tract length normalization for speaker independent acoustic-to-articulatory

speech inversion," in *INTERSPEECH*, San Francisco, CA, USA, 2016, pp. 455–459.

[105] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 4624–4627.

[106] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *INTERSPEECH*, Pittsburgh, MI, USA, 2006, pp. 2446–2449.

[107] T. Toda, L.-H. Chen, F. Villavicencio, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *Proc. INTERSPEECH*, San Francisco, CA, USA, Sep. 2016, pp. 1632–1636.

[108] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods," *CoRR arXiv preprint arXiv:1804.04262*, 2018.

[109] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[110] D. T. Chappell and J. H. Hansen, "Speaker-specific pitch contour modeling and modification," in *Proc. ICASSP*, Seattle, USA, May 1998, pp. 885–888.

[111] Z. Wu, T. Kinnunen, E. Chng, and H. Li, "Text-independent F0 transformation with non-parallel data for voice conversion," in *Proc. INTERSPEECH*, Makuhari, Japan, Sep. 2010, pp. 1732–1735.

[112] L.-H. Chen, L.-J. Liu, Z.-H. Ling, Y. Jiang, and L.-R. Dai, "The USTC system for Voice Conversion Challenge 2016: Neural network based approaches for spec-

trum, aperiodicity and F0 conversion," in *Proc. INTERSPEECH*, San Francisco, CA, USA, Sep. 2016, pp. 1642–1646.

[113] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 3844–3848.

[114] C.-C. Hsu, "PyWorldVocoder - A Python wrapper for World Vocoder." [Online]. Available: https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

[115] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, vol. 9, Sardinia, Italy, May 2010, pp. 249–256.

[116] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR arXiv preprint arXiv:1412.6980*, 2014.

[117] K. Kobayashi and T. Toda, "sprocket: Open-source voice conversion software," in *Proc. Odyssey*, Les Sables d'Olonne, France, Jun. 2018, pp. 203–210.

[118] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Commun.*, vol. 51, no. 10, pp. 920–932, 2009.

[119] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 19, no. 1, pp. 153–165, 2011.

[120] D. Erro, I. Sainz, E. Navas, and I. Hernaez, “Harmonics plus noise model based vocoder for statistical parametric speech synthesis,” *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 2, pp. 184–194, 2014.

[121] G. Degottex, P. Lanchantin, and M. Gales, “A pulse model in log-domain for a uniform synthesizer,” in *Proc. 9th ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, Sep. 2016, pp. 230–236.

[122] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, and R. A. Saurous, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 4779–4783.

[123] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, “WaveNet vocoder with limited training data for voice conversion,” in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 1983–1987.

[124] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black, “Style transfer through back-translation,” *CoRR arXiv preprint arXiv:1804.09000*, 2018.

[125] T. Hayashi, S. Watanabe, Y. Zhang, T. Toda, T. Hori, R. F. Astudillo, and K. Takeda, “Back-translation-style data augmentation for end-to-end ASR,” *CoRR arXiv preprint arXiv:1807.10893*, 2018.

[126] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” *CoRR arXiv preprint arXiv:1406.1078*, 2014.

[127] M. Morise, "A high-performance fundamental frequency estimator from speech signals," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 2321–2325.

[128] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Commun.*, vol. 67, pp. 1–7, 2015.

[129] M. Morise, "Error evaluation of an F0-adaptive spectral envelope estimator in robustness against the additive noise and F0 error," *IEICE Trans. Inf. Syst.*, vol. E98-D, no. 7, pp. 1405–1408, Jul. 2015.

[130] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 57–65, 2016.

[131] T. Hayashi, "WaveNet-Vocoder implementation with PyTorch." [Online]. Available: https://github.com/kan-bayashi/PytorchWaveNetVocoder

[132] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[133] K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5664–5668.

[134] P. L. Tobing, "dtw-c." [Online]. Available: https://github.com/patrickltobing/dtw_c

[135] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 3976–3980.

[136] P. L. Tobing, T. Hayashi, and T. Toda, "Investigation of shallow WaveNet vocoder with Laplacian distribution output," in *Proc. IEEE ASRU*, Sentosa, Singapore, Dec. 2019.

[137] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," *CoRR arXiv preprint arXiv:1910.11480*, 2019.

[138] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. Speech Audio Process.*, vol. 18, no. 5, pp. 944–953, 2010.

[139] H. Benisty, D. Malah, and K. Crammer, "Non-parallel voice conversion using joint optimization of alignment by temporal context and spectral distortion," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 7909–7913.

[140] M. Zhang, J. Tao, J. Tian, and X. Wang, "Text-independent voice conversion based on state mapped codebook," in *Proc. ICASSP*, Las Vegas, NV, USA, Mar. 2008, pp. 4605–4608.

[141] P. Song, W. Zheng, and L. Zhao, "Non-parallel training for voice conversion based on adaptation method," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 6905–6909.

[142] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality non-parallel voice conversion based on cycle-consistent adversarial network," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5279–5283.

[143] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, , and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein genera-

tive adversarial networks," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 3364–3368.

[144] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5274–5278.

[145] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," *CoRR arXiv preprint arXiv:1808.05092*, 2018.

[146] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," *CoRR arXiv preprint arXiv:1903.07593*, 2019.

[147] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using WaveNet autoencoders," *CoRR arXiv preprint arXiv:1901.08810*, 2019.

[148] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *CoRR arXiv preprint arXiv:1704.04222*, 2017.

[149] A. van den Oord and O. Vinyals, "Neural discrete representation learning," in *Adv. NIPS*, Long Beach, CA, USA, Dec. 2017, pp. 6306–6315.

[150] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Adv. NIPS*, Long Beach, CA, USA, Dec. 2017, pp. 1878–1889.

[151] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, "Non-parallel voice conversion using i-vector PLDA: Towards unifying speaker verification and transformation," in *Proc. ICASSP*, New Orleans, LA, USA, Mar. 2017, pp. 5535–5539.

[152] H. Dai, B. Dai, Y.-M. Zhang, S. Li, and L. Song, "Recurrent hidden semi-Markov model," in *Proc. ICLR*, Toulon, France, Apr. 2017.

[153] H. Liu, L. He, H. Bai, B. Dai, K. Bai, and Z. Xu, "Structured inference for recurrent hidden semi-Markov model," in *Proc. IJCAI*, Stockholm, Sweden, Jul. 2018, pp. 2447–2453.

[154] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. ICASSP*, Minneapolis, MN, USA, Apr. 1993, pp. 554–557.

[155] K. Kobayashi, T. Toda, and S. Nakamura, "F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential," in *Proc. IEEE SLT*, San Diego, CA, USA, Dec. 2016, pp. 693–700.

[156] Z. Dong, B. A. Seybold, K. P. Murphy, and H. H. Bui, "Collapsed amortized variational inference for switching nonlinear dynamical systems," *CoRR arXiv preprint arXiv:1910.09588*, 2019.

List of Publications

Journal Papers

1. P. L. Tobing, K. Kobayashi, and T. Toda, “Articulatory controllable speech modification based on statistical inversion and production mappings,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2337–2350, 2017.

2. P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, “Voice conversion with CycleRNN-based spectral mapping and finely tuned WaveNet vocoder,” IEEE Access, vol. 7, pp. 171114–171125, Dec. 2019.

International Conferences

1. P. L. Tobing, T. Toda, G. Neubig, S. Sakti, S. Nakamura, and A. Purwarianti, “Articulatory controllable speech modification based on statistical feature mapping with Gaussian mixture models,” Proc. INTERSPEECH, Singapore, Sep., 2014, pp. 2298–2302.

2. P. L. Tobing, K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “Articulatory controllable speech modification based on Gaussian mixture models with

direct waveform modification using spectrum differential," Proc. INTERSPEECH, Dresden, Germany, Sep., 2015, pp. 3350–3354.

3. P. L. Tobing, T. Toda, H. Kameoka, and S. Nakamura, "Acoustic-to-articulatory inversion mapping based on latent trajectory Gaussian mixture model," Proc. INTERSPEECH, San Fransisco, USA, Sep., 2016, pp. 953–957.

4. P. L. Tobing, H. Kameoka, and T. Toda, "Deep acoustic-to-articulatory inversion mapping with latent trajectory modeling," Proc. APSIPA, Kuala Lumpur, Malaysia, Dec., 2017, pp. 1274–1277.

5. P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "NU voice conversion system for the Voice Conversion Challenge 2018," Proc. Speaker Odyssey, Les Sables d ’ Olonne, France, Jun. 2018, pp. 219–226.

6. P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "An evaluation of deep spectral mappings and WaveNet vocoder for voice conversion," Proc. IEEE SLT, Athens, Greece, Dec. 2018, pp. 297–303.

7. P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Voice conversion with cyclic recurrent neural network and fine-tuned WaveNet vocoder," Proc. ICASSP, Brighton, UK, May 2019, pp. 6815–6819.

8. P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational autoencoder," Proc. INTERSPEECH, Graz, Austria, Sep. 2019, pp. 674–678.

9. P. L. Tobing, T. Hayashi, and T. Toda, "Investigation of shallow WaveNet vocoder with Laplacian distribution output," in Proc. IEEE ASRU, Sentosa, Singapore, Dec. 2019, pp. 176–183.

* Seven other papers were published as a co-author.

Domestic Conferences

1. P. L. Tobing, T. Toda, G. Neubig, S. Sakti, S. Nakamura, and A. Purwarianti, “Articulatory controllable speech modification based on statistical feature mapping with Gaussian mixture models,” 電子情報通信学会技術研究報告, vol. 114, No. 365, pp. 57–62, Dec. 2014.

2. P. L. Tobing, K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “Articulatory controllable speech modification based on Gaussian mixture models with direct waveform modification using spectrum differential,” 日本音響学会講演論文集, pp. 267–268, Mar. 2015.

3. P. L. Tobing, K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “An evaluation of articulatory controllable speech modification based on Gaussian mixture models with direct waveform modification,” 日本音響学会講演論文集, pp. 221—222, Sep. 2015.

4. P. L. Tobing, T. Toda, H. Kameoka, and S. Nakamura, “An investigation of acoustic-to-articulatory inversion mapping with latent trajectory Gaussian mixture model,” 日本音響学会講演論文集, pp. 227—228, Mar. 2016.

5. P. L. Tobing, H. Kameoka, and T. Toda, “Acoustic-to-articulatory inversion mapping with variational latent trajectory Gaussian mixture model,” 電子情報通信学会技術研究報告, vol. 116, no. 475, pp. 291—296, Mar. 2017.

6. P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, “Development of NU voice conversion system 2018,” 電子情報通信学会技術研究報告, vol. 117, no. 517, pp. 203—208, Mar. 2018.

7. P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, “Voice conversion with cyclic recurrent neural network for WaveNet fine-tuning,” 日本音響学会講演論文集, pp. 1319—1320, Mar. 2019.

* Four other papers were published as the first author.

Awards

1. 日本音響学会第 11 回 学生優秀発表賞, Mar. 2015.

2. NEC C&C 2018 年度外国人研究員助成事業